Soc504: Generalized Linear Models

Brandon Stewart[1]

Princeton

February 22 - March 15, 2017

[1] These slides are heavily influenced by Gary King with some material from Teppei Yamamoto, Patrick Lam and Yuri Zhukov. Some individual vignettes are built from the collective effort of generations of teaching fellows for Gov2001 at Harvard.

Followup

Followup

- Questions?

Followup

- Questions?
- Replication Stories?

Binary Outcomes

- Binary outcome variable:

$$Y_i \in \{0, 1\}$$

Binary Outcomes

- Binary outcome variable:

$$Y_i \ \in \{0, 1\}$$

- Examples in social science:

Binary Outcomes

- Binary outcome variable:

$$Y_i \in \{0, 1\}$$

- Examples in social science: **numerous!**

Binary Outcomes

- Binary outcome variable:

$$Y_i \ \in \{0, 1\}$$

- Examples in social science: **numerous!**
  - Turnout ($1 =$ vote; $0 =$ abstain)

Binary Outcomes

- Binary outcome variable:

$$Y_i \in \{0, 1\}$$

- Examples in social science: **numerous!**
  - Turnout ($1 =$ vote; $0 =$ abstain)
  - Education ($1 =$ completed hs; $0 =$ dropped out)

Binary Outcomes

- Binary outcome variable:

$$Y_i \in \{0, 1\}$$

- Examples in social science: **numerous!**
  - Turnout ($1 =$ vote; $0 =$ abstain)
  - Education ($1 =$ completed hs; $0 =$ dropped out)
  - Conflict ($1 =$ civil war; $0 =$ no civil war)

Binary Outcomes

- Binary outcome variable:

$$Y_i \in \{0, 1\}$$

- Examples in social science: **numerous!**
  - Turnout ($1 =$ vote; $0 =$ abstain)
  - Education ($1 =$ completed hs; $0 =$ dropped out)
  - Conflict ($1 =$ civil war; $0 =$ no civil war)
  - Eviction ($1 =$ evicted; $0 =$ not evicted)

Binary Outcomes

- Binary outcome variable:

$$Y_i \in \{0, 1\}$$

- Examples in social science: **numerous!**
  - Turnout ($1 =$ vote; $0 =$ abstain)
  - Education ($1 =$ completed hs; $0 =$ dropped out)
  - Conflict ($1 =$ civil war; $0 =$ no civil war)
  - Eviction ($1 =$ evicted; $0 =$ not evicted)
  - etc. etc.

How Do We Model a Binary Outcome Variable?

- For a continuous outcome variable, we model the ____________________ function with predictors X_i:

How Do We Model a Binary Outcome Variable?

- For a continuous outcome variable, we model the conditional expectation function with predictors X_i:

$$\mathbb{E}(Y_i \mid X_i) \ = \ X_i^\top \beta \quad \text{(linear regression)}$$

How Do We Model a Binary Outcome Variable?

- For a continuous outcome variable, we model the conditional expectation function with predictors X_i:

$$\mathbb{E}(Y_i \mid X_i) \;=\; X_i^\top \beta \quad \text{(linear regression)}$$

- When Y_i is binary, $\mathbb{E}(Y_i \mid X_i) =$

How Do We Model a Binary Outcome Variable?

- For a continuous outcome variable, we model the conditional expectation function with predictors X_i:

$$\mathbb{E}(Y_i \mid X_i) \; = \; X_i^\top \beta \quad \text{(linear regression)}$$

- When Y_i is binary, $\mathbb{E}(Y_i \mid X_i) = \Pr(Y_i = 1 \mid X_i)$

How Do We Model a Binary Outcome Variable?

- For a continuous outcome variable, we model the conditional expectation function with predictors X_i:

$$\mathbb{E}(Y_i \mid X_i) \;=\; X_i^\top \beta \quad \text{(linear regression)}$$

- When Y_i is binary, $\mathbb{E}(Y_i \mid X_i) = \Pr(Y_i = 1 \mid X_i)$
- Thus, we model the conditional probability of $Y_i = 1$:

$$\Pr(Y_i = 1 \mid X_i) \;=\; g(X_i^\top \beta)$$

How Do We Model a Binary Outcome Variable?

- For a continuous outcome variable, we model the conditional expectation function with predictors X_i:

$$\mathbb{E}(Y_i \mid X_i) \;=\; X_i^\top \beta \quad \text{(linear regression)}$$

- When Y_i is binary, $\mathbb{E}(Y_i \mid X_i) = \Pr(Y_i = 1 \mid X_i)$
- Thus, we model the conditional probability of $Y_i = 1$:

$$\Pr(Y_i = 1 \mid X_i) \;=\; g(X_i^\top \beta)$$

- There are many possible binary outcome models, depending on the choice of $g(\cdot)$

Linear Probability Model (LPM)

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \;=\; X_i^\top \beta$$

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \; = \; X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim \textit{Bernoulli}(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \;=\; X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim Bernoulli(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

- Advantages:

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \; = \; X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim \mathit{Bernoulli}(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

- Advantages:
  - Easy to estimate: Regress Y_i on X_i

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \;=\; X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim Bernoulli(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

- Advantages:
  - Easy to estimate: Regress Y_i on X_i
  - Easy to interpret: $\beta = ATE$ if $X_i \in \{0, 1\}$ and exogenous

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \;=\; X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim Bernoulli(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

- Advantages:
  - Easy to estimate: Regress Y_i on X_i
  - Easy to interpret: $\beta = ATE$ if $X_i \in \{0, 1\}$ and exogenous

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \ = \ X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim Bernoulli(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

- Advantages:
  - Easy to estimate: Regress Y_i on X_i
  - Easy to interpret: $\beta = ATE$ if $X_i \in \{0, 1\}$ and exogenous
- Disadvantages:

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \;=\; X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim \textit{Bernoulli}(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

- Advantages:
  - Easy to estimate: Regress Y_i on X_i
  - Easy to interpret: $\beta = ATE$ if $X_i \in \{0, 1\}$ and exogenous
- Disadvantages:
  - Estimated probability can go outside of $[0, 1]$

Linear Probability Model (LPM)

- The simplest choice: $g(X_i^\top \beta) = X_i^\top \beta$
- This gives the linear probability model (LPM):

$$\Pr(Y_i = 1 \mid X_i) \ = \ X_i^\top \beta$$

or equivalently

$$\begin{aligned} Y_i &\sim \mathit{Bernoulli}(\pi_i) \\ \pi_i &= X_i^\top \beta \end{aligned}$$

- Advantages:
  - Easy to estimate: Regress Y_i on X_i
  - Easy to interpret: $\beta = ATE$ if $X_i \in \{0, 1\}$ and exogenous
- Disadvantages:
  - Estimated probability can go outside of $[0, 1]$
  - Always heteroskedastic

Logit and Probit Models

- Want: $0 \leq \ g(X_i^\top \beta) \leq 1$ for any X_i

Logit and Probit Models

- Want: $0 \leq \; g(X_i^\top \beta) \leq 1$ for any X_i
- Solution: Use a CDF

$$\pi_i \;=\; g(X_i^\top \beta) \;=\; F(X_i^\top \beta)$$

Logit and Probit Models

- Want: $0 \leq g(X_i^\top \beta) \leq 1$ for any X_i
- Solution: Use a CDF

$$\pi_i = g(X_i^\top \beta) = F(X_i^\top \beta)$$

Note: F is *not* the CDF of Y_i, which is ________

Logit and Probit Models

- Want: $0 \leq \; g(X_i^\top \beta) \leq 1$ for any X_i
- Solution: Use a CDF

$$\pi_i \;=\; g(X_i^\top \beta) \;=\; F(X_i^\top \beta)$$

Note: F is *not* the CDF of Y_i, which is Bernoulli

Logit and Probit Models

- Want: $0 \leq \; g(X_i^\top \beta) \leq 1$ for any X_i
- Solution: Use a CDF

$$\pi_i \;=\; g(X_i^\top \beta) \;=\; F(X_i^\top \beta)$$

Note: F is *not* the CDF of Y_i, which is Bernoulli
(using a CDF is just a convenient way to ensure $0 \leq \pi_i \leq 1$)

Logit and Probit Models

- Want: $0 \leq \; g(X_i^\top \beta) \leq 1$ for any X_i
- Solution: Use a CDF

$$\pi_i \; = \; g(X_i^\top \beta) \; = \; F(X_i^\top \beta)$$

Note: F is *not* the CDF of Y_i, which is Bernoulli
(using a CDF is just a convenient way to ensure $0 \leq \pi_i \leq 1$)

- Logit: Logistic CDF (a.k.a. inverse logit function)

$$\pi_i \; = \; \text{logit}^{-1}(X_i^\top \beta) \; \equiv \; \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)} \; = \; \frac{1}{1 + \exp(-X_i^\top \beta)}$$

Logit and Probit Models

- Want: $0 \leq \ g(X_i^\top \beta) \leq 1$ for any X_i
- Solution: Use a CDF

$$\pi_i \ = \ g(X_i^\top \beta) \ = \ F(X_i^\top \beta)$$

Note: F is *not* the CDF of Y_i, which is Bernoulli
(using a CDF is just a convenient way to ensure $0 \leq \pi_i \leq 1$)

- Logit: Logistic CDF (a.k.a. inverse logit function)

$$\pi_i \ = \ \text{logit}^{-1}(X_i^\top \beta) \ \equiv \ \frac{\exp(X_i^\top \beta)}{1 + \exp(X_i^\top \beta)} \ = \ \frac{1}{1 + \exp(-X_i^\top \beta)}$$

- Probit: Standard normal CDF

$$\pi_i \ = \ \Phi(X_i^\top \beta)$$

Binary Variable Regression Models

The logistic regression (or "logit") model:

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

$$Y_i \sim Y_{\text{Bern}}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \begin{cases} \pi_i & \text{for } y = 1 \\ 1-\pi_i & \text{for } y = 0 \end{cases}$$

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

$$Y_i \sim Y_{\text{Bern}}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \begin{cases} \pi_i & \text{for } y = 1 \\ 1-\pi_i & \text{for } y = 0 \end{cases}$$

2. Systematic Component:

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

$$Y_i \sim Y_{\text{Bern}}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \begin{cases} \pi_i & \text{for } y = 1 \\ 1-\pi_i & \text{for } y = 0 \end{cases}$$

2. Systematic Component:

$$\Pr(Y_i = 1|\beta) \equiv E(Y_i) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$$

Binary Variable Regression Models

The logistic regression (or "logit") model:

1. Stochastic component:

$$Y_i \sim Y_{\text{Bern}}(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i} = \begin{cases} \pi_i & \text{for } y = 1 \\ 1-\pi_i & \text{for } y = 0 \end{cases}$$

2. Systematic Component:

$$\Pr(Y_i = 1|\beta) \equiv E(Y_i) \equiv \pi_i = \frac{1}{1+e^{-x_i\beta}}$$

3. Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X

Binary Variable Regression Models

The probability density of all the data:

Binary Variable Regression Models

The probability density of all the data:

$$\mathbb{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

Binary Variable Regression Models

The probability density of all the data:

$$\mathbb{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

Binary Variable Regression Models

The probability density of all the data:

$$\mathbb{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \{y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\}$$

Binary Variable Regression Models

The probability density of all the data:

$$\mathbb{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

$$\begin{aligned}
\ln L(\pi|y) &= \sum_{i=1}^{n} \{y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \left\{-y_i \ln\left(1+e^{-x_i\beta}\right) + (1-y_i)\ln\left(1-\frac{1}{1+e^{-x_i\beta}}\right)\right\}
\end{aligned}$$

Binary Variable Regression Models

The probability density of all the data:

$$\mathbb{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

$$\begin{aligned}
\ln L(\pi|y) &= \sum_{i=1}^{n} \{y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \left\{-y_i \ln\left(1+e^{-x_i\beta}\right) + (1-y_i)\ln\left(1-\frac{1}{1+e^{-x_i\beta}}\right)\right\} \\
&= -\sum_{i=1}^{n} \ln\left(1+e^{(1-2y_i)x_i\beta}\right).
\end{aligned}$$

Binary Variable Regression Models

The probability density of all the data:

$$\mathbb{P}(y|\pi) = \prod_{i=1}^{n} \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

The log-likelihood:

$$\begin{aligned}
\ln L(\pi|y) &= \sum_{i=1}^{n} \{y_i \ln \pi_i + (1-y_i)\ln(1-\pi_i)\} \\
&= \sum_{i=1}^{n} \left\{-y_i \ln\left(1+e^{-x_i\beta}\right) + (1-y_i)\ln\left(1-\frac{1}{1+e^{-x_i\beta}}\right)\right\} \\
&= -\sum_{i=1}^{n} \ln\left(1+e^{(1-2y_i)x_i\beta}\right).
\end{aligned}$$

What do we do with this?

Interpreting Functional Forms

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i \beta}}$$

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i \beta}}$$

Methods:

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i \beta}}$$

Methods:

1. Graphs.

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i \beta}}$$

Methods:

1. Graphs.
   (a) Can use desired instead of observed X's

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

1. Graphs.
    (a) Can use desired instead of observed X's
    (b) Can try entire surface plot for a small number of X's

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i \beta}}$$

Methods:

1. Graphs.
   (a) Can use desired instead of observed X's
   (b) Can try entire surface plot for a small number of X's
   (c) Marginal effects: Can hold "other variables" constant at their means, a typical value, or at their observed values

Interpreting Functional Forms

Running Example is logit:

$$\pi_i = \frac{1}{1 + e^{-x_i\beta}}$$

Methods:

1. Graphs.
   (a) Can use desired instead of observed X's
   (b) Can try entire surface plot for a small number of X's
   (c) Marginal effects: Can hold "other variables" constant at their means, a typical value, or at their observed values
   (d) Average effects: compute effects for every observation and average

Interpreting Functional Forms

2. Fitted Values for selected combinations of X's, or "typical" people or types:

Interpreting Functional Forms

2. Fitted Values for selected combinations of X's, or "typical" people or types:

Sex	Age	Home	Income	Pr(vote)
Male	20	Chicago	\$33,000	0.20
Female	27	New York City	\$43,000	0.28
Male	50	Madison, WI	\$55,000	0.72
⋮				

Interpreting Functional Forms

2. Fitted Values for selected combinations of X's, or "typical" people or types:

Sex	Age	Home	Income	Pr(vote)
Male	20	Chicago	\$33,000	0.20
Female	27	New York City	\$43,000	0.28
Male	50	Madison, WI	\$55,000	0.72
⋮				

We may also want to include uncertainty (fundamental and estimation uncertainty)

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
   (a) Define X_s ($\underline{s}$tarting point) and X_e ($\underline{e}$nding point) as $k \times 1$ vectors of values of X. Usually all values are the same but one.

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
   (a) Define X_s (s̲tarting point) and X_e (e̲nding point) as $k \times 1$ vectors of values of X. Usually all values are the same but one.
   (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
   (a) Define X_s (starting point) and X_e (ending point) as $k \times 1$ vectors of values of X. Usually all values are the same but one.
   (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$
   (c) $D = \frac{1}{1+e^{-X_e\hat{\beta}}} - \frac{1}{1+e^{-X_s\hat{\beta}}}$

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
   (a) Define X_s (s̲tarting point) and X_e (e̲nding point) as $k \times 1$ vectors of values of X. Usually all values are the same but one.
   (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$
   (c) $D = \frac{1}{1+e^{-X_e \hat{\beta}}} - \frac{1}{1+e^{-X_s \hat{\beta}}}$
   (d) Better (and necessary to compute se's): do by simulation (we'll repeat the details soon)

Interpreting Functional Forms

3. First Differences (called Risk Differences in epidemiology)
   (a) Define X_s (starting point) and X_e (ending point) as $k \times 1$ vectors of values of X. Usually all values are the same but one.
   (b) First difference $= g(X_e, \hat{\beta}) - g(X_s, \hat{\beta})$
   (c) $D = \frac{1}{1+e^{-X_e \hat{\beta}}} - \frac{1}{1+e^{-X_s \hat{\beta}}}$
   (d) Better (and necessary to compute se's): do by simulation (we'll repeat the details soon)

Variable	From		To	*FirstDifference*
Sex	Male	$\rightarrow$	Female	.05
Age	65	$\rightarrow$	75	$-.10$
Home	NYC	$\rightarrow$	Madison, WI	.26
Income	\$35,000	$\rightarrow$	\$75,000	.14

Interpreting Functional Forms

4. Derivatives (i.e. a source of heuristics for talks)

Interpreting Functional Forms

4. Derivatives (i.e. a source of heuristics for talks)

$$\frac{\partial \pi_i}{\partial X_j} = \frac{\partial \frac{1}{1+e^{-X\beta}}}{\partial X_j} = \hat{\beta}_j \hat{\pi}_i (1 - \hat{\pi}_i)$$

Interpreting Functional Forms

4. Derivatives (i.e. a source of heuristics for talks)

$$\frac{\partial \pi_i}{\partial X_j} = \frac{\partial \frac{1}{1+e^{-X\beta}}}{\partial X_j} = \hat{\beta}_j \hat{\pi}_i (1 - \hat{\pi}_i)$$

   (a) Max value of logit derivative: $\hat{\beta} \times 0.5(1 - 0.5) = \hat{\beta}/4$

Interpreting Functional Forms

4. Derivatives (i.e. a source of heuristics for talks)

$$\frac{\partial \pi_i}{\partial X_j} = \frac{\partial \frac{1}{1+e^{-X\beta}}}{\partial X_j} = \hat{\beta}_j \hat{\pi}_i (1 - \hat{\pi}_i)$$

   (a) Max value of logit derivative: $\hat{\beta} \times 0.5(1-0.5) = \hat{\beta}/4$

   (b) Max value for probit $[\pi_i = \Phi(X_i\beta)]$ derivative: $\hat{\beta} \times 0.4$

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*
- Let

$$
\begin{aligned}
Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\
Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{where} \quad \mathbb{E}(\epsilon_i) = 0
\end{aligned}
$$

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*
- Let

$$\begin{aligned} Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\ Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{where} \quad \mathbb{E}(\epsilon_i) = 0 \end{aligned}$$

- This is also called a random utility model, where

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*
- Let

$$\begin{aligned} Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\ Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{where} \quad \mathbb{E}(\epsilon_i) = 0 \end{aligned}$$

- This is also called a random utility model, where
  - Y_i^*: Utility from choosing $Y_i = 1$ instead of $Y_i = 0$

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*
- Let

$$
\begin{aligned}
Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\
Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{where} \quad \mathbb{E}(\epsilon_i) = 0
\end{aligned}
$$

- This is also called a random utility model, where
  - Y_i^*: Utility from choosing $Y_i = 1$ instead of $Y_i = 0$
  - $X_i^\top \beta$: Systematic component of utility

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*
- Let

$$
\begin{aligned}
Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\
Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{where} \quad \mathbb{E}(\epsilon_i) = 0
\end{aligned}
$$

- This is also called a random utility model, where
  - Y_i^*: Utility from choosing $Y_i = 1$ instead of $Y_i = 0$
  - $X_i^\top \beta$: Systematic component of utility
  - ϵ_i: Stochastic (random) component of utility

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*
- Let

$$
\begin{aligned}
Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\
Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{where} \quad \mathbb{E}(\epsilon_i) = 0
\end{aligned}
$$

- This is also called a random utility model, where
  - Y_i^*: Utility from choosing $Y_i = 1$ instead of $Y_i = 0$
  - $X_i^\top \beta$: Systematic component of utility
  - ϵ_i: Stochastic (random) component of utility
- Make distributional assumptions about ϵ_i:

Latent Variable Interpretation

- Logit models can also be interpreted in terms of a latent variable Y_i^*
- Let

$$\begin{aligned} Y_i &= \begin{cases} 1 & \text{if } Y_i^* > 0 \\ 0 & \text{if } Y_i^* \leq 0 \end{cases} \\ Y_i^* &= X_i^\top \beta + \epsilon_i \quad \text{where} \quad \mathbb{E}(\epsilon_i) = 0 \end{aligned}$$

- This is also called a random utility model, where
  - Y_i^*: Utility from choosing $Y_i = 1$ instead of $Y_i = 0$
  - $X_i^\top \beta$: Systematic component of utility
  - ϵ_i: Stochastic (random) component of utility
- Make distributional assumptions about ϵ_i:
  - $\epsilon_i \overset{\text{i.i.d.}}{\sim}$ Logistic $\implies$ Logit

Latent Variable: Walkthrough

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$Y_i^* \sim P(y_i^* | \mu_i)$$

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$Y_i^* \sim P(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$\begin{aligned} Y_i^* &\sim P(y_i^*|\mu_i) \\ \mu_i &= x_i\beta \end{aligned}$$

with observation mechanism:

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$
\begin{aligned}
Y_i^* &\sim P(y_i^*|\mu_i) \\
\mu_i &= x_i\beta
\end{aligned}
$$

with observation mechanism:

$$y_i =$$

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$\begin{aligned} Y_i^* &\sim P(y_i^*|\mu_i) \\ \mu_i &= x_i\beta \end{aligned}$$

with observation mechanism:

$$y_i = \begin{cases} 1 & y^* \leq \tau \text{ if } i \text{ is alive} \\ 0 & y^* > \tau \text{ if } i \text{ is dead} \end{cases}$$

Latent Variable: Walkthrough

Let Y^* be a continuous unobserved variable. Health, propensity to vote, etc.

A model:

$$\begin{aligned} Y_i^* &\sim P(y_i^*|\mu_i) \\ \mu_i &= x_i\beta \end{aligned}$$

with observation mechanism:

$$y_i = \begin{cases} 1 & y^* \leq \tau \text{ if } i \text{ is alive} \\ 0 & y^* > \tau \text{ if } i \text{ is dead} \end{cases}$$

Since Y^* is unobserved anyway, define the threshold as $\tau = 0$. (Plus the same independence assumption, which from now on is assumed implicit.)

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.
2. If only y_i is observed, and Y^* is standardized logistic (which looks close to the normal),

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.
2. If only y_i is observed, and Y^* is standardized logistic (which looks close to the normal),

$$P(y_i^*|\mu_i) = \text{STL}(y^*|\mu_i) = \frac{\exp(y_i^* - \mu_i)}{[1 + \exp(y_i^* - \mu_i)]^2}$$

Same Logit Model, Different Justification and Interpretation

1. If Y^* is observed and $P(\cdot)$ is normal, this is a regression.
2. If only y_i is observed, and Y^* is standardized logistic (which looks close to the normal),

$$P(y_i^*|\mu_i) = \text{STL}(y^*|\mu_i) = \frac{\exp(y_i^* - \mu_i)}{[1 + \exp(y_i^* - \mu_i)]^2}$$

then we get a logit model.

Same Logit Model, Different Justification and Interpretation

3. The derivation:

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\Pr(Y_i = 1|\mu_i) = \Pr(Y_i^* \leq 0)$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\begin{aligned}\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0) \\ &= \int_{-\infty}^{0} \text{STL}(y_i^*|\mu_i) dy_i^*\end{aligned}$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$
\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0) \\
&= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i) dy_i^* \\
&= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]}
\end{aligned}
$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0) \\
&= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i) dy_i^* \\
&= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]} \\
&= [1 + \exp(-X_i\beta)]^{-1}
\end{aligned}$$

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\begin{aligned}
\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0) \\
&= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i)dy_i^* \\
&= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]} \\
&= [1 + \exp(-X_i\beta)]^{-1}
\end{aligned}$$

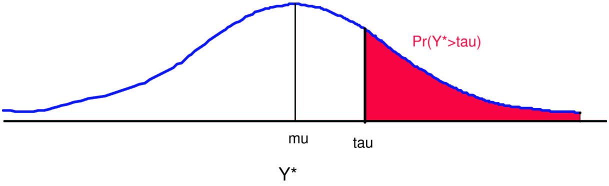
The same functional form!

Same Logit Model, Different Justification and Interpretation

3. The derivation:

$$\begin{aligned}\Pr(Y_i = 1|\mu_i) &= \Pr(Y_i^* \leq 0)\\ &= \int_{-\infty}^{0} \mathrm{STL}(y_i^*|\mu_i)dy_i^* \\ &= F_{stl}(0|\mu_i) \qquad \text{[the CDF of the STL]}\\ &= [1+\exp(-X_i\beta)]^{-1}\end{aligned}$$

The same functional form!



The Probit Model

4. For the Probit Model, we modify:

The Probit Model

4. For the Probit Model, we modify:

$$\mathbb{P}(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

The Probit Model

4. For the Probit Model, we modify:

$$\mathbb{P}(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

with the same observation mechanism, implying

The Probit Model

4. For the Probit Model, we modify:

$$\mathbb{P}(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

with the same observation mechanism, implying

$$\Pr(Y_i = 1|\mu) = \int_{-\infty}^{0} N(y_i^*|\mu_i, 1) dy_i^* = \Phi(X_i\beta)$$

The Probit Model

4. For the Probit Model, we modify:

$$\mathbb{P}(y_i^*|\mu_i) = N(y_i^*|\mu_i, 1)$$

with the same observation mechanism, implying

$$\Pr(Y_i = 1|\mu) = \int_{-\infty}^{0} N(y_i^*|\mu_i, 1)dy_i^* = \Phi(X_i\beta)$$

5. $\implies$ interpret β as regression coefficients of Y^* on X: $\hat{\beta}_1$ is what happens to Y^* on average (or μ_i) when X_1 goes up by one unit, holding constant the other explanatory variables (and conditional on the model). In probit, one unit of Y^* is one standard deviation.

An Econometric Interpretation: Utility Maximization

An Econometric Interpretation: Utility Maximization

- Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.

An Econometric Interpretation: Utility Maximization

- Let U_i^D be the utility for the Democratic candidate;
  and U_i^R be the utility for the Republican candidate.
- Assume U_i^D and U_i^R are independent

An Econometric Interpretation: Utility Maximization

- Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
- Assume U_i^D and U_i^R are independent
- Assume $U_i^k \sim \mathbb{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.

An Econometric Interpretation: Utility Maximization

- Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
- Assume U_i^D and U_i^R are independent
- Assume $U_i^k \sim \mathbb{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
- Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.

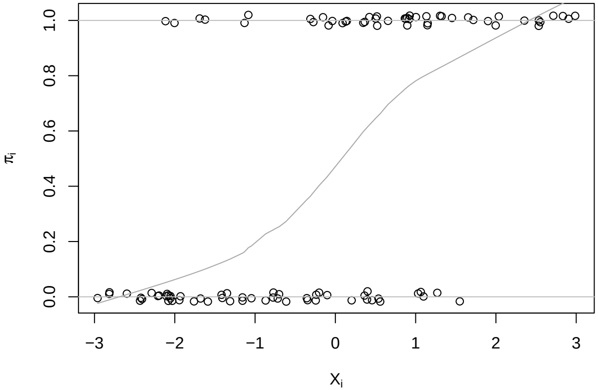
An Econometric Interpretation: Utility Maximization

- Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
- Assume U_i^D and U_i^R are independent
- Assume $U_i^k \sim \mathbb{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
- Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.
- If $\mathbb{P}(\cdot)$ is normal, we get a Probit model

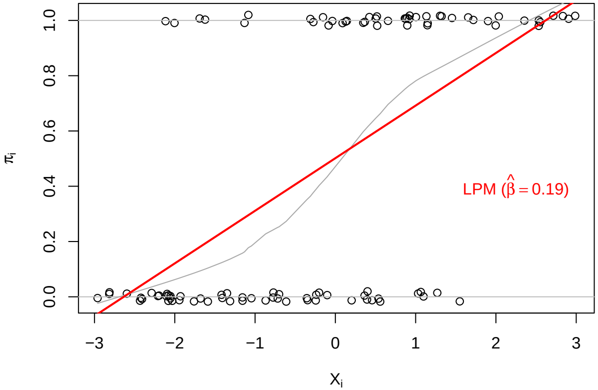
An Econometric Interpretation: Utility Maximization

- Let U_i^D be the utility for the Democratic candidate; and U_i^R be the utility for the Republican candidate.
- Assume U_i^D and U_i^R are independent
- Assume $U_i^k \sim \mathbb{P}(U_i^k|\eta_i^k)$ for $k = \{D, R\}$.
- Let $Y^* \equiv U_i^D - U_i^R$ and apply the same interpretation as above: If $y^* > 0$, choose the Democrat, otherwise, choose the Republican.
- If $\mathbb{P}(\cdot)$ is normal, we get a Probit model
- If $\mathbb{P}(\cdot)$ is generalized extreme value, we get logit.

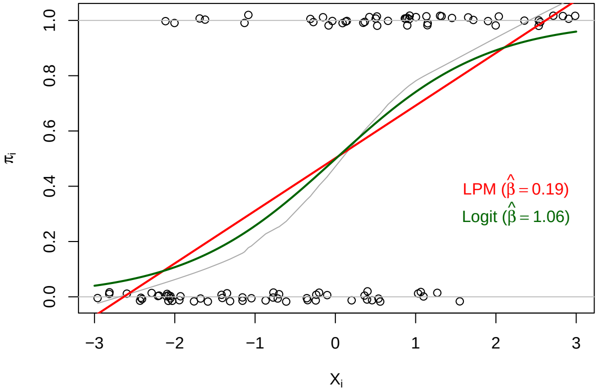
Comparing LPM, Logit, and Probit



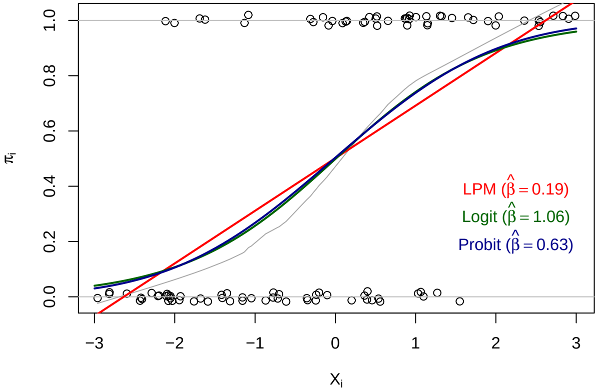
Comparing LPM, Logit, and Probit



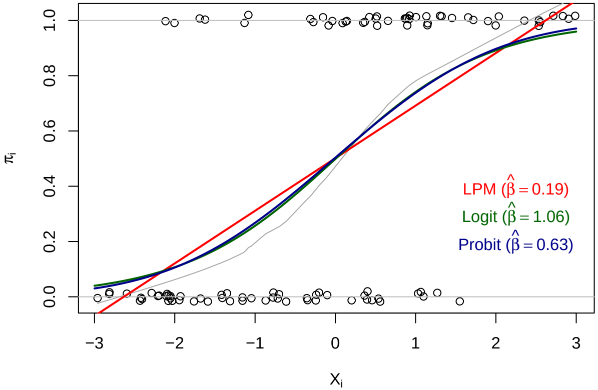
Comparing LPM, Logit, and Probit



Comparing LPM, Logit, and Probit



Comparing LPM, Logit, and Probit



- LPM goes outside of $[0, 1]$ for extreme values of X_i
- LPM underestimates the marginal effect near center and overpredicts near extremes
- Logit has *slightly* fatter tails than probit, but no practical difference
- Note that $\hat{\beta}$ are completely different between the models

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
*$p \leq .05$, two-tailed. **$p \leq .01$, two-tailed.

How Not to Present Statistical Results

TABLE 1

Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.

NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.

$^*p \leq .05$, two-tailed. $^{**}p \leq .01$, two-tailed.

1. This one is typical of current practice, not that unusual.

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
*$p \leq .05$, two-tailed. **$p \leq .01$, two-tailed.

1. This one is typical of current practice, not that unusual.
2. What do these numbers mean?

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
*$p \leq .05$, two-tailed. **$p \leq .01$, two-tailed.

1. This one is typical of current practice, not that unusual.
2. What do these numbers mean?
3. Why so much whitespace? Can you connect cols A and B?

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
$*p \leq .05$, two-tailed. $**p \leq .01$, two-tailed.

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
*$p \leq .05$, two-tailed. **$p \leq .01$, two-tailed.

4. What does the star-gazing add?

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
*$p \leq .05$, two-tailed. **$p \leq .01$, two-tailed.

4. What does the star-gazing add?
5. Can any be interpreted as causal estimates?

How Not to Present Statistical Results

TABLE 1
Predicting Which Ethnic Group Conquered Most of Bosnia

Attention to Bosnia crisis	.609**
Age	.007**
Education	.289**
Family income	.151**
Race (non-White/White)	.695**
Gender (female/male)	.789**
Region (South/non-South)	.076
Network coverage	.000
Education × Time	–.003*
Time in months	.078**
Constant	–9.257**
Number	7,021
–2 log-likelihood	7,215.231
Goodness of fit	6,789.45
Cox & Snell R^2	.212
Nagelkerke R^2	.295
Overall correct classification (%)	73.96

SOURCE: *Times Mirror* polls from September 1992, January 1993, September 1993, January 1994, and June 1995.
NOTE: Unstandardized coefficients for logistic regression. Dependent variable is knowledge of which group conquered most of Bosnia.
*$p \leq .05$, two-tailed. **$p \leq .01$, two-tailed.

4. What does the star-gazing add?
5. Can any be interpreted as causal estimates?
6. Can you compute a quantity of interest from these numbers?

Interpretation and Presentation

Interpretation and Presentation

1. Statistical presentations should

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.
   (d) Include no superfluous information, long lists of coefficients no one understands, star gazing, etc.

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.
   (d) Include no superfluous information, long lists of coefficients no one understands, star gazing, etc.
2. For example: Other things being equal, an additional year of education would increase your annual income by $1,500 on average, plus or minus about $500.

Interpretation and Presentation

1. Statistical presentations should
   (a) Convey numerically precise estimates of the quantities of substantive interest,
   (b) Include reasonable measures of uncertainty about those estimates,
   (c) Require little specialized knowledge to understand.
   (d) Include no superfluous information, long lists of coefficients no one understands, star gazing, etc.
2. For example: Other things being equal, an additional year of education would increase your annual income by $1,500 on average, plus or minus about $500.
3. Your work should satisfy a reader who hasn't taken this course

Reading

- King, Tomz, Wittenberg, "Making the Most of Statistical Analyses: Improving Interpretation and Presentation" *American Journal of Political Science*, Vol. 44, No. 2 (March, 2000): 341-355.
- Hamner and Kalkan (2013). Behind the Curve: Clarifying the Best Approach to Calculating Predicted Probabilities and Marginal Effects from Limited Dependent Variable Models. *American Journal of Political Science*.
- Greenhill, Ward, and Sacks (2011). The Separation Plot: A new visual method for evaluating the fit of binary models. *American Journal of Political Science*.

Quantities of Interest

- How to interpret β in binary outcome models?
  - In LPM, β is the marginal effect ($\beta_j = \partial \mathbb{E}[Y_i \mid X] / \partial X_{ij}$)

Quantities of Interest

- How to interpret β in binary outcome models?
  - In LPM, β is the marginal effect ($\beta_j = \partial \mathbb{E}[Y_i \mid X] / \partial X_{ij}$)
  - In logit, β is the log odds ratio

$$\beta_j = \log \left\{ \frac{\Pr(Y_i = 1 \mid X_{ij} = 1) / \Pr(Y_i = 0 \mid X_{ij} = 1)}{\Pr(Y_i = 1 \mid X_{ij} = 0) / \Pr(Y_i = 0 \mid X_{ij} = 0)} \right\}$$

Quantities of Interest

- How to interpret β in binary outcome models?
  - In LPM, β is the marginal effect ($\beta_j = \partial \mathbb{E}[Y_i \mid X] / \partial X_{ij}$)
  - In logit, β is the log odds ratio

$$\beta_j = \log \left\{ \frac{\Pr(Y_i = 1 \mid X_{ij} = 1) / \Pr(Y_i = 0 \mid X_{ij} = 1)}{\Pr(Y_i = 1 \mid X_{ij} = 0) / \Pr(Y_i = 0 \mid X_{ij} = 0)} \right\}$$

  - In probit, no direct substantive interpretation of β

Quantities of Interest

- How to interpret β in binary outcome models?
  - In LPM, β is the marginal effect ($\beta_j = \partial \mathbb{E}[Y_i \mid X] / \partial X_{ij}$)
  - In logit, β is the log odds ratio

$$\beta_j = \log \left\{ \frac{\Pr(Y_i = 1 \mid X_{ij} = 1) / \Pr(Y_i = 0 \mid X_{ij} = 1)}{\Pr(Y_i = 1 \mid X_{ij} = 0) / \Pr(Y_i = 0 \mid X_{ij} = 0)} \right\}$$

  - In probit, no direct substantive interpretation of β
  - In general, it is a bad practice to just present a coefficients table!

Quantities of Interest

- How to interpret β in binary outcome models?
  - In LPM, β is the marginal effect ($\beta_j = \partial \mathbb{E}[Y_i \mid X] / \partial X_{ij}$)
  - In logit, β is the log odds ratio

$$\beta_j = \log \left\{ \frac{\Pr(Y_i = 1 \mid X_{ij} = 1) / \Pr(Y_i = 0 \mid X_{ij} = 1)}{\Pr(Y_i = 1 \mid X_{ij} = 0) / \Pr(Y_i = 0 \mid X_{ij} = 0)} \right\}$$

  - In probit, no direct substantive interpretation of β
  - In general, it is a bad practice to just present a coefficients table!
  - Instead, always try to present your results in terms of an easy-to-interpret quantity

Analytic Quantities of Interest

Analytic Quantities of Interest

1. Predicted probability when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \;=\; \pi(x)$$

Analytic Quantities of Interest

1. Predicted probability when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \;=\; \pi(x)$$

2. Average Treatment Effect (ATE):
   - Given the model, ϵ_i is i.i.d.
     $\Longrightarrow T_i$ is conditionally ignorable given W_i

Analytic Quantities of Interest

1. Predicted probability when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \ = \ \pi(x)$$

2. Average Treatment Effect (ATE):
   - Given the model, ϵ_i is i.i.d.
     $\Longrightarrow T_i$ is conditionally ignorable given W_i
   - Thus, ATE can be identified as

$$\begin{aligned}\tau &= \mathbb{E}\left[\Pr(Y_i = 1 \mid T_i = 1, W_i) - \Pr(Y_i = 1 \mid T_i = 0, W_i)\right] \\ &= \mathbb{E}\left[\pi(T_i = 1, W_i) - \pi(T_i = 0, W_i)\right]\end{aligned}$$

   where $\mathbb{E}$ is taken with respect to both ϵ and W

Analytic Quantities of Interest

1. Predicted probability when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \;=\; \pi(x)$$

2. Average Treatment Effect (ATE):
   - Given the model, ϵ_i is i.i.d.
     $\Longrightarrow T_i$ is conditionally ignorable given W_i
   - Thus, ATE can be identified as

$$\begin{aligned}\tau &= \mathbb{E}\left[\Pr(Y_i = 1 \mid T_i = 1, W_i) - \Pr(Y_i = 1 \mid T_i = 0, W_i)\right] \\ &= \mathbb{E}\left[\pi(T_i = 1, W_i) - \pi(T_i = 0, W_i)\right]\end{aligned}$$

   where $\mathbb{E}$ is taken with respect to both ϵ and W

3. Marginal effects: For a continuous predictor X_{ij},

$$\frac{\partial \mathbb{E}[Y_i \mid X]}{\partial X_{ij}} = \begin{cases} \beta_j \cdot \text{logit}^{-1}(X_i^\top \beta)\left(1 - \text{logit}^{-1}(X_i^\top \beta)\right) & \text{(for logit)} \\ \beta_j \cdot \phi(X_i^\top \beta) & \text{(for probit)} \end{cases}$$

Analytic Quantities of Interest

1. Predicted probability when $X_i = x$:

$$\Pr(Y_i = 1 \mid X_i = x) \;=\; \pi(x)$$

2. Average Treatment Effect (ATE):
   - Given the model, ϵ_i is i.i.d. $\implies T_i$ is conditionally ignorable given W_i
   - Thus, ATE can be identified as

$$\begin{aligned}\tau &= \mathbb{E}\left[\Pr(Y_i = 1 \mid T_i = 1, W_i) - \Pr(Y_i = 1 \mid T_i = 0, W_i)\right] \\ &= \mathbb{E}\left[\pi(T_i = 1, W_i) - \pi(T_i = 0, W_i)\right]\end{aligned}$$

   where $\mathbb{E}$ is taken with respect to both ϵ and W

3. Marginal effects: For a continuous predictor X_{ij},

$$\frac{\partial \mathbb{E}[Y_i \mid X]}{\partial X_{ij}} = \begin{cases} \beta_j \cdot \text{logit}^{-1}(X_i^\top \beta)\left(1 - \text{logit}^{-1}(X_i^\top \beta)\right) & \text{(for logit)} \\ \beta_j \cdot \phi(X_i^\top \beta) & \text{(for probit)} \end{cases}$$

Note: Depends on all X_i, so must pick a particular value

Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- Y_i: Civil conflict
- T_i: Political instability
- W_i: Geography (log % mountainous)

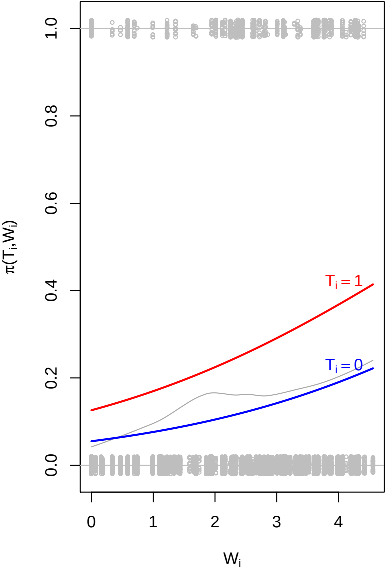
Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- Y_i: Civil conflict
- T_i: Political instability
- W_i: Geography (log % mountainous)

Estimated model:

$$\widehat{\Pr(Y_i = 1 \mid T_i, W_i)} = \text{logit}^{-1}(-2.84 + 0.91\,T_i + 0.35\,W_i)$$



Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

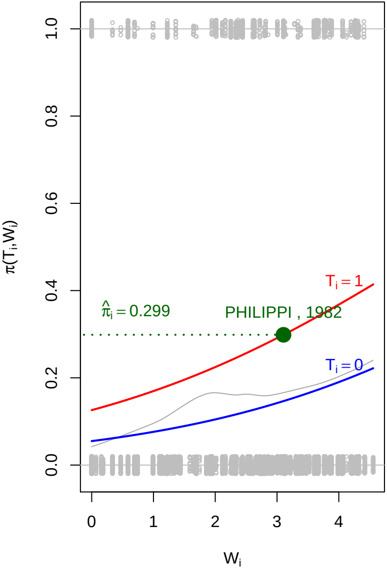
- Y_i: Civil conflict
- T_i: Political instability
- W_i: Geography (log % mountainous)

Estimated model:

$$\widehat{\Pr(Y_i = 1 \mid T_i, W_i)} = \text{logit}^{-1}(-2.84 + 0.91\,T_i + 0.35\,W_i)$$

Predicted probability:

$$\hat{\pi}(T_i = 1, W_i = 3.10) = 0.299$$



Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- Y_i: Civil conflict
- T_i: Political instability
- W_i: Geography (log % mountainous)

Estimated model:

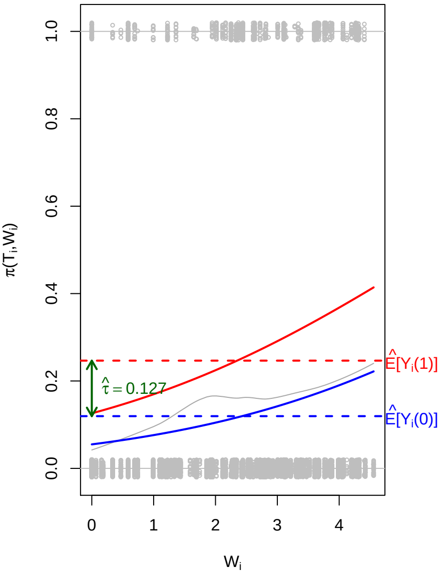
$$\begin{aligned}&\widehat{\Pr(Y_i = 1 \mid T_i, W_i)}\\ &= \text{logit}^{-1}(-2.84 + 0.91\,T_i + 0.35\,W_i)\end{aligned}$$

Predicted probability:

$$\hat{\pi}(T_i = 1, W_i = 3.10) \;=\; 0.299$$

ATE:

$$\begin{aligned}\hat{\tau} &= \frac{1}{n}\sum_{i=1}^{n}\{\hat{\pi}(1, W_i) - \hat{\pi}(0, W_i)\}\\ &= 0.127\end{aligned}$$



Example: Civil Conflict and Political Instability

Fearon & Laitin (2003):

- Y_i: Civil conflict
- T_i: Political instability
- W_i: Geography (log % mountainous)

Estimated model:

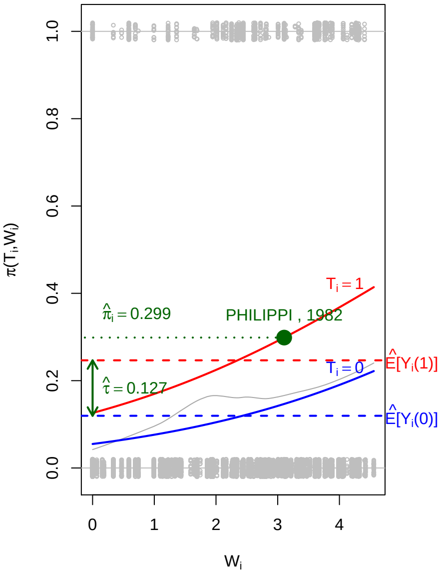
$$\widehat{\Pr(Y_i = 1 \mid T_i, W_i)} \\ = \text{logit}^{-1}(-2.84 + 0.91\,T_i + 0.35\,W_i)$$

Predicted probability:

$$\hat{\pi}(T_i = 1, W_i = 3.10) \; = \; 0.299$$

ATE:

$$\hat{\tau} \; = \; \frac{1}{n}\sum_{i=1}^{n}\{\hat{\pi}(1, W_i) - \hat{\pi}(0, W_i)\} \\ = \; 0.127$$



Variance Estimation

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest
- Predicted probability: $\hat{\pi}(x) = \frac{\exp(x^\top \hat{\beta}_{MLE})}{1+\exp(x^\top \hat{\beta}_{MLE})}$ for logit

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest
- Predicted probability: $\hat{\pi}(x) = \frac{\exp(x^\top \hat{\beta}_{MLE})}{1+\exp(x^\top \hat{\beta}_{MLE})}$ for logit
- ATE: $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)}{1+\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)} - \frac{\exp(W_i^\top \hat{\beta}_W)}{1+\exp(W_i^\top \hat{\beta}_W)} \right)$ for logit

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest
- Predicted probability: $\hat{\pi}(x) = \frac{\exp(x^\top \hat{\beta}_{MLE})}{1+\exp(x^\top \hat{\beta}_{MLE})}$ for logit
- ATE: $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)}{1+\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)} - \frac{\exp(W_i^\top \hat{\beta}_W)}{1+\exp(W_i^\top \hat{\beta}_W)} \right)$ for logit
- How we compute standard errors for quantities like $\hat{\pi}$ and $\hat{\tau}$?

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest
- Predicted probability: $\hat{\pi}(x) = \frac{\exp(x^\top \hat{\beta}_{MLE})}{1+\exp(x^\top \hat{\beta}_{MLE})}$ for logit
- ATE: $\hat{\tau} = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)}{1+\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)} - \frac{\exp(W_i^\top \hat{\beta}_W)}{1+\exp(W_i^\top \hat{\beta}_W)} \right)$ for logit
- How we compute standard errors for quantities like $\hat{\pi}$ and $\hat{\tau}$?
- Three approaches:

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest
- Predicted probability: $\hat{\pi}(x) = \frac{\exp(x^\top \hat{\beta}_{MLE})}{1+\exp(x^\top \hat{\beta}_{MLE})}$ for logit
- ATE: $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)}{1+\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)} - \frac{\exp(W_i^\top \hat{\beta}_W)}{1+\exp(W_i^\top \hat{\beta}_W)} \right)$ for logit
- How we compute standard errors for quantities like $\hat{\pi}$ and $\hat{\tau}$?
- Three approaches:
  1. Analytical approximation: the Delta method

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e., \ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest
- Predicted probability: $\hat{\pi}(x) = \frac{\exp(x^\top \hat{\beta}_{MLE})}{1+\exp(x^\top \hat{\beta}_{MLE})}$ for logit
- ATE: $\hat{\tau} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)}{1+\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)} - \frac{\exp(W_i^\top \hat{\beta}_W)}{1+\exp(W_i^\top \hat{\beta}_W)} \right)$ for logit
- How we compute standard errors for quantities like $\hat{\pi}$ and $\hat{\tau}$?
- Three approaches:
    1. Analytical approximation: the Delta method
    2. Simulating from sampling distributions

Variance Estimation

- The variance estimates can be used for calculating confidence intervals for logit/probit β: $[\hat{\beta}_{MLE} - z_{\alpha/2} \cdot s.e.,\ \hat{\beta}_{MLE} + z_{\alpha/2} \cdot s.e.]$
- But β itself is (typically) not of direct substantive interest
- Predicted probability: $\hat{\pi}(x) = \frac{\exp(x^\top \hat{\beta}_{MLE})}{1+\exp(x^\top \hat{\beta}_{MLE})}$ for logit
- ATE: $\hat{\tau} = \frac{1}{n} \sum_{i=1}^n \left(\frac{\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)}{1+\exp(\hat{\beta}_T + W_i^\top \hat{\beta}_W)} - \frac{\exp(W_i^\top \hat{\beta}_W)}{1+\exp(W_i^\top \hat{\beta}_W)} \right)$ for logit
- How we compute standard errors for quantities like $\hat{\pi}$ and $\hat{\tau}$?
- Three approaches:
  1. Analytical approximation: the Delta method
  2. Simulating from sampling distributions
  3. Resampling: the bootstrap (parametric or nonparametric)

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
- For each draw θ^*, compute $f(\theta^*)$ by plugging in

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
- For each draw θ^*, compute $f(\theta^*)$ by plugging in
- Variance in the distribution of θ^* should transfer to $f(\theta^*)$

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
- For each draw θ^*, compute $f(\theta^*)$ by plugging in
- Variance in the distribution of θ^* should transfer to $f(\theta^*)$
- This leads to the algorithm of King, Tomz and Wittenberg (2000):

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
- For each draw θ^*, compute $f(\theta^*)$ by plugging in
- Variance in the distribution of θ^* should transfer to $f(\theta^*)$
- This leads to the algorithm of King, Tomz and Wittenberg (2000):
  1. Draw R copies of $\hat{\theta}_r$ from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
- For each draw θ^*, compute $f(\theta^*)$ by plugging in
- Variance in the distribution of θ^* should transfer to $f(\theta^*)$
- This leads to the algorithm of King, Tomz and Wittenberg (2000):
  1. Draw R copies of $\hat{\theta}_r$ from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
  2. For each $\hat{\theta}_r$, compute $f(\hat{\theta}_r)$

Monte Carlo Approximation

- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
- For each draw θ^*, compute $f(\theta^*)$ by plugging in
- Variance in the distribution of θ^* should transfer to $f(\theta^*)$
- This leads to the algorithm of King, Tomz and Wittenberg (2000):
  1. Draw R copies of $\hat{\theta}_r$ from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
  2. For each $\hat{\theta}_r$, compute $f(\hat{\theta}_r)$

  3a. To obtain s.e. of $f(\hat{\theta})$, use the sample standard deviation of $\{f(\hat{\theta}_1), ..., f(\hat{\theta}_R)\}$

Monte Carlo Approximation

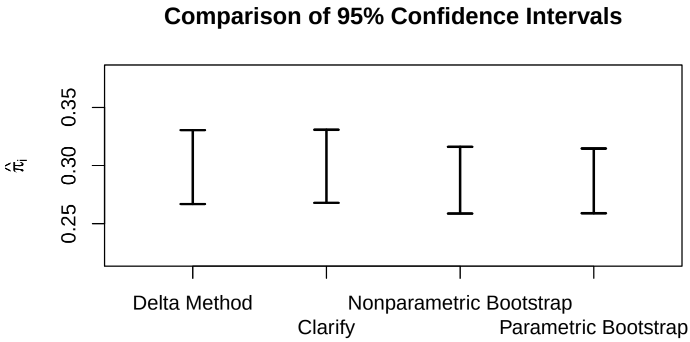
- For MLE, we know that $\hat{\theta} \overset{\text{approx.}}{\sim} \mathcal{N}(\theta, \mathbb{V}(\hat{\theta}))$
- We can simulate this distribution by sampling from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
- For each draw θ^*, compute $f(\theta^*)$ by plugging in
- Variance in the distribution of θ^* should transfer to $f(\theta^*)$
- This leads to the algorithm of King, Tomz and Wittenberg (2000):
  1. Draw R copies of $\hat{\theta}_r$ from $\mathcal{N}(\hat{\theta}, \widehat{\mathbb{V}(\hat{\theta})})$
  2. For each $\hat{\theta}_r$, compute $f(\hat{\theta}_r)$

  3a. To obtain s.e. of $f(\hat{\theta})$, use the sample standard deviation of $\{f(\hat{\theta}_1), ..., f(\hat{\theta}_R)\}$

  3b. To compute 95% CI, use 2.5/97.5 percentiles of $\{f(\hat{\theta}_1), ..., f(\hat{\theta}_R)\}$ as the lower/upper bounds

Example: Civil Conflict and Political Instability

Confidence Intervals for $\hat{\pi}(T_i = 1, W_i = 3.10)$:



	Delta Method	Clarify	Nonpara. B.	Para. B.
Normal Approximation	Yes	Yes	No	No
Simulations	No	Yes	Yes	Yes
Derivation-Free	No	Yes	Yes	No
Computation Speed	Instant	Fast	Very Slow	Slow

Simulation from any model must reflect all uncertainty

Simulation from any model must reflect all uncertainty

$Y_i \sim f(\theta_i, \alpha)$ stochastic

Simulation from any model must reflect all uncertainty

$Y_i \sim f(\theta_i, \alpha)$ stochastic

$\theta_i = g(x_i, \beta)$ systematic

Simulation from any model must reflect all uncertainty

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(x_i, \beta) \qquad \text{systematic}$$

Must simulate anything with uncertainty:

Simulation from any model must reflect all uncertainty

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(x_i, \beta) \qquad \text{systematic}$$

Must simulate anything with uncertainty:

1. Estimation uncertainty: Lack of knowledge of β and α. (Due to inadequacies in your research design: n is not infinite.)

Simulation from any model must reflect all uncertainty

$$Y_i \sim f(\theta_i, \alpha) \qquad \text{stochastic}$$
$$\theta_i = g(x_i, \beta) \qquad \text{systematic}$$

Must simulate anything with uncertainty:

1. Estimation uncertainty: Lack of knowledge of β and α. (Due to inadequacies in your research design: n is not infinite.)
2. Fundamental uncertainty: Represented by the stochastic component. (Due to the nature of nature!)

Strategy for Simulating from Generalized Linear Models

Strategy for Simulating from Generalized Linear Models

All of the models we've talked about so far (and for the next few weeks) belong to the class of **generalized linear models (GLM)**.

Strategy for Simulating from Generalized Linear Models

All of the models we've talked about so far (and for the next few weeks) belong to the class of **generalized linear models (GLM)**.

Three elements of a GLM

Strategy for Simulating from Generalized Linear Models

All of the models we've talked about so far (and for the next few weeks) belong to the class of **generalized linear models (GLM)**.

Three elements of a GLM

- A distribution for Y (stochastic component)

Strategy for Simulating from Generalized Linear Models

All of the models we've talked about so far (and for the next few weeks) belong to the class of **generalized linear models (GLM)**.

Three elements of a GLM

- A distribution for Y (stochastic component)
- A linear predictor $X\beta$ (systematic component)

Strategy for Simulating from Generalized Linear Models

All of the models we've talked about so far (and for the next few weeks) belong to the class of **generalized linear models (GLM)**.

Three elements of a GLM

- A distribution for Y (stochastic component)
- A linear predictor $X\beta$ (systematic component)
- A link function that relates the linear predictor to the mean of the distribution. (systematic component)

Strategy for Simulating from Generalized Linear Models

All of the models we've talked about so far (and for the next few weeks) belong to the class of **generalized linear models (GLM)**.

Three elements of a GLM

- A distribution for Y (stochastic component)
- A linear predictor $X\beta$ (systematic component)
- A link function that relates the linear predictor to the mean of the distribution. (systematic component)

(Note: the language is slightly different for the latent variable with observation mechanism but the result is the same)

Complete Recipe

Complete Recipe

1. Specify a distribution for Y

Complete Recipe

1. Specify a distribution for Y
2. Specify a linear predictor

Complete Recipe

1. Specify a distribution for Y
2. Specify a linear predictor
3. Specify a link function

Complete Recipe

1. Specify a distribution for Y
2. Specify a linear predictor
3. Specify a link function
4. Estimate Parameters via Maximum Likelihood

Complete Recipe

1. Specify a distribution for Y
2. Specify a linear predictor
3. Specify a link function
4. Estimate Parameters via Maximum Likelihood
5. Simulate or Calculate Quantities of Interest

Complete Recipe

1. Specify a distribution for Y
2. Specify a linear predictor
3. Specify a link function
4. Estimate Parameters via Maximum Likelihood
5. Simulate or Calculate Quantities of Interest

Let's do this together for a particular example.

The Data: Political Assassinations

Taken from Olken and Jones (2009), "Hit or Miss? The Effect of Assassinations on Institutions and War", American Economic Journal: Macroeconomics.

The Data: Political Assassinations

Taken from Olken and Jones (2009), "Hit or Miss? The Effect of Assassinations on Institutions and War", American Economic Journal: Macroeconomics.

Dataframe is called as and contains information on assassination attempts, success or failure, and various covariates.

The Data: Political Assassinations

Taken from Olken and Jones (2009), "Hit or Miss? The Effect of Assassinations on Institutions and War", American Economic Journal: Macroeconomics.

Dataframe is called as and contains information on assassination attempts, success or failure, and various covariates.

```
> as[as$country == "United States" & as$year == "1975",]
      country year leadername age tenure attempt
United States 1975       Ford  62    510    TRUE
      survived result dem_score civil_war war
             1     24        10         0   0
          pop  energy solo weapon
       215973 2208506    1    gun
```

Observations are country-year-leaders, so some country-years have multiple observations.

The Data: Political Assassinations

Taken from Olken and Jones (2009), "Hit or Miss? The Effect of Assassinations on Institutions and War", American Economic Journal: Macroeconomics.

Dataframe is called as and contains information on assassination attempts, success or failure, and various covariates.

```
> as[as$country == "United States" & as$year == "1975",]
      country year leadername age tenure attempt
United States 1975       Ford  62    510    TRUE
      survived result dem_score civil_war war
             1     24        10         0   0
          pop  energy solo weapon
       215973 2208506    1    gun
```

Observations are country-year-leaders, so some country-years have multiple observations.

Let's try to predict assassination attempts with some of our covariates.

1. Specify a distribution for Y

1. Specify a distribution for Y

Assume our data was generated from some distribution.

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded:

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary:

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count:

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration:

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration: Exponential

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration: Exponential
- Ordered Categories:

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration: Exponential
- Ordered Categories: Normal with observation mechanism

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration: Exponential
- Ordered Categories: Normal with observation mechanism
- Unordered Categories:

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration: Exponential
- Ordered Categories: Normal with observation mechanism
- Unordered Categories: Multinomial

1. Specify a distribution for Y

Assume our data was generated from some distribution.

Examples:

- Continuous and Unbounded: Normal
- Binary: Bernoulli
- Event Count: Poisson
- Duration: Exponential
- Ordered Categories: Normal with observation mechanism
- Unordered Categories: Multinomial

What fits our application?

2. Specify a linear predictor

We are interested in allowing some parameter of the distribution θ to vary as a (linear) function of covariates. So we specify a linear predictor.

2. Specify a linear predictor

We are interested in allowing some parameter of the distribution θ to vary as a (linear) function of covariates. So we specify a linear predictor.

$$X\beta = \beta_0 + x_1\beta_1 + x_2\beta_2 + \cdots + x_k\beta_k$$

What's in our model?

We wish to predict assassination attempts for country-year-leaders.

What's in our model?

We wish to predict assassination attempts for country-year-leaders.

- `tenure`: number of days in office
- `age`: age of leader, in years
- `dem_score`: polity score, -10 to 10
- `civil_war`: is there currently a civil war?
- `war`: is country in an international conflict?
- `pop`: the country's population, in thousands
- `energy`: energy usage

3. Specify a link function

3. Specify a link function

The link function relates the linear predictor to some parameter θ of the distribution for Y (usually the mean).

3. Specify a link function

The link function relates the linear predictor to some parameter θ of the distribution for Y (usually the mean).

Let $g(\cdot)$ be the link function and let $E(Y) = \theta$ be the mean of distribution for Y.

3. Specify a link function

The link function relates the linear predictor to some parameter θ of the distribution for Y (usually the mean).

Let $g(\cdot)$ be the link function and let $E(Y) = \theta$ be the mean of distribution for Y.

$$g(\theta) = X\beta$$

3. Specify a link function

The link function relates the linear predictor to some parameter θ of the distribution for Y (usually the mean).

Let $g(\cdot)$ be the link function and let $E(Y) = \theta$ be the mean of distribution for Y.

$$\begin{aligned} g(\theta) &= X\beta \\ \theta &= g^{-1}(X\beta) \end{aligned}$$

3. Specify a link function

The link function relates the linear predictor to some parameter θ of the distribution for Y (usually the mean).

Let $g(\cdot)$ be the link function and let $E(Y) = \theta$ be the mean of distribution for Y.

$$\begin{aligned} g(\theta) &= X\beta \\ \theta &= g^{-1}(X\beta) \end{aligned}$$

Note that we usually use the **inverse link function** $g^{-1}(X\beta)$ rather than the link function.

3. Specify a link function

The link function relates the linear predictor to some parameter θ of the distribution for Y (usually the mean).

Let $g(\cdot)$ be the link function and let $E(Y) = \theta$ be the mean of distribution for Y.

$$\begin{aligned} g(\theta) &= X\beta \\ \theta &= g^{-1}(X\beta) \end{aligned}$$

Note that we usually use the **inverse link function** $g^{-1}(X\beta)$ rather than the link function.

Together with the linear predictor this forms the systematic component that we've been talking about all along.

Example Link Functions

Example Link Functions

Identity:

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Probit:

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Probit:

- Link: $\Phi^{-1}(\pi) = X\beta$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Probit:

- Link: $\Phi^{-1}(\pi) = X\beta$
- Inverse Link: $\pi = \Phi(X\beta)$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Probit:

- Link: $\Phi^{-1}(\pi) = X\beta$
- Inverse Link: $\pi = \Phi(X\beta)$

Log:

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Probit:

- Link: $\Phi^{-1}(\pi) = X\beta$
- Inverse Link: $\pi = \Phi(X\beta)$

Log:

- Link: $\ln(\lambda) = X\beta$

Example Link Functions

Identity:

- Link: $\mu = X\beta$

Inverse:

- Link: $\lambda^{-1} = X\beta$
- Inverse Link: $\lambda = (X\beta)^{-1}$

Logit:

- Link: $\ln\left(\frac{\pi}{1-\pi}\right) = X\beta$
- Inverse Link: $\pi = \frac{1}{1+e^{-X\beta}}$

Probit:

- Link: $\Phi^{-1}(\pi) = X\beta$
- Inverse Link: $\pi = \Phi(X\beta)$

Log:

- Link: $\ln(\lambda) = X\beta$
- Inverse Link: $\lambda = \exp(X\beta)$

Logit or Probit?

Logit or Probit?

"The question of which distribution to use is a natural one... There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds ...[A]s a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference." -*Econometric Analysis*, Greene. pg. 774.

Logit or Probit?

"The question of which distribution to use is a natural one... There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds ...[A]s a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference." -*Econometric Analysis*, Greene. pg. 774.

Let's do probit.

Logit or Probit?

"The question of which distribution to use is a natural one... There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds ...[A]s a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference." -*Econometric Analysis*, Greene. pg. 774.

Let's do probit. Why?

Logit or Probit?

"The question of which distribution to use is a natural one... There are practical reasons for favoring one or the other in some cases for mathematical convenience, but it is difficult to justify the choice of one distribution or another on theoretical grounds ...[A]s a general proposition, the question is unresolved. In most applications, the choice between these two seems not to make much difference." -*Econometric Analysis*, Greene. pg. 774.

Let's do probit. Why? Mostly to avoid giving away the problem set.

4. Estimate Parameters via ML

4. Estimate Parameters via ML

a. Write down the likelihood

4. Estimate Parameters via ML

a. Write down the likelihood
b. Estimate all the parameters by maximizing the likelihood.

4. Estimate Parameters via ML

a. Write down the likelihood
b. Estimate all the parameters by maximizing the likelihood.
   - In this case it would be the coefficients β

4. Estimate Parameters via ML

a. Write down the likelihood

b. Estimate all the parameters by maximizing the likelihood.
  - In this case it would be the coefficients β
  - In the regression case it would be $\theta = \{\beta, \gamma\}$ where γ is a reparametrization of the variance.

4. Estimate Parameters via ML

a. Write down the likelihood
b. Estimate all the parameters by maximizing the likelihood.
   - In this case it would be the coefficients β
   - In the regression case it would be $\theta = \{\beta, \gamma\}$ where γ is a reparametrization of the variance.
c. Obtain an estimate of the variance by inverting the negative Hessian

Step 4a: Write Down the Likelihood

The model:

1. $Y_i \sim f_{\text{bern}}(y_i|\pi_i)$.
2. $\pi_i = \Phi(X_i\beta)$ where Φ is the CDF of the standard normal distribution.
3. Y_i and Y_j are independent for all $i \neq j$.

Step 4a: Write Down the Likelihood

The model:

1. $Y_i \sim f_{\text{bern}}(y_i|\pi_i)$.
2. $\pi_i = \Phi(X_i\beta)$ where Φ is the CDF of the standard normal distribution.
3. Y_i and Y_j are independent for all $i \neq j$.

Like all CDF's, Φ has range 0 to 1, so it bounds our π_i to the correct space:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(\frac{z^2}{2}) dz.$$

Step 4a: Write Down the Likelihood

We can then derive the log-likelihood for β:

Step 4a: Write Down the Likelihood

We can then derive the log-likelihood for β:

$$L(\beta|\mathbf{y}) \propto \prod_{i=1}^{n} f_{\text{bern}}(y_i|\pi_i)$$

Step 4a: Write Down the Likelihood

We can then derive the log-likelihood for β:

$$
\begin{aligned}
L(\beta|\mathbf{y}) &\propto \prod_{i=1}^{n} f_{\text{bern}}(y_i|\pi_i) \\
&= \prod_{i=1}^{n} (\pi_i)^{y_i}(1-\pi_i)^{(1-y_i)}
\end{aligned}
$$

Step 4a: Write Down the Likelihood

We can then derive the log-likelihood for β:

$$\begin{aligned} L(\beta|\mathbf{y}) &\propto \prod_{i=1}^{n} f_{\text{bern}}(y_i|\pi_i) \\ &= \prod_{i=1}^{n} (\pi_i)^{y_i}(1-\pi_i)^{(1-y_i)} \end{aligned}$$

Therefore:

Step 4a: Write Down the Likelihood

We can then derive the log-likelihood for β:

$$\begin{aligned} L(\beta|\mathbf{y}) &\propto \prod_{i=1}^{n} f_{\mathrm{bern}}(y_i|\pi_i) \\ &= \prod_{i=1}^{n} (\pi_i)^{y_i}(1-\pi_i)^{(1-y_i)} \end{aligned}$$

Therefore:

$$\ln L(\beta|\mathbf{y}) \propto \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)$$

Step 4a: Write Down the Likelihood

We can then derive the log-likelihood for β:

$$
\begin{aligned}
L(\beta|\mathbf{y}) &\propto \prod_{i=1}^{n} f_{\text{bern}}(y_i|\pi_i) \\
&= \prod_{i=1}^{n} (\pi_i)^{y_i}(1-\pi_i)^{(1-y_i)}
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\ln L(\beta|\mathbf{y}) &\propto \sum_{i=1}^{n} y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i) \\
&= \sum_{i=1}^{n} y_i \ln(\Phi(X_i\beta)) + (1-y_i)\ln(1-\Phi(X_i\beta))
\end{aligned}
$$

Step 4b: Maximize the Likelihood

First implement a function of the likelihood:

Step 4b: Maximize the Likelihood

First implement a function of the likelihood:

```
ll.probit <- function(beta, y=y, X=X){
    phi <- pnorm(X%*%beta, log = TRUE)
    opp.phi <- pnorm(X%*%beta, log = TRUE, lower.tail = FALSE)
    logl <- sum(y*phi + (1-y)*opp.phi)
    return(logl)
}
```

Step 4b: Maximize the Likelihood

First implement a function of the likelihood:

```
ll.probit <- function(beta, y=y, X=X){
    phi <- pnorm(X%*%beta, log = TRUE)
    opp.phi <- pnorm(X%*%beta, log = TRUE, lower.tail = FALSE)
    logl <- sum(y*phi + (1-y)*opp.phi)
    return(logl)
}
```

Notes:

Step 4b: Maximize the Likelihood

First implement a function of the likelihood:

```
ll.probit <- function(beta, y=y, X=X){
    phi <- pnorm(X%*%beta, log = TRUE)
    opp.phi <- pnorm(X%*%beta, log = TRUE, lower.tail = FALSE)
    logl <- sum(y*phi + (1-y)*opp.phi)
    return(logl)
}
```

Notes:

1. the STN CDF is evaluated with pnorm. R's pre-programmed log of the CDF has greater range than `log(pnorm(Z))` (try Z=-50).

Step 4b: Maximize the Likelihood

First implement a function of the likelihood:

```
ll.probit <- function(beta, y=y, X=X){
    phi <- pnorm(X%*%beta, log = TRUE)
    opp.phi <- pnorm(X%*%beta, log = TRUE, lower.tail = FALSE)
    logl <- sum(y*phi + (1-y)*opp.phi)
    return(logl)
}
```

Notes:

1. the STN CDF is evaluated with pnorm. R's pre-programmed log of the CDF has greater range than `log(pnorm(Z))` (try Z=-50).
2. if `lower.tail = FALSE` gives $Pr(Z \geq z)$.

Step 4b: Maximize the Likelihood

First implement a function of the likelihood:

```
ll.probit <- function(beta, y=y, X=X){
    phi <- pnorm(X%*%beta, log = TRUE)
    opp.phi <- pnorm(X%*%beta, log = TRUE, lower.tail = FALSE)
    logl <- sum(y*phi + (1-y)*opp.phi)
    return(logl)
}
```

Notes:

1. the STN CDF is evaluated with pnorm. R's pre-programmed log of the CDF has greater range than `log(pnorm(Z))` (try Z=-50).
2. if `lower.tail = FALSE` gives $Pr(Z \geq z)$.
3. uses a logical test to check that an intercept column has been added

Step 4b: Maximize the Likelihood

```
y <- as$attempt
X <- as[,c("tenure","age","dem_score","civil_war",
          "war","pop","energy")]

yX <- na.omit(cbind(y,X))
y <- yX[,1]; X <- cbind(1,yX[,-1])

opt <- optim(par = rep(0,ncol(X)), fn = ll.probit, y=y, X=X,
      method = "BFGS", control = list(fnscale = -1,
      maxit = 1000), hessian = TRUE)
opt$par
[1] -1.836166916  0.029141779 -0.005751652 -0.012433873  0.066287021  0.317451778
[8]  0.026859441
```

Step 4c: Estimate the Variance-Covariance Matrix

```
vcov <- solve(-opt$hessian)
```

Step 4c: Estimate the Variance-Covariance Matrix

```
vcov <- solve(-opt$hessian)
```

Now we can draw approximate the sampling distribution of beta.

```
MASS::mvrnorm(n=1, mu=opt$par, Sigma=vcov)
[1] -1.819984146 -0.001830225 -0.005933452 -0.012464456  0.122449059  0.380434336  (
[8]  0.008879418
```

Step 4c: Estimate the Variance-Covariance Matrix

```
vcov <- solve(-opt$hessian)
```

Now we can draw approximate the sampling distribution of beta.

```
MASS::mvrnorm(n=1, mu=opt$par, Sigma=vcov)
[1] -1.819984146 -0.001830225 -0.005933452 -0.012464456  0.122449059  0.380434336  
[8]  0.008879418
```

This is stochastic so we do it again and get a different answer:

```
MASS::mvrnorm(n=1, mu=opt$par, Sigma=vcov)
[1] -1.792636772  0.081477117 -0.006457063 -0.013436530  0.019081307  0.255634394  
[8]  0.073840550
```

5. Quantities of Interest

5. Quantities of Interest

What not to do...

5. Quantities of Interest

What not to do...

	est	SE
Intercept	-1.8362	1.4012
tenure	0.0291	0.2937
age	-0.0058	0.0251
dem_score	-0.0124	0.0445
civil_war	0.0663	0.8918
war	0.3175	1.0141
pop	0.0409	0.2368
energy	0.0269	0.2432

```
ses <- sqrt(diag(solve(-opt$hessian)))
table.dat <- cbind(opt$par, ses)
rownames(table.dat) <- colnames(X)
xtable::xtable(table.dat, digits = 4)
```

5. Quantities of Interest

5. Quantities of Interest

1. Simulate parameters from multivariate normal.

5. Quantities of Interest

1. Simulate parameters from multivariate normal.
2. Run $X\beta$ through inverse link function to get the original parameter (typically the distribution mean)

5. Quantities of Interest

1. Simulate parameters from multivariate normal.
2. Run $X\beta$ through inverse link function to get the original parameter (typically the distribution mean)
3. Draw from distribution of Y for predicted values.

5. Quantities of Interest

General considerations:

5. Quantities of Interest

General considerations:

a. Incorporating estimation uncertainty.

5. Quantities of Interest

General considerations:

a. Incorporating estimation uncertainty.

b. Incorporating fundamental uncertainty when making predictions.

5. Quantities of Interest

General considerations:

a. Incorporating estimation uncertainty.
b. Incorporating fundamental uncertainty when making predictions.
c. Establishing appropriate baseline values for QOI, and considering plausible changes in those values.

Expected Values

For this model we will be interested in estimating the predicted probability of an assassination attempt at some level for the covariate values. In general, $E[y|X]$.

Expected Values

For this model we will be interested in estimating the predicted probability of an assassination attempt at some level for the covariate values. In general, $E[y|X]$.

Let's consider a potentially high risk situations (we'll call them "highrisk","X_{HR}") then we can manipulate the risk factors:

Expected Values

For this model we will be interested in estimating the predicted probability of an assassination attempt at some level for the covariate values. In general, $E[y|X]$.

Let's consider a potentially high risk situations (we'll call them "highrisk","X_{HR}") then we can manipulate the risk factors:

Var.	Value
tenure	-0.30
age	54.00
dem_score	-3.00
civil_war	0.00
war	0.00
pop	-0.18
energy	-0.23

Expected Values

What's the estimated probability of an assassination at X_{HR}?

Expected Values

What's the estimated probability of an assassination at X_{HR}?

Draw $\tilde{\beta}$

```
beta.draws <- MASS::mvrnorm(10000, mu = opt$par, Sigma = vcov)
dim(beta.draws)
[1] 10000     8
```

Now we simulate the outcome (warning: inefficient code!)

```
nsims <- 10000
p.ests <- vector(length=nrow(beta.draws))
for(i in 1:nsims){
  p.ass.att <- pnorm(highrisk%*%beta.draws[i,])
  outcomes <- rbinom(nsims2, 1, p.ass.att)
  p.ests[i] <- mean(outcomes)
}
> mean(p.ests)
[1] 0.0166266
> quantile(p.ests, .025); quantile(p.ests, .975)
 2.5%    97.5%
0.0134  0.0201
```

Expected Values

What are the steps that I just took?

Expected Values

What are the steps that I just took?

1. simulate from the estimated sampling distribution of $\hat{\beta}$ to incorporate estimation uncertainty.

Expected Values

What are the steps that I just took?

1. simulate from the estimated sampling distribution of $\hat{\beta}$ to incorporate estimation uncertainty.
2. start our for-loop which will do steps 3-5 each time.

Expected Values

What are the steps that I just took?

1. simulate from the estimated sampling distribution of $\hat{\beta}$ to incorporate estimation uncertainty.
2. start our for-loop which will do steps 3-5 each time.
3. combine one $\tilde{\beta}$ draw with X_{HR} as $X_{HR}\tilde{\beta}$, then plug into $\Phi()$ to get probability of attempt for that $\tilde{\beta}$ draw.

Expected Values

What are the steps that I just took?

1. simulate from the estimated sampling distribution of $\hat{\beta}$ to incorporate estimation uncertainty.
2. start our for-loop which will do steps 3-5 each time.
3. combine one $\tilde{\beta}$ draw with X_{HR} as $X_{HR}\tilde{\beta}$, then plug into $\Phi()$ to get probability of attempt for that $\tilde{\beta}$ draw.
4. draw a bunch of outcomes from the $Bernoulli(\Phi(X_{HR}\tilde{\beta}))$.

Expected Values

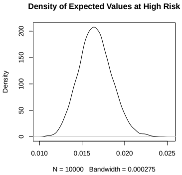
What are the steps that I just took?

1. simulate from the estimated sampling distribution of $\hat{\beta}$ to incorporate estimation uncertainty.
2. start our for-loop which will do steps 3-5 each time.
3. combine one $\tilde{\beta}$ draw with X_{HR} as $X_{HR}\tilde{\beta}$, then plug into $\Phi()$ to get probability of attempt for that $\tilde{\beta}$ draw.
4. draw a bunch of outcomes from the $Bernoulli(\Phi(X_{HR}\tilde{\beta}))$.
5. average over those draws to get one simulated $E[y|X_{HR}]$.

Expected Values

What are the steps that I just took?

1. simulate from the estimated sampling distribution of $\hat{\beta}$ to incorporate estimation uncertainty.
2. start our for-loop which will do steps 3-5 each time.
3. combine one $\tilde{\beta}$ draw with X_{HR} as $X_{HR}\tilde{\beta}$, then plug into $\Phi()$ to get probability of attempt for that $\tilde{\beta}$ draw.
4. draw a bunch of outcomes from the $Bernoulli(\Phi(X_{HR}\tilde{\beta}))$.
5. average over those draws to get one simulated $E[y|X_{HR}]$.
6. return to step 3.



Expected Values: A Shortcut

There is a shorter way to come up with the same answer, but it requires some care in its application.

Expected Values: A Shortcut

There is a shorter way to come up with the same answer, but it requires some care in its application.

```
beta.draws <- mvrnorm(10000, mu = opt$par, Sigma
              = solve(-opt$hessian))
p.ests2 <- pnorm(highrisk%*%t(beta.draws))

> mean(p.ests2)
[1] 0.01659705
> quantile(p.ests2, .025); quantile(p.ests2, .975)
      2.5%        97.5%
0.01395935   0.01955867
```

Expected Values: A Shortcut

There is a shorter way to come up with the same answer, but it requires some care in its application.

```
beta.draws <- mvrnorm(10000, mu = opt$par, Sigma
              = solve(-opt$hessian))
p.ests2 <- pnorm(highrisk%*%t(beta.draws))

> mean(p.ests2)
[1] 0.01659705
> quantile(p.ests2, .025); quantile(p.ests2, .975)
      2.5%        97.5%
0.01395935   0.01955867
```

This shortcut works because $E[y|X_{HR}] = \pi_{HR}$; i.e. the parameter is the expected value of the outcome.

When wouldn't this work?

When wouldn't this work?

Ex.: suppose that $y_i \sim Expo(\lambda_i)$ where $\lambda_i = exp(X_i\beta)$. We could find our likelihood, insert our parameterization of λ_i for each i, and then maximize to find $\hat{\beta}$ as usual.

When wouldn't this work?

Ex.: suppose that $y_i \sim Expo(\lambda_i)$ where $\lambda_i = exp(X_i \beta)$. We could find our likelihood, insert our parameterization of λ_i for each i, and then maximize to find $\hat{\beta}$ as usual.

Thus, for some baseline set of covariates X_{BL}, we now have a simulated sampling distribution for λ_{BL} which has a mean at $E[exp(X_{BL}\hat{\beta})]$.

When wouldn't this work?

Ex.: suppose that $y_i \sim Expo(\lambda_i)$ where $\lambda_i = exp(X_i\beta)$. We could find our likelihood, insert our parameterization of λ_i for each i, and then maximize to find $\hat{\beta}$ as usual.

Thus, for some baseline set of covariates X_{BL}, we now have a simulated sampling distribution for λ_{BL} which has a mean at $E[exp(X_{BL}\hat{\beta})]$.

Its not too hard to show that if $y \sim Expo(\lambda)$, then $E[y] = \frac{1}{\lambda}$. The temptation is then to declare that because $E[\hat{\lambda}_{BL}] = E[exp(X_{BL}\hat{\beta})]$ then $\widehat{E[y]} = 1/E[\exp(X_{BL}\hat{\beta})]$.

When wouldn't this work?

Ex.: suppose that $y_i \sim Expo(\lambda_i)$ where $\lambda_i = exp(X_i\beta)$. We could find our likelihood, insert our parameterization of λ_i for each i, and then maximize to find $\hat{\beta}$ as usual.

Thus, for some baseline set of covariates X_{BL}, we now have a simulated sampling distribution for λ_{BL} which has a mean at $E[exp(X_{BL}\hat{\beta})]$.

Its not too hard to show that if $y \sim Expo(\lambda)$, then $E[y] = \frac{1}{\lambda}$. The temptation is then to declare that because $E[\hat{\lambda}_{BL}] = E[exp(X_{BL}\hat{\beta})]$ then $\widehat{E[y]} = 1/E[\exp(X_{BL}\hat{\beta})]$.

It turns out this is not the case because $E[1/\hat{\lambda}] \neq 1/E[\hat{\lambda}]$. The first averages over the sampling distribution of the means of y. The second averages over the sampling distribution of $\hat{\lambda}$ then plugs into the formula for the mean of y.

When wouldn't this work?

Why this annoying wrinkle?

When wouldn't this work?

Why this annoying wrinkle? Jensen's inequality: given a random variable X, $E[g(X)] \neq g(E[X])$ (it's $\geq$ if $g(\cdot)$ is concave; $\leq$ if $g(\cdot)$ is convex).

When wouldn't this work?

Why this annoying wrinkle? Jensen's inequality: given a random variable X, $E[g(X)] \neq g(E[X])$ (it's $\geq$ if $g(\cdot)$ is concave; $\leq$ if $g(\cdot)$ is convex).

Why can we use our shortcut with the Probit model?

When wouldn't this work?

Why this annoying wrinkle? Jensen's inequality: given a random variable X, $E[g(X)] \neq g(E[X])$ (it's $\geq$ if $g(\cdot)$ is concave; $\leq$ if $g(\cdot)$ is convex).

Why can we use our shortcut with the Probit model? If $Y \sim Bern(\pi)$ then $E[Y] = \pi$. Our guess would then be that $\widehat{E[Y]} = E[\Phi(X_{HR}\hat{\beta})]$ which is fine because $1 \cdot \ E[\Phi(X_{HR}\hat{\beta})] = E[1 \cdot \Phi(X_{HR}\hat{\beta})]$.

When wouldn't this work?

Why this annoying wrinkle? Jensen's inequality: given a random variable X, $E[g(X)] \neq g(E[X])$ (it's $\geq$ if $g(\cdot)$ is concave; $\leq$ if $g(\cdot)$ is convex).

Why can we use our shortcut with the Probit model? If $Y \sim Bern(\pi)$ then $E[Y] = \pi$. Our guess would then be that $\widehat{E[Y]} = E[\Phi(X_{HR}\hat{\beta})]$ which is fine because $1 \cdot \; E[\Phi(X_{HR}\hat{\beta})] = E[1 \cdot \Phi(X_{HR}\hat{\beta})]$.

$\boxed{\text{\underline{Rule of thumb}: if } E[Y] = \theta\text{, you are safe taking the shortcut.}}$

More Expected Values

What if I want a bunch of these to see how expected values change with some variable?

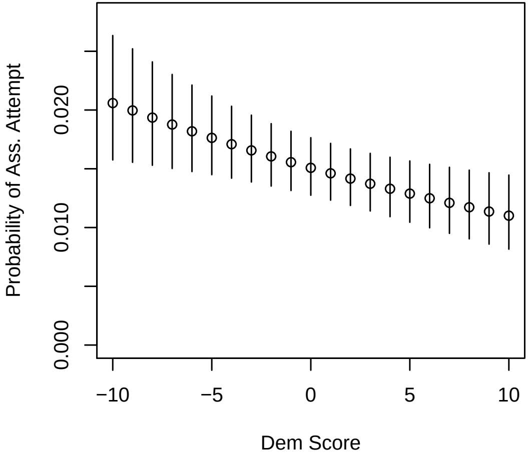
More Expected Values

What if I want a bunch of these to see how expected values change with some variable?

```
dem.rng <- -10:10
p.ests <- matrix(data = NA, ncol = length(dem.rng),
                 nrow=10000)

for(j in 1:length(dem.rng)){
  highrisk.dem <- highrisk
  highrisk.dem["dem_score"] <- dem.rng[j]
  p.ests[,j] <- pnorm(highrisk.dem%*%t(beta.draws))
}

plot(dem.rng, apply(p.ests,2,mean), ylim = c(0,.028))
segments(x0 = dem.rng, x1 = dem.rng,
         y0 = apply(p.ests, 2, quantile, .025),
         y1 = apply(p.ests, 2, quantile, .975))
```


Probability of Ass. Attempt
0.000
0.010
0.020
−10
−5
0
5
10
Dem Score

Predicted Values

What if someone asks me to predict whether an assassination will take place? If I were to simulate from the distribution of $\hat{\beta}$, and then draw 1 value from the stochastic component for each simulation I would get a predictive distribution for $y|X$.

Predicted Values

What if someone asks me to predict whether an assassination will take place? If I were to simulate from the distribution of $\hat{\beta}$, and then draw 1 value from the stochastic component for each simulation I would get a predictive distribution for $y|X$.
Q: What will it look like?

Predicted Values

What if someone asks me to predict whether an assassination will take place? If I were to simulate from the distribution of $\hat{\beta}$, and then draw 1 value from the stochastic component for each simulation I would get a predictive distribution for $y|X$.
Q: What will it look like?

There is no need to actually conduct the simulation, though. The simulated outcomes will be $Bern(\widehat{E[y|X]}) = Bern(.166)$. How is this different than the linear regression case?

First Differences

Compare expected values of the outcome for two different scenarios, usually all predictors held constant but one. Recall from our regression results that war seemed to have a big positive effect on probability of an assassination attempt.

First Differences

Compare expected values of the outcome for two different scenarios, usually all predictors held constant but one. Recall from our regression results that war seemed to have a big positive effect on probability of an assassination attempt.

So let's find:

$$E[y|X_{War}] - E[y|X_{Nowar}].$$

First Differences

Compare expected values of the outcome for two different scenarios, usually all predictors held constant but one. Recall from our regression results that war seemed to have a big positive effect on probability of an assassination attempt.

So let's find:

$$E[y|X_{War}] - E[y|X_{Nowar}].$$

Each of these are just fitted values for the probability parameter, with all covariates at the highrisk values except war, which we control.

First Differences

```
highrisk.war <- highrisk
highrisk.war["war"] <- 1
highrisk.nowar <- highrisk
highrisk.nowar["war"] <- 0
```

First Differences

```
highrisk.war <- highrisk
highrisk.war["war"] <- 1
highrisk.nowar <- highrisk
highrisk.nowar["war"] <- 0

fd.ests <- pnorm(highrisk.war%*%t(beta.draws)) -
           pnorm(highrisk.nowar%*%t(beta.draws))

> mean(fd.ests)
[1] 0.01891
> quantile(fd.ests, .025); quantile(fd.ests, .975)
   2.5%      97.5%
0.00578    0.03609
```

Simulating (Parameter) Estimation Uncertainty

To take one random draw of all the parameters $\gamma = (\vec{\beta}, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

Simulating (Parameter) Estimation Uncertainty

To take one random draw of all the parameters $\gamma = (\vec{\beta}, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\mathsf{V}}(\hat{\gamma})$.

Simulating (Parameter) Estimation Uncertainty

To take one random draw of all the parameters $\gamma = (\vec{\beta}, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\mathrm{V}}(\hat{\gamma})$.
2. Draw the vector γ from the multivariate normal distribution:

Simulating (Parameter) Estimation Uncertainty

To take one random draw of all the parameters $\gamma = (\vec{\beta}, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\mathsf{V}}(\hat{\gamma})$.
2. Draw the vector γ from the multivariate normal distribution:

$$\gamma \sim \mathsf{N}\left(\hat{\gamma}, \hat{\mathsf{V}}(\hat{\gamma})\right)$$

Simulating (Parameter) Estimation Uncertainty

To take one random draw of all the parameters $\gamma = (\vec{\beta}, \alpha)$ from their "sampling distribution" (or "posterior distribution" with a flat prior):

1. Estimate the model by maximizing the likelihood function, record the point estimates $\hat{\gamma}$ and variance matrix $\hat{\mathsf{V}}(\hat{\gamma})$.
2. Draw the vector γ from the multivariate normal distribution:

$$\gamma \sim \mathsf{N}\left(\hat{\gamma}, \hat{\mathsf{V}}(\hat{\gamma})\right)$$

   Denote the draw $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$, which has *k* elements.

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

To simulate one predicted value, follow these steps:

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

To simulate one predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

To simulate one predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose a predicted value to compute, defined by one value for each explanatory variable as the vector X_c.

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

To simulate one predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose a predicted value to compute, defined by one value for each explanatory variable as the vector X_c.
3. Extract simulated $\tilde{\beta}$ from $\tilde{\gamma}$; compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from systematic component)

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

To simulate one predicted value, follow these steps:

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose a predicted value to compute, defined by one value for each explanatory variable as the vector X_c.
3. Extract simulated $\tilde{\beta}$ from $\tilde{\gamma}$; compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from systematic component)
4. Simulate outcome variable $\tilde{Y}_c \sim f(\tilde{\theta}_c, \tilde{\alpha})$ (from stochastic component)

Simulating the Distribution of Predicted Values, $\sim Y$

Predicted values can be for:

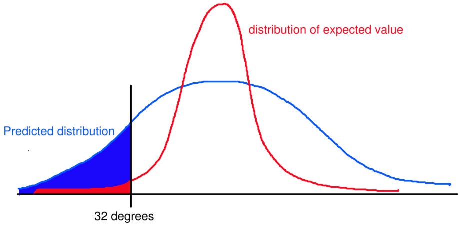
1. Forecasts: about the future
2. Farcasts: about some area for which you have no y
3. Nowcasts: about the current data (perhaps to reproduce it to see whether it fits)

To simulate one predicted value, follow these steps:

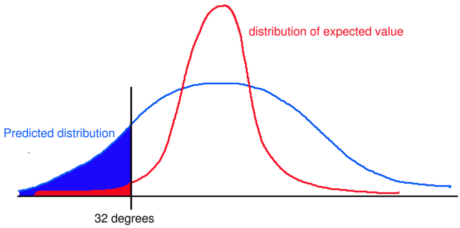
1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose a predicted value to compute, defined by one value for each explanatory variable as the vector X_c.
3. Extract simulated $\tilde{\beta}$ from $\tilde{\gamma}$; compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from systematic component)
4. Simulate outcome variable $\tilde{Y}_c \sim f(\tilde{\theta}_c, \tilde{\alpha})$ (from stochastic component)

Repeat algorithm say $M = 1000$ times, to produce 1000 predicted values. Use these to compute a histogram for the full posterior, the average, variance, percentile values, or others.

The Distribution of Expected v. Predicted Values

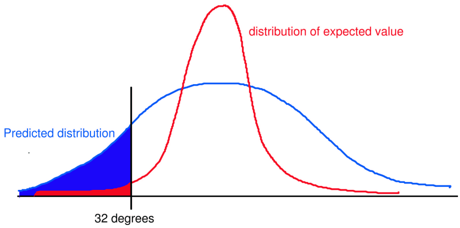


The Distribution of Expected v. Predicted Values



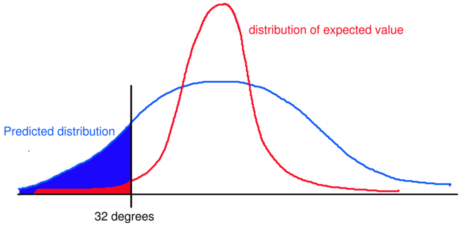
1. Predicted values: draws of Y that are or could be observed

The Distribution of Expected v. Predicted Values



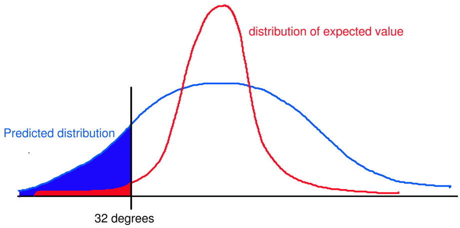
1. Predicted values: draws of Y that are or could be observed
2. Expected values: draws of fixed features of the distribution of Y, such as $E(Y)$.

The Distribution of Expected v. Predicted Values

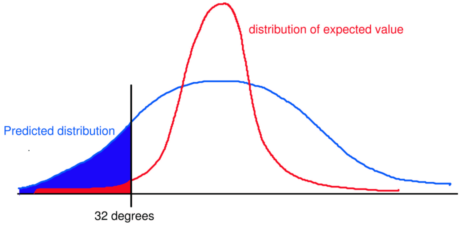


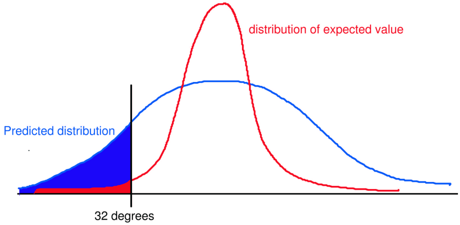
1. Predicted values: draws of Y that are or could be observed
2. Expected values: draws of fixed features of the distribution of Y, such as $E(Y)$.
3. Predicted values: include estimation and fundamental uncertainty.

The Distribution of Expected v. Predicted Values

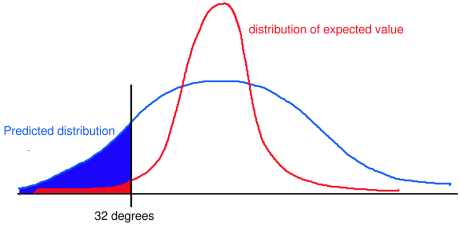


1. Predicted values: draws of Y that are or could be observed
2. Expected values: draws of fixed features of the distribution of Y, such as $E(Y)$.
3. Predicted values: include estimation and fundamental uncertainty.
4. Expected values: average away fundamental uncertainty

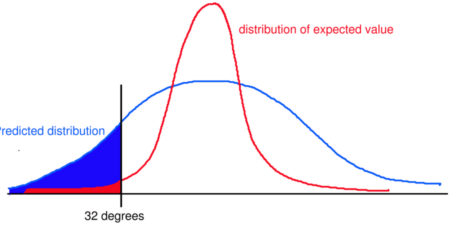

distribution of expected value
Predicted distribution
32 degrees



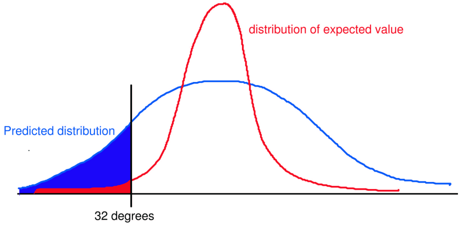
5. The variance of expected values (but not predicted values) go to 0 and n gets large.



5. The variance of expected values (but not predicted values) go to 0 and n gets large.
6. Example use of predicted value distribution: probability of temperature colder than 32° tomorrow. (Predicted temperature is uncertain because we have to estimate it and because of natural fluctuations.)



5. The variance of expected values (but not predicted values) go to 0 and n gets large.
6. Example use of predicted value distribution: probability of temperature colder than 32° tomorrow. (Predicted temperature is uncertain because we have to estimate it and because of natural fluctuations.)
7. Example use of expected value distribution: probability the average temperature on days like tomorrow will be colder than 32°. (Expected temperature is only uncertain because we have to estimate it; natural fluctuations in temperature doesn't affect the average.)



5. The variance of expected values (but not predicted values) go to 0 and n gets large.
6. Example use of predicted value distribution: probability of temperature colder than 32° tomorrow. (Predicted temperature is uncertain because we have to estimate it and because of natural fluctuations.)
7. Example use of expected value distribution: probability the average temperature on days like tomorrow will be colder than 32°. (Expected temperature is only uncertain because we have to estimate it; natural fluctuations in temperature doesn't affect the average.)
8. Which to use for causal effects & first differences?

Simulating the Distribution of Expected Values: An Algorithm

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)
3. Taking the one set of simulated $\tilde{\beta}$ from $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from the systematic component)

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)
3. Taking the one set of simulated $\tilde{\beta}$ from $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from the systematic component)
4. Draw m values of the outcome variable $\tilde{Y}_c^{(k)}$ ($k = 1, \ldots, m$) from the stochastic component $f(\tilde{\theta}_c, \tilde{\alpha})$. (This step simulates fundamental uncertainty.)

Simulating the Distribution of Expected Values: An Algorithm

1. Draw one value of $\tilde{\gamma} = \text{vec}(\tilde{\beta}, \tilde{\alpha})$.
2. Choose one value for each explanatory variable (X_c is a vector)
3. Taking the one set of simulated $\tilde{\beta}$ from $\tilde{\gamma}$, compute $\tilde{\theta}_c = g(X_c, \tilde{\beta})$ (from the systematic component)
4. Draw m values of the outcome variable $\tilde{Y}_c^{(k)}$ ($k = 1, \dots, m$) from the stochastic component $f(\tilde{\theta}_c, \tilde{\alpha})$. (This step simulates fundamental uncertainty.)
5. Average over the fundamental uncertainty by calculating the mean of the m simulations to yield one simulated expected value $\tilde{E}(Y_c) = \sum_{k=1}^{m} \tilde{Y}_c^{(k)}/m$.

Simulating Expected Values: Notes

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.
3. Repeat entire algorithm M times (say 1000), with results differing only due to estimation uncertainty

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.
3. Repeat entire algorithm M times (say 1000), with results differing only due to estimation uncertainty
4. Use to compute a histogram, average, standard error, confidence interval, etc.

Simulating Expected Values: Notes

1. When $m = 1$, this algorithm produces predicted values.
2. With large m, this algorithm better represents and averages over the fundamental uncertainty.
3. Repeat entire algorithm M times (say 1000), with results differing only due to estimation uncertainty
4. Use to compute a histogram, average, standard error, confidence interval, etc.
5. When $E(Y_c) = \theta_c$, we can skip the last two steps. E.g., in the logit model, once we simulate π_i, we don't need to draw Y and then average to get back to π_i. (If you're unsure, do it anyway!)

Simulating First Differences

Simulating First Differences

To draw one simulated first difference:

Simulating First Differences

To draw one simulated first difference:

1. Choose vectors X_s, the starting point, X_e, the ending point.

Simulating First Differences

To draw one simulated first difference:

1. Choose vectors X_s, the starting point, X_e, the ending point.
2. Apply the expected value algorithm twice, once for X_s and X_e (but reuse the random draws).

Simulating First Differences

To draw one simulated first difference:

1. Choose vectors X_s, the starting point, X_e, the ending point.
2. Apply the expected value algorithm twice, once for X_s and X_e (but reuse the random draws).
3. Take the difference in the two expected values.

Simulating First Differences

To draw one simulated first difference:

1. Choose vectors X_s, the <u>s</u>tarting point, X_e, the <u>e</u>nding point.
2. Apply the expected value algorithm twice, once for X_s and X_e (but reuse the random draws).
3. Take the difference in the two expected values.
4. (To save computation time, and improve approximation, use the same simulated β in each.)

Tricks for Simulating Parameters

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to
   - make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.)

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to
   - make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.)
   - make the maximization algorithm work faster without constraints

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to
   - make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.)
   - make the maximization algorithm work faster without constraints
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to
   - make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.)
   - make the maximization algorithm work faster without constraints
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   - $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to
   - make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.)
   - make the maximization algorithm work faster without constraints
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   - $\sigma^2 = e^{\eta}$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^{η})
   - For a probability, $\pi = [1 + e^{-\eta}]^{-1}$ (a logit transformation).

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to
   - make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.)
   - make the maximization algorithm work faster without constraints
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   - $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)
   - For a probability, $\pi = [1 + e^{-\eta}]^{-1}$ (a logit transformation).
   - For $-1 \leq \rho \leq 1$, use $\rho = (e^{2\eta} - 1)/(e^{2\eta} + 1)$ (Fisher's Z transformation)

Tricks for Simulating Parameters

1. Simulate all parameters (in γ), including ancillary parameters, together, unless you know they are orthogonal.
2. Reparameterize to unbounded scale to
   - make $\hat{\gamma}$ converge more quickly in n (and so work better with small n) to a multivariate normal. (MLEs don't change, but the posteriors do.)
   - make the maximization algorithm work faster without constraints
3. To do this, all estimated parameters should be unbounded and logically symmetric. E.g.,
   - $\sigma^2 = e^\eta$ (i.e., wherever you see σ^2, in your log-likelihood function, replace it with e^η)
   - For a probability, $\pi = [1 + e^{-\eta}]^{-1}$ (a logit transformation).
   - For $-1 \leq \rho \leq 1$, use $\rho = (e^{2\eta} - 1)/(e^{2\eta} + 1)$ (Fisher's Z transformation)

   In all 3 cases, η is unbounded: estimate it, simulate from it, and reparameterize back to the scale you care about.

Tricks for Simulating Quantities of Interest

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)

2. Simulating functions of Y

   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.

   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.

   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.
3. Check the approximation error of your simulation algorithm: Run it twice, check the number of digits of precision that don't change. If its not enough for your tables, increase M (or m) and try again.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.
3. Check the approximation error of your simulation algorithm: Run it twice, check the number of digits of precision that don't change. If its not enough for your tables, increase M (or m) and try again.
4. Analytical calculations and other tricks can speed simulation, or precision.

Tricks for Simulating Quantities of Interest

1. Unless you're sure, always compute simulations of Y and use that as a basis for creating simulations of other quantities. (This will get all information from the model in the simulations.)
2. Simulating functions of Y
   (a) If some function of Y, such as $\ln(Y)$, is used, simulate $\ln(Y)$ and then apply the inverse function $\exp(\ln(Y))$ to reveal Y.
   (b) The usual, but wrong way: Regress $\ln(Y)$ on X, compute predicted value $\widehat{\ln(Y)}$ and exponentiate.
   (c) Its wrong because the regression estimates $E[\ln(Y)]$, but $E[\ln(Y)] \neq \ln[E(Y)]$, so $\exp(E[\ln(Y)]) \neq Y$
   (d) More generally, $E(g[Y]) \neq g[E(Y)]$, unless $g[\cdot]$ is linear.
3. Check the <u>approximation error</u> of your simulation algorithm: Run it twice, check the number of digits of precision that don't change. If its not enough for your tables, increase M (or m) and try again.
4. Analytical calculations and other tricks can speed simulation, or precision.
5. Canned Software Options: Clarify in Stata, Zelig in R

Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

1. Logit of reported turnout on Age, Age^2, Education, Income, and Race

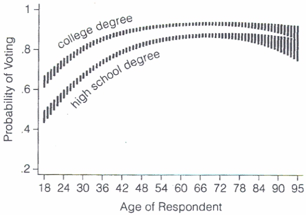
Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

1. Logit of reported turnout on Age, Age^2, Education, Income, and Race
2. Quantity of Interest: (nonlinear) effect of age on $\Pr(vote|X)$, holding constant Income and Race.

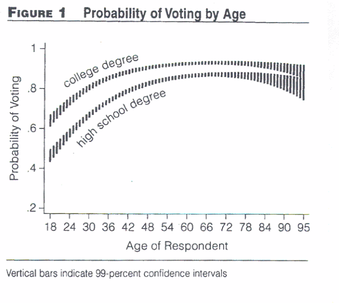
Replication of Rosenstone and Hansen from King, Tomz and Wittenberg (2000)

1. Logit of reported turnout on Age, Age^2, Education, Income, and Race
2. Quantity of Interest: (nonlinear) effect of age on $\Pr(vote|X)$, holding constant Income and Race.
3. Use $M = 1000$ and compute 99% CI:

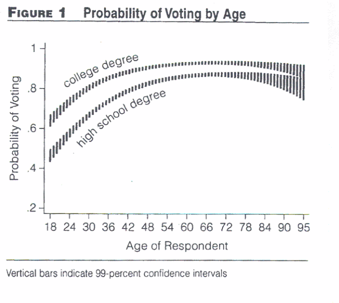
FIGURE 1 Probability of Voting by Age



Vertical bars indicate 99-percent confidence intervals

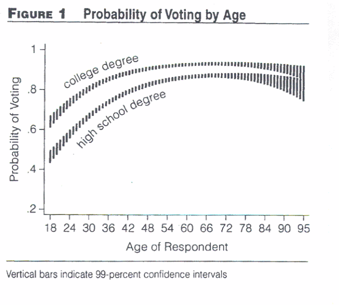


To create this graph, simulate:



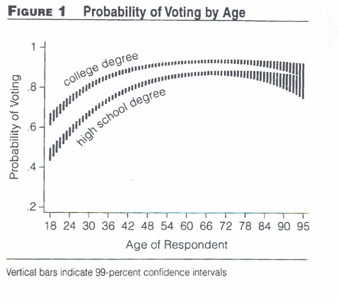
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white



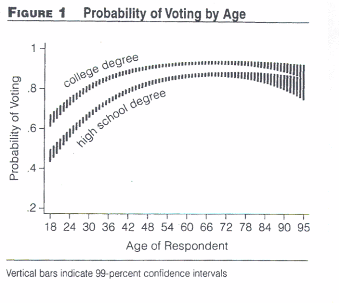
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression



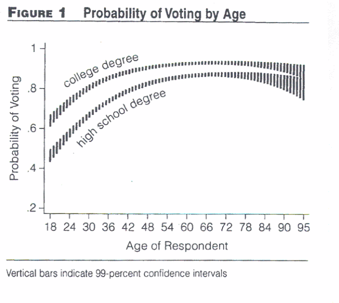
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s



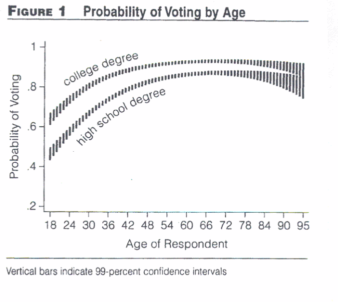
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$



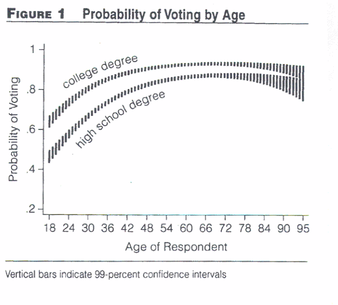
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort in numerical order



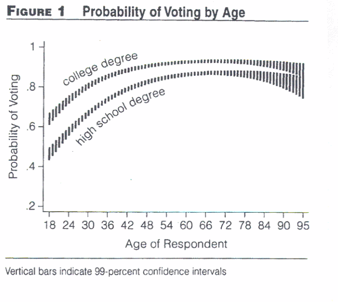
To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort in numerical order
6. Take 5th and 995th values as the 99% confidence interval



To create this graph, simulate:

1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort in numerical order
6. Take 5th and 995th values as the 99% confidence interval
7. Plot a vertical line on the graph at age=24 representing the CI.



To create this graph, simulate:

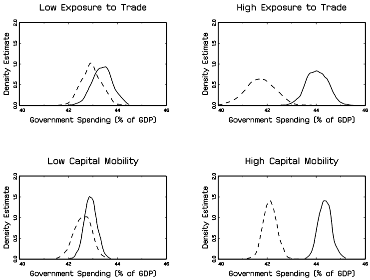
1. Set age=24, education=high school, income=average, Race=white
2. Run logistic regression
3. Simulate 1000 $\tilde{\beta}$'s
4. Compute 1000 $\tilde{\pi}_i = [1 + e^{x_i \tilde{\beta}}]^{-1}$
5. Sort in numerical order
6. Take 5th and 995th values as the 99% confidence interval
7. Plot a vertical line on the graph at age=24 representing the CI.
8. Repeat for other ages and for college degree.

Replication of Garrett (King, Tomz and Wittenberg 2000)

- Dependent variable: Government Spending as % of GDP
- Key explanatory variable: left-labor power (high = solid line; low = dashed)
- Garrett used only point estimates to distinguish the eight quantities represented above. What new information do we learn with this approach?
- Left-labor power only has a clear effect when exposure to trade or capital mobility is high.

See Last Semester's Slides

Replication of Garrett (King, Tomz and Wittenberg 2000)



- Dependent variable: Government Spending as % of GDP
- Key explanatory variable: left-labor power (high = solid line; low = dashed)
- Garrett used only point estimates to distinguish the eight quantities represented above. What new information do we learn with this approach?
- Left-labor power only has a clear effect when exposure to trade or capital mobility is high.

See Last Semester's Slides

Model Diagnostics for Binary Outcome Models

- How do you quantify the goodness of fit of your model?

Model Diagnostics for Binary Outcome Models

- How do you quantify the goodness of fit of your model?
- *Note*: Goodness of fit may or may not be very important

Model Diagnostics for Binary Outcome Models

- How do you quantify the goodness of fit of your model?
- *Note*: Goodness of fit may or may not be very important
- A model that fits well is likely to be a good *predictive* model

Model Diagnostics for Binary Outcome Models

- How do you quantify the goodness of fit of your model?
- *Note*: Goodness of fit may or may not be very important
- A model that fits well is likely to be a good *predictive* model
- But it does not guarantee that the model is a good *causal* model

Model Diagnostics for Binary Outcome Models

- How do you quantify the goodness of fit of your model?

- *Note*: Goodness of fit may or may not be very important
- A model that fits well is likely to be a good *predictive* model
- But it does not guarantee that the model is a good *causal* model
- Pseudo-R^2: a generalization of R^2 to outside of the linear world

$$\tilde{R}^2 \;=\; 1 - \frac{\ell(\hat{\beta}_{MLE})}{\ell(\bar{y})} \;\in\; [0,1]$$

$\ell(\bar{y})$: log-likelihood of the null model, which sets $\hat{\pi}_i = \bar{y}$ for all i

Model Diagnostics for Binary Outcome Models

- How do you quantify the goodness of fit of your model?

- *Note*: Goodness of fit may or may not be very important
- A model that fits well is likely to be a good *predictive* model
- But it does not guarantee that the model is a good *causal* model
- Pseudo-R^2: a generalization of R^2 to outside of the linear world

$$\tilde{R}^2 \ = \ 1 - \frac{\ell(\hat{\beta}_{MLE})}{\ell(\bar{y})} \ \in \ [0, 1]$$

$\ell(\bar{y})$: log-likelihood of the null model, which sets $\hat{\pi}_i = \bar{y}$ for all i

- This one is due to McFadden (1974); many other variants exist

How do you know which model is better?

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set; compare to the test set.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
   (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set; compare to the test set.
   (e) Comparisons to average prediction and full distribution.

How do you know which model is better?

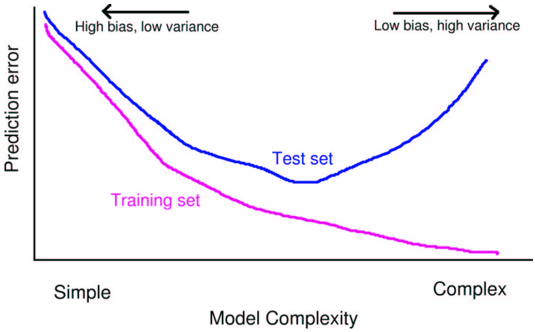
1. Out-of-sample forecasts (or farcasts)
   (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.
   (b) Set aside some (test) data.
   (c) Fit your model to the rest (the training data).
   (d) Make predictions with training set; compare to the test set.
   (e) Comparisons to average prediction and full distribution.
   (f) E.g., if a set of predictions have $\Pr(y = 1) = 0.2$, then 20% of these observations in the test set should be 1s.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
    (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.
    (b) Set aside some (test) data.
    (c) Fit your model to the rest (the training data).
    (d) Make predictions with training set; compare to the test set.
    (e) Comparisons to average prediction and full distribution.
    (f) E.g., if a set of predictions have $\Pr(y = 1) = 0.2$, then 20% of these observations in the test set should be 1s.
    (g) The best test sets are really out of sample, not even available yet.

How do you know which model is better?

1. Out-of-sample forecasts (or farcasts)
    (a) Your job: find the underlying (persistent) structure, not the idiosyncratic features of any one data set.
    (b) Set aside some (test) data.
    (c) Fit your model to the rest (the training data).
    (d) Make predictions with training set; compare to the test set.
    (e) Comparisons to average prediction and full distribution.
    (f) E.g., if a set of predictions have $\Pr(y = 1) = 0.2$, then 20% of these observations in the test set should be 1s.
    (g) The best test sets are really out of sample, not even available yet.
    (h) If the world changes, an otherwise good model will fail. But it's still the right test.



(See Trevor Hastie et al. 2001. The Elements of Statistical Learning, Springer, Chapter 7: Fig 7.1.)

(i) Binary variable predictions require a normative decision.

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.
- Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.
- Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
  - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.
- Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
  - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
  - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.
- Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
  - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
  - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
- Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(i) Binary variable predictions require a normative decision.
  - Let C be number of times more costly misclassifying a 1 is than a 0.
  - C must be chosen independently of the data.
  - C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
  - People often choose $C = 1$, but without justification.
  - Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
    - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
    - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
  - Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.
- Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
  - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
  - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
- Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves

- Compute %1s and %0s correctly predicted for every possible value of C

(i) Binary variable predictions require a normative decision.
- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.
- Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
  - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
  - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
- Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves
- Compute %1s and %0s correctly predicted for every possible value of C
- Plot %1s by %0s

(i) Binary variable predictions require a normative decision.

- Let C be number of times more costly misclassifying a 1 is than a 0.
- C must be chosen independently of the data.
- C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
- People often choose $C = 1$, but without justification.
- Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
  - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
  - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
- Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves

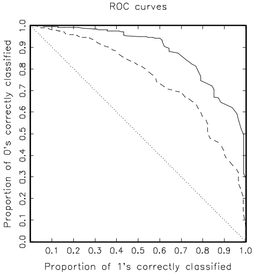
- Compute %1s and %0s correctly predicted for every possible value of C
- Plot %1s by %0s
- Overlay curves for several models on the same graph.

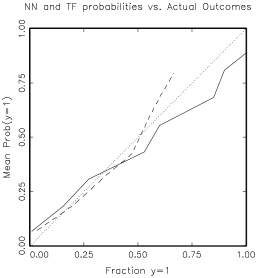
(i) Binary variable predictions require a normative decision.
  - Let C be number of times more costly misclassifying a 1 is than a 0.
  - C must be chosen independently of the data.
  - C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
  - People often choose $C = 1$, but without justification.
  - Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
    - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
    - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
  - Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves
  - Compute %1s and %0s correctly predicted for every possible value of C
  - Plot %1s by %0s
  - Overlay curves for several models on the same graph.
  - If one curve is above another the whole way, then that model dominates the other. It's better no matter your normative decision (about C)

(i) Binary variable predictions require a normative decision.
  - Let C be number of times more costly misclassifying a 1 is than a 0.
  - C must be chosen independently of the data.
  - C could come from your philosophical justification, survey of policy makers, a review of the literature, etc.
  - People often choose $C = 1$, but without justification.
  - Decision theory: choose $Y = 1$ when $\hat{\pi} > 1/(1 + C)$ and 0 otherwise.
    - If $C = 1$, predict $y = 1$ when $\hat{\pi} > 0.5$
    - If $C = 2$, predict $y = 1$ when $\hat{\pi} > 1/3$
  - Only with C chosen can we compute (a) % of 1s correctly predicted and (b) % of 0s correctly predicted, and (c) patterns in errors in different subsets of the data or forecasts.

(j) If you can't justify a choice for C, use ROC (receiver-operator characteristic) curves
  - Compute %1s and %0s correctly predicted for every possible value of C
  - Plot %1s by %0s
  - Overlay curves for several models on the same graph.
  - If one curve is above another the whole way, then that model dominates the other. It's better no matter your normative decision (about C)
  - Otherwise, one model is better than the other in only given specificed ranges of C (i.e., for only some normative perspectives).





In sample ROC, on left (from Gary King and Langche Zeng. "Improving Forecasts of State Failure," World Politics, Vol. 53, No. 4 (July, 2001): 623-58)

4. Cross-validation

4. Cross-validation
    (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets

4. Cross-validation
    (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
    (b) Useful for smaller data sets; real test sets are better.

4. Cross-validation
   (a) The idea: set aside k observations as the “test set”: evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
   (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)

4. Cross-validation
   (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
   (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables

4. Cross-validation
   (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
   (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables
   (a) The usual regression diagnostics

4. Cross-validation
    (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
    (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables
    (a) The usual regression diagnostics
    (b) E.G., plots of $e = y - \hat{y}$ by X, Y or $\hat{y}$

4. Cross-validation
   (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
   (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables
   (a) The usual regression diagnostics
   (b) E.G., plots of $e = y - \hat{y}$ by X, Y or $\hat{y}$
   (c) Check more than the means. E.g., plot e by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.

4. Cross-validation
    (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
    (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables
    (a) The usual regression diagnostics
    (b) E.G., plots of $e = y - \hat{y}$ by X, Y or $\hat{y}$
    (c) Check more than the means. E.g., plot e by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.
    (d) For graphics:

4. Cross-validation
   (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
   (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables
   (a) The usual regression diagnostics
   (b) E.G., plots of $e = y - \hat{y}$ by X, Y or $\hat{y}$
   (c) Check more than the means. E.g., plot e by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.
   (d) For graphics:
      - ⋆ transform bounded variables

4. Cross-validation
    (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
    (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables
    (a) The usual regression diagnostics
    (b) E.G., plots of $e = y - \hat{y}$ by X, Y or $\hat{y}$
    (c) Check more than the means. E.g., plot e by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.
    (d) For graphics:
        - ★ transform bounded variables
        - ★ transform heteroskedastic results

4. Cross-validation
   (a) The idea: set aside k observations as the "test set": evaluate; and then set aside another set of k observations. Repeat multiple times; report performance averaged over subsets
   (b) Useful for smaller data sets; real test sets are better.
5. Fit, in general: Look for all possible observable implications of a model, and compare to observations. (Think. Be creative here!)
6. Fit: continuous variables
   (a) The usual regression diagnostics
   (b) E.G., plots of $e = y - \hat{y}$ by X, Y or $\hat{y}$
   (c) Check more than the means. E.g., plot e by $\hat{y}$ and draw a line at 0 and at $\pm 1, 2$ se's. 66%, 95% of the observations should fall between the lines.
   (d) For graphics:
      - ★ transform bounded variables
      - ★ transform heteroskedastic results
      - ★ highlight key results; label everything

7. Fit: dichotomous variables

7. Fit: dichotomous variables

   (a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,. . . , $[0.9, 1]$.

7. Fit: dichotomous variables

(a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,. . . , $[0.9, 1]$.

(b) From the observations in each bin, compute (a) the mean predictions (probably near 0.05, 0.15, etc.) and (b) the average fraction of 1s.

7. Fit: dichotomous variables

(a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$,..., $[0.9, 1]$.

(b) From the observations in each bin, compute (a) the mean predictions (probably near 0.05, 0.15, etc.) and (b) the average fraction of 1s.

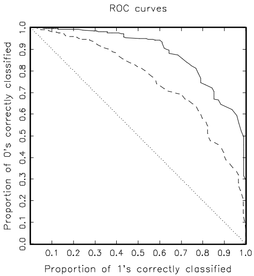
(c) Plot (a) by (b) and look for systematic deviation from 45° line.

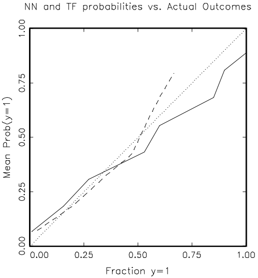
7. Fit: dichotomous variables

(a) Sort estimated probabilities into bins of say 0.1 width: $[0, 0.1)$, $[0.1, 0.2)$, $\ldots$, $[0.9, 1]$.

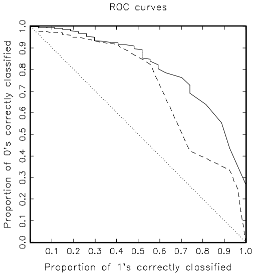
(b) From the observations in each bin, compute (a) the mean predictions (probably near 0.05, 0.15, etc.) and (b) the average fraction of 1s.

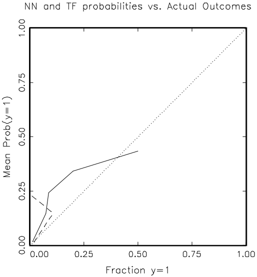
(c) Plot (a) by (b) and look for systematic deviation from 45° line.





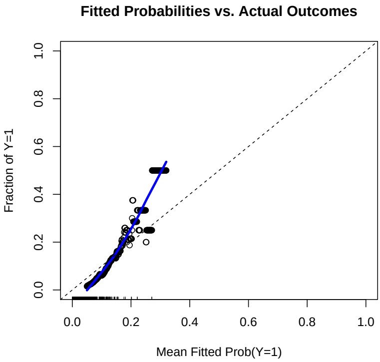
In sample calibration graph on right (from Gary King and Langche Zeng. "Improving Forecasts of State Failure," World Politics, Vol. 53, No. 4 (July, 2001): 623-58)

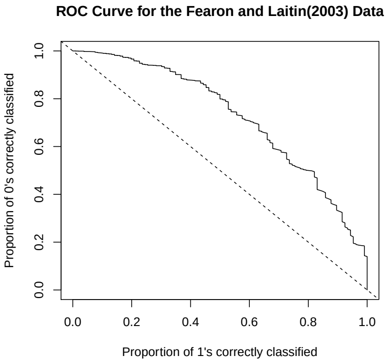




Out of sample calibration graph on right.

Out-of-Sample with Cross-Validation





New Developments: Separation plots

Greenhill, Ward and Sacks (2011)

Table 1 Sample Data

Country	Actual Outcome (y)	Fitted Value ($\hat{p}$)
A	0	0.774
B	0	0.364
C	1	0.997
D	0	0.728
E	1	0.961
F	1	0.422

New Developments: Separation plots

Greenhill, Ward and Sacks (2011)

Table 4 Rearrangement (and Coloring) of the Data Presented in Table 1 for Use in the Separation Plot

Country	Fitted Value ($\hat{p}$)	Actual Outcome (y)
B	0.364	0
F	0.422	1
D	0.728	0
A	0.774	0
E	0.961	1
C	0.997	1

New Developments: Separation plots

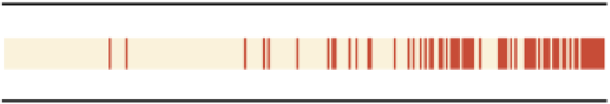
Greenhill, Ward and Sacks (2011)

FIGURE 2 Separation Plot Representing the Data Presented in Table 1

New Developments: Separation plots

Greenhill, Ward and Sacks (2011)

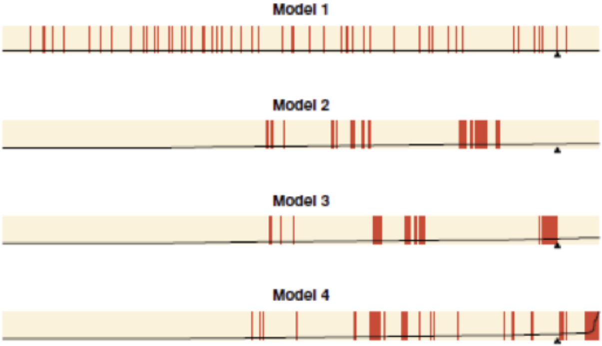
FIGURE 3 Separation Plot for a Larger Data Set



New Developments: Separation plots

Greenhill, Ward and Sacks (2011)

FIGURE 7 Separation Plots Used in the Development of a Model of Insurgency in the Asia-Pacific Region, 1998–2004

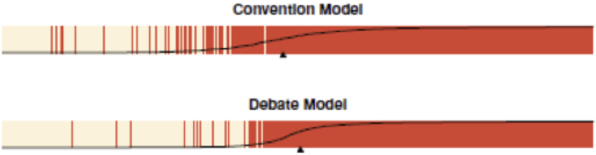


Note: For comparison, Models 1–4 have AUC scores of 0.500, 0.714, 0.744, and 0.816; Brier scores of 0.065, 0.063, 0.062, and 0.057; and ePCP scores of 0.869, 0.875, 0.876, and 0.887.

New Developments: Separation plots

Greenhill, Ward and Sacks (2011)

FIGURE 8 Separation Plots for the Hillygus and Jackman (2003) Models of Voting Intentions in the 2000 Presidential Election



Note: The upper plot shows the results of the survey conducted in the period following the party conventions, while the lower plot shows the results of the survey conducted after the presidential debates. Both models make an excellent fit to the data. (For comparison, the convention and debate models have AUC scores of 0.964 and 0.982; Brier scores of 0.071 and 0.045; and ePCP scores of 0.859 and 0.909.)

New Developments: Separation plots

Greenhill, Ward and Sacks (2011)

FIGURE 9 **Comparison of the Separation Plots Produced by Replicating Model 1 of Fearon and Laitin (2003) and by Reestimating the Model with Logged GDP per Capita as the Only Covariate**

Note: For comparison, Model 1 and the GDP-only model have AUC scores of 0.760 and 0.671; Brier scores of 0.016 and 0.016; and ePCP scores of 0.968 and 0.967.

New Developments

Hanmer and Kalkan distinguish between:

New Developments

Hanmer and Kalkan distinguish between:

- average case (most common)

New Developments

Hanmer and Kalkan distinguish between:

- average case (most common)
- observed value (they argue better)

The Case for Observed Case

The Case for Observed Case

Consider the case of voting for Bush in 2004.

The Case for Observed Case

Consider the case of voting for Bush in 2004. Average case is:

- a white

The Case for Observed Case

Consider the case of voting for Bush in 2004. Average case is:

- a white
- 48 year old woman

The Case for Observed Case

Consider the case of voting for Bush in 2004. Average case is:

- a white
- 48 year old woman
- identifies as independent and a political moderate

The Case for Observed Case

Consider the case of voting for Bush in 2004. Average case is:

- a white
- 48 year old woman
- identifies as independent and a political moderate
- with an associates degree

The Case for Observed Case

Consider the case of voting for Bush in 2004. Average case is:

- a white
- 48 year old woman
- identifies as independent and a political moderate
- with an associates degree
- believes economic performance has been constant

The Case for Observed Case

Consider the case of voting for Bush in 2004. Average case is:

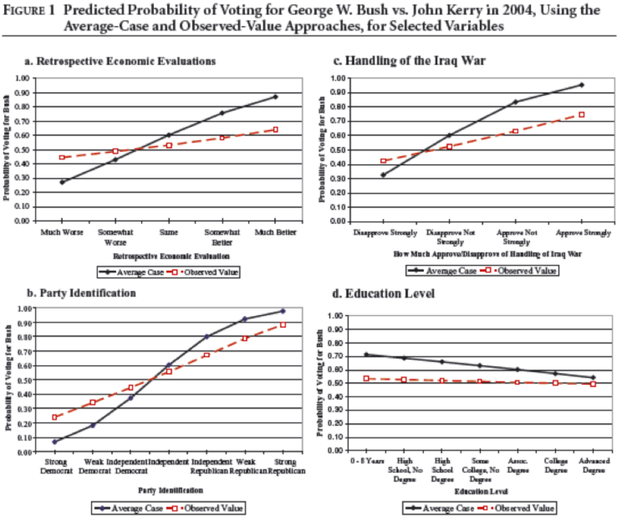
- a white
- 48 year old woman
- identifies as independent and a political moderate
- with an associates degree
- believes economic performance has been constant
- disapproves of the Iraq war but not strongly

The Case for Observed Case

Consider the case of voting for Bush in 2004. Average case is:

- a white
- 48 year old woman
- identifies as independent and a political moderate
- with an associates degree
- believes economic performance has been constant
- disapproves of the Iraq war but not strongly
- with an income between $45K and $50K

The Case for Observed Case



Notes: Data are from the 2004 ANES, using respondents who first answered the standard turnout question. Results are based on estimates from the model reported in SI Section B Table 1.

The Case for Observed Case

Try to come up with an argument for why the average-case method will tend to produce bigger changes than the observed-case method.

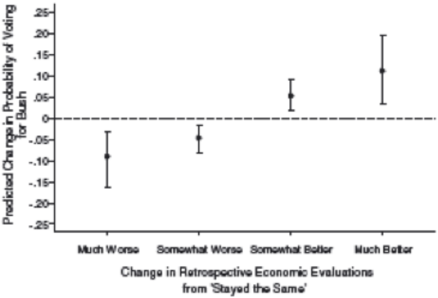
The Case for Observed Case

Try to come up with an argument for why the average-case method will tend to produce bigger changes than the observed-case method.

Average cases are likely to be in the "middle" of the data where the predicted probabilities are changing the fastest. Think about Bill O'Reilly and Rachel Maddow. They are going to show up in the observed-case method but not the average-case method.

The Case for Observed Case

FIGURE 3 Predicted Effects (First Differences) of Changing Retrospective Economic Evaluations on the Probability of Voting for George W. Bush vs. John Kerry in 2004, Using the Observed-Value Approach, with 95% Confidence Intervals



Notes: Data are from the 2004 ANES, using respondents who first answered the standard turnout question. Results are from statistical simulation.

UPM Remainder of Chapter 5.

Optionally: Greenhill et al. 2011, Hamner and Kalkan 2013

Also Helpful: Mood 2010, Berry, DeMeritt and Esarey 2010

Modeling Ordered Outcomes

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
  - "Likert scale" in survey questions ("strongly agree", "agree", etc.)

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
  - "Likert scale" in survey questions ("strongly agree", "agree", etc.)
  - Party positions (extreme left, center left, center, right, extreme right)

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
  - "Likert scale" in survey questions ("strongly agree", "agree", etc.)
  - Party positions (extreme left, center left, center, right, extreme right)
  - Levels of democracy (autocracy, anocracy, democracy)

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
  - "Likert scale" in survey questions ("strongly agree", "agree", etc.)
  - Party positions (extreme left, center left, center, right, extreme right)
  - Levels of democracy (autocracy, anocracy, democracy)
  - Health status (healthy, sick, dying, dead)

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
  - "Likert scale" in survey questions ("strongly agree", "agree", etc.)
  - Party positions (extreme left, center left, center, right, extreme right)
  - Levels of democracy (autocracy, anocracy, democracy)
  - Health status (healthy, sick, dying, dead)
- Why not use continuous outcome models?

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
  - "Likert scale" in survey questions ("strongly agree", "agree", etc.)
  - Party positions (extreme left, center left, center, right, extreme right)
  - Levels of democracy (autocracy, anocracy, democracy)
  - Health status (healthy, sick, dying, dead)
- Why not use continuous outcome models?
  $\longrightarrow$ Don't want to assume equal distances between levels

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
  - "Likert scale" in survey questions ("strongly agree", "agree", etc.)
  - Party positions (extreme left, center left, center, right, extreme right)
  - Levels of democracy (autocracy, anocracy, democracy)
  - Health status (healthy, sick, dying, dead)
- Why not use continuous outcome models?
  $\longrightarrow$ Don't want to assume equal distances between levels
- Why not use categorical outcome models?

Modeling Ordered Outcomes

- Suppose that we have an outcome which is one of J choices that are ordered in a substantively meaningful way
- Examples:
    - "Likert scale" in survey questions ("strongly agree", "agree", etc.)
    - Party positions (extreme left, center left, center, right, extreme right)
    - Levels of democracy (autocracy, anocracy, democracy)
    - Health status (healthy, sick, dying, dead)
- Why not use continuous outcome models?
  $\longrightarrow$ Don't want to assume equal distances between levels
- Why not use categorical outcome models?
  $\longrightarrow$ Don't want to waste information about ordering

Ordered Dependent Variable Models

Ordered Dependent Variable Models

The model

Ordered Dependent Variable Models

The model

$$Y_i^* \sim \mathrm{STN}(y_i^*|\mu_i)$$

Ordered Dependent Variable Models

The model

$$Y_i^* \sim \mathsf{STN}(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Ordered Dependent Variable Models

The model

$$
\begin{aligned}
Y_i^* &\sim \mathsf{STN}(y_i^*|\mu_i) \\
\mu_i &= x_i\beta
\end{aligned}
$$

Observation mechanism

Ordered Dependent Variable Models

The model

$$Y_i^* \sim \mathrm{STN}(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Observation mechanism

$$y_{ij} =$$

Ordered Dependent Variable Models

The model

$$Y_i^* \sim \mathsf{STN}(y_i^*|\mu_i)$$
$$\mu_i = x_i\beta$$

Observation mechanism

$$y_{ij} = \begin{cases} 1 & \text{if } \tau_{j-1,i} \leq y_i^* \leq \tau_{j,i} \\ 0 & \text{otherwise} \end{cases}$$

Ordered Dependent Variable Models

The model

$$
\begin{aligned}
Y_i^* &\sim \mathrm{STN}(y_i^*|\mu_i) \\
\mu_i &= x_i\beta
\end{aligned}
$$

Observation mechanism

$$
y_{ij} = \begin{cases} 1 & \text{if } \tau_{j-1,i} \leq y_i^* \leq \tau_{j,i} \\ 0 & \text{otherwise} \end{cases}
$$

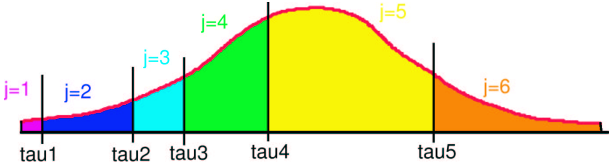
Ordered Dependent Variable Models

The model

$$Y_i^* \sim \text{STN}(y_i^* | \mu_i)$$
$$\mu_i = x_i \beta$$

Observation mechanism

$$y_{ij} = \begin{cases} 1 & \text{if } \tau_{j-1,i} \leq y_i^* \leq \tau_{j,i} \\ 0 & \text{otherwise} \end{cases}$$



Ordered Logit and Probit Models

- Again, the latent variable representation: $Y_i^* \ = \ X_i^\top \beta + \epsilon_i$

Ordered Logit and Probit Models

- Again, the latent variable representation: $Y_i^* = X_i^\top \beta + \epsilon_i$
- Assume that Y_i^* gives rise to Y_i based on the following scheme:

$$Y_i = \begin{cases} 1 & \text{if} \quad -\infty(=\psi_0) < Y_i^* \leq \psi_1, \\ 2 & \text{if} \quad \psi_1 < Y_i^* \leq \psi_2, \\ \vdots & \qquad\qquad \vdots \\ J & \text{if} \quad \psi_{J-1} < Y_i^* \leq \infty(=\psi_J) \end{cases}$$

where $\psi_1, ..., \psi_{J-1}$ are the threshold parameters to be estimated

Ordered Logit and Probit Models

- Again, the latent variable representation: $Y_i^* = X_i^\top \beta + \epsilon_i$
- Assume that Y_i^* gives rise to Y_i based on the following scheme:

$$Y_i = \begin{cases} 1 & \text{if} \quad -\infty(=\psi_0) < Y_i^* \leq \psi_1, \\ 2 & \text{if} \quad \psi_1 < Y_i^* \leq \psi_2, \\ \vdots & \qquad \vdots \\ J & \text{if} \quad \psi_{J-1} < Y_i^* \leq \infty(=\psi_J) \end{cases}$$

  where $\psi_1, ..., \psi_{J-1}$ are the threshold parameters to be estimated
- If X_i contains an intercept, one of the ψ's must be fixed for identifiability (typically $\psi_1 = 0$)

Ordered Logit and Probit Models

- Again, the latent variable representation: $Y_i^* \;=\; X_i^\top \beta + \epsilon_i$
- Assume that Y_i^* gives rise to Y_i based on the following scheme:

$$Y_i \;=\; \begin{cases} 1 & \text{if} \quad -\infty(=\psi_0) \;<\; Y_i^* \;\leq\; \psi_1, \\ 2 & \text{if} \quad \psi_1 \;<\; Y_i^* \;\leq\; \psi_2, \\ \vdots & \qquad\qquad\qquad \vdots \\ J & \text{if} \quad \psi_{J-1} \;<\; Y_i^* \;\leq\; \infty(=\psi_J) \end{cases}$$

  where $\psi_1, ..., \psi_{J-1}$ are the threshold parameters to be estimated
- If X_i contains an intercept, one of the ψ's must be fixed for identifiability (typically $\psi_1 = 0$)
- $\epsilon_j \overset{\text{i.i.d.}}{\sim}$ logistic $\Rightarrow$ the ordered logit model:

$$\Pr(Y_i \leq j \mid X_i) \;=\; \frac{\exp(\psi_j - X_i^\top \beta)}{1 + \exp(\psi_j - X_i^\top \beta)}$$

Ordered Logit and Probit Models

- Again, the latent variable representation: $Y_i^* = X_i^\top \beta + \epsilon_i$
- Assume that Y_i^* gives rise to Y_i based on the following scheme:

$$Y_i = \begin{cases} 1 & \text{if} \quad -\infty(=\psi_0) < Y_i^* \leq \psi_1, \\ 2 & \text{if} \quad \psi_1 < Y_i^* \leq \psi_2, \\ \vdots & \qquad\qquad \vdots \\ J & \text{if} \quad \psi_{J-1} < Y_i^* \leq \infty(=\psi_J) \end{cases}$$

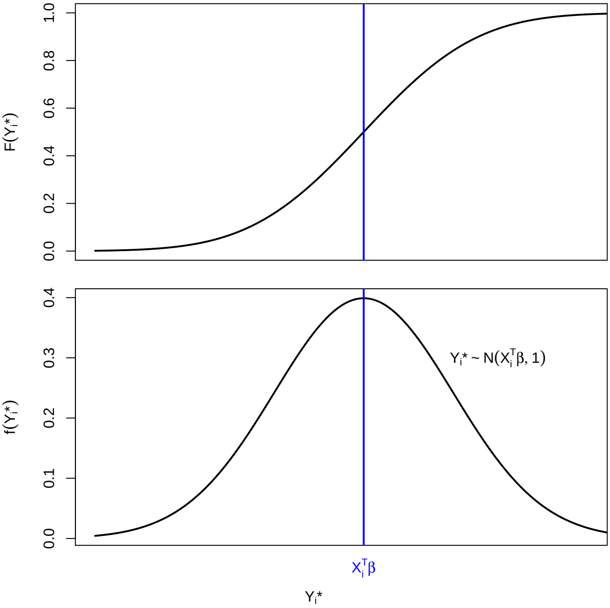
  where $\psi_1, ..., \psi_{J-1}$ are the threshold parameters to be estimated
- If X_i contains an intercept, one of the ψ's must be fixed for identifiability (typically $\psi_1 = 0$)
- $\epsilon_j \overset{\text{i.i.d.}}{\sim}$ logistic $\Rightarrow$ the ordered logit model:

$$\Pr(Y_i \leq j \mid X_i) = \frac{\exp(\psi_j - X_i^\top \beta)}{1 + \exp(\psi_j - X_i^\top \beta)}$$

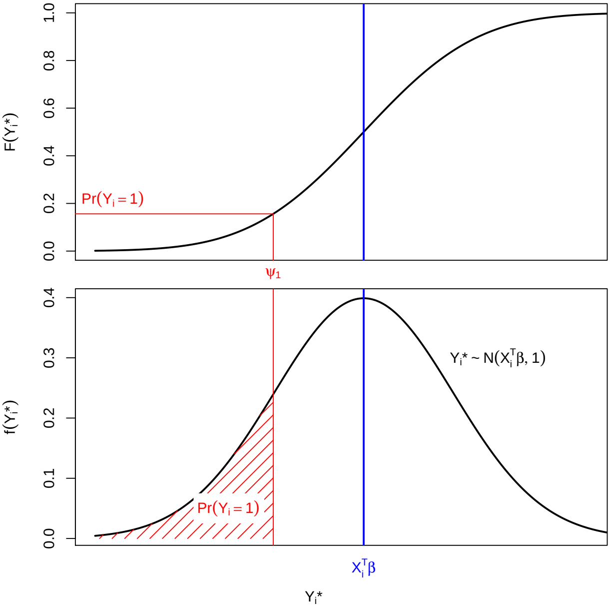
- $\epsilon_j \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \Rightarrow$ the ordered probit model:

$$\Pr(Y_i \leq j \mid X_i) = \Phi\left(\psi_j - X_i^\top \beta\right)$$

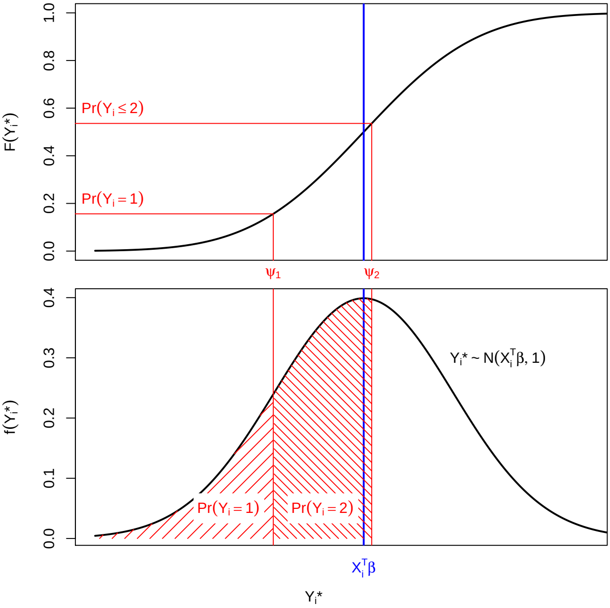
Ordered Logit and Probit Models



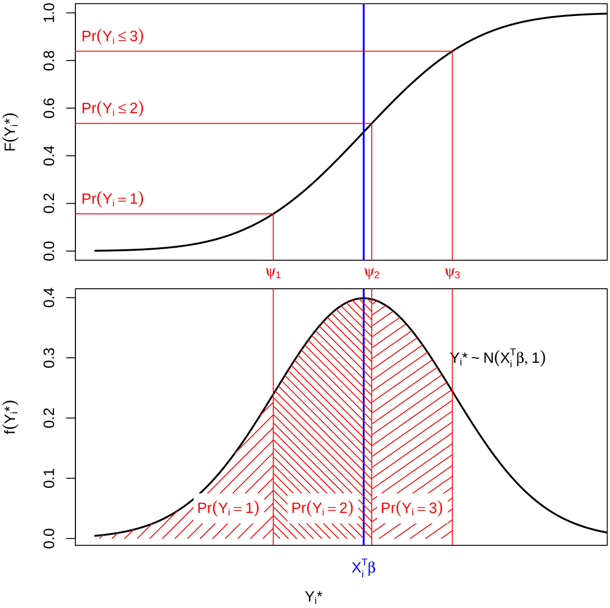
Ordered Logit and Probit Models



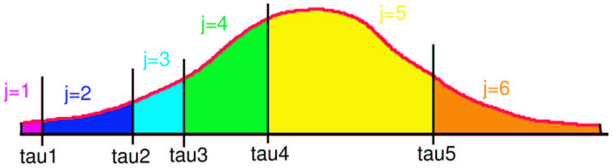
Ordered Logit and Probit Models



Ordered Logit and Probit Models

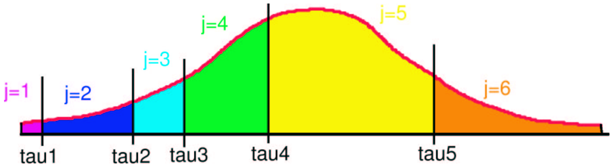


Ordered Dependent Variable Models: Connections



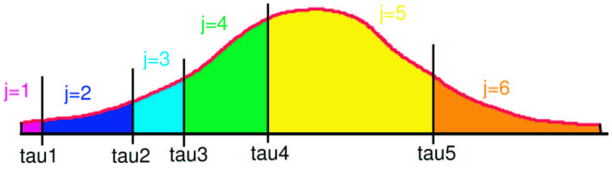
Ordered Dependent Variable Models: Connections

1. $Y_i^* \sim \text{STL}(y_i^*|\mu_i) \rightarrow$ ordinal logit
   $Y_i^* \sim \text{STN}(y_i^*|\mu_i) \rightarrow$ ordinal probit



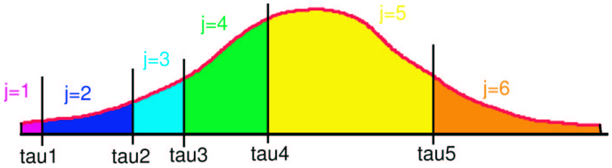
Ordered Dependent Variable Models: Connections

1. $Y_i^* \sim \text{STL}(y_i^*|\mu_i) \rightarrow$ ordinal logit
   $Y_i^* \sim \text{STN}(y_i^*|\mu_i) \rightarrow$ ordinal probit
2. Alternate representation: dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.



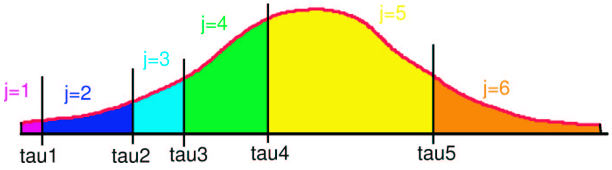
Ordered Dependent Variable Models: Connections

1. $Y_i^* \sim \text{STL}(y_i^*|\mu_i) \rightarrow$ ordinal logit
   $Y_i^* \sim \text{STN}(y_i^*|\mu_i) \rightarrow$ ordinal probit
2. Alternate representation: dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.
3. If Y_i^* is observed, the probit version is a linear-normal regression model



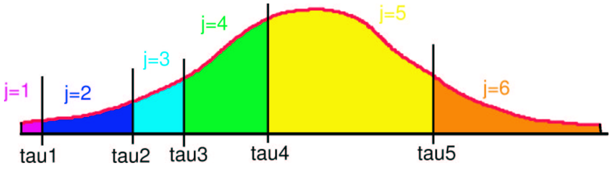
Ordered Dependent Variable Models: Connections

1. $Y_i^* \sim \text{STL}(y_i^*|\mu_i) \rightarrow$ ordinal logit
   $Y_i^* \sim \text{STN}(y_i^*|\mu_i) \rightarrow$ ordinal probit
2. Alternate representation: dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.
3. If Y_i^* is observed, the probit version is a linear-normal regression model
4. If a dichotomous realization of Y^* is observed, its a logit/probit model



Ordered Dependent Variable Models: Connections

1. $Y_i^* \sim \text{STL}(y_i^*|\mu_i) \rightarrow$ ordinal logit
   $Y_i^* \sim \text{STN}(y_i^*|\mu_i) \rightarrow$ ordinal probit
2. Alternate representation: dichotomous variable Y_{ji} for each category j, only one of which is 1; the others are 0.
3. If Y_i^* is observed, the probit version is a linear-normal regression model
4. If a dichotomous realization of Y^* is observed, its a logit/probit model
5. This is the same model, and the same parameters are being estimated; only the observation mechanism differs.



Deriving the likelihood function

Deriving the likelihood function

First the probability of each observation, then the joint probability.

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\Pr(Y_{ij} = 1) = \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j)$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}\Pr(Y_{ij} = 1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\ &= \int_{\tau_{j-1}}^{\tau_j} \mathrm{STN}(y_i^* | \mu_i) dy_i^*\end{aligned}$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}\Pr(Y_{ij} = 1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\ &= \int_{\tau_{j-1}}^{\tau_j} \text{STN}(y_i^* | \mu_i) dy_i^* \\ &= F_{stn}(\tau_j | \mu_i) - F_{stn}(\tau_{j-1} | \mu_i)\end{aligned}$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ij} = 1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \text{STN}(y_i^* | \mu_i) dy_i^* \\
&= F_{stn}(\tau_j | \mu_i) - F_{stn}(\tau_{j-1} | \mu_i) \\
&= F_{stn}(\tau_j | x_i \beta) - F_{stn}(\tau_{j-1} | x_i \beta)
\end{aligned}$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ij}=1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \mathrm{STN}(y_i^*|\mu_i)dy_i^* \\
&= F_{stn}(\tau_j|\mu_i) - F_{stn}(\tau_{j-1}|\mu_i) \\
&= F_{stn}(\tau_j|x_i\beta) - F_{stn}(\tau_{j-1}|x_i\beta)
\end{aligned}$$

The joint probability is then:

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ij} = 1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \text{STN}(y_i^*|\mu_i) dy_i^* \\
&= F_{stn}(\tau_j|\mu_i) - F_{stn}(\tau_{j-1}|\mu_i) \\
&= F_{stn}(\tau_j|x_i\beta) - F_{stn}(\tau_{j-1}|x_i\beta)
\end{aligned}$$

The joint probability is then:

$$P(Y) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \Pr(Y_{ij} = 1)^{y_{ij}} \right]$$

Deriving the likelihood function

First the probability of each observation, then the joint probability.

$$\begin{aligned}
\Pr(Y_{ij} = 1) &= \Pr(\tau_{j-1} \leq Y_i^* \leq \tau_j) \\
&= \int_{\tau_{j-1}}^{\tau_j} \text{STN}(y_i^* | \mu_i) dy_i^* \\
&= F_{stn}(\tau_j | \mu_i) - F_{stn}(\tau_{j-1} | \mu_i) \\
&= F_{stn}(\tau_j | x_i \beta) - F_{stn}(\tau_{j-1} | x_i \beta)
\end{aligned}$$

The joint probability is then:

$$P(Y) = \prod_{i=1}^{n} \left[\prod_{j=1}^{J} \Pr(Y_{ij} = 1)^{y_{ij}} \right]$$

Bracketed portion has only one active component for each i.

Deriving the likelihood function

The Log-likelihood:

Deriving the likelihood function

The Log-likelihood:

$$\ln L(\beta, \tau | y) = \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln \Pr(Y_{ij} = 1)$$

Deriving the likelihood function

The Log-likelihood:

$$
\begin{aligned}
\ln L(\beta, \tau | y) &= \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln \Pr(Y_{ij} = 1) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln \left[F_{stn}(\tau_j | x_i \beta) - F_{stn}(\tau_{j-1} | x_i \beta) \right]
\end{aligned}
$$

Deriving the likelihood function

The Log-likelihood:

$$\begin{aligned}\ln L(\beta, \tau | y) &= \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln \Pr(Y_{ij} = 1) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{J} y_{ij} \ln \left[F_{stn}(\tau_j | x_i \beta) - F_{stn}(\tau_{j-1} | x_i \beta)\right]\end{aligned}$$

(Constraints during optimization make this more complicated to do from scratch: $\tau_{j-1} < \tau_j$, $\forall j$)

Interpretation: Ordinal Probit

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.
3. One first difference has an effect on all J probabilities.

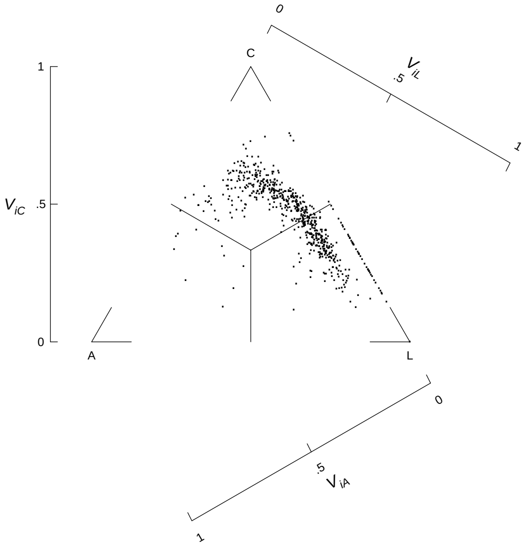
Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.
3. One first difference has an effect on all J probabilities.
4. When one probability goes up, at least one of the others must go down.

Interpretation: Ordinal Probit

1. Coefficients are the linear effect of X on Y^* in standard deviation units
2. Predictions from the model are J probabilities that sum to 1.
3. One first difference has an effect on all J probabilities.
4. When one probability goes up, at least one of the others must go down.
5. Can use ternary diagrams if $J = 3$

Representing 3 variables, with $Y_j \in [0, 1]$ and $\sum_{j=1}^{3} Y_j = 1$



Calculating Quantities of Interest

- Predicted probability:

$$\begin{aligned}\pi_{ij}(X_i) &\equiv \Pr(Y_i = j \mid X_i) = \Pr(Y_i \leq j \mid X_i) - \Pr(Y_i \leq j-1 \mid X_i) \\ &= \begin{cases} \frac{\exp(\psi_j - X_i^\top \beta)}{1+\exp(\psi_j - X_i^\top \beta)} - \frac{\exp(\psi_{j-1} - X_i^\top \beta)}{1+\exp(\psi_{j-1} - X_i^\top \beta)} & \text{for logit} \\ \Phi\left(\psi_j - X_i^\top \beta\right) - \Phi\left(\psi_{j-1} - X_i^\top \beta\right) & \text{for probit} \end{cases}\end{aligned}$$

Calculating Quantities of Interest

- Predicted probability:

$$\begin{aligned}\pi_{ij}(X_i) &\equiv \Pr(Y_i = j \mid X_i) \;=\; \Pr(Y_i \le j \mid X_i) - \Pr(Y_i \le j-1 \mid X_i) \\ &= \begin{cases} \frac{\exp(\psi_j - X_i^\top \beta)}{1+\exp(\psi_j - X_i^\top \beta)} - \frac{\exp(\psi_{j-1} - X_i^\top \beta)}{1+\exp(\psi_{j-1} - X_i^\top \beta)} & \text{for logit} \\ \Phi\left(\psi_j - X_i^\top \beta\right) - \Phi\left(\psi_{j-1} - X_i^\top \beta\right) & \text{for probit} \end{cases}\end{aligned}$$

- ATE (APE): $\tau_j \;=\; \mathbb{E}\left[\pi_j(T_i = 1, W_i) - \pi_j(T_i = 0, W_i)\right]$

Calculating Quantities of Interest

- Predicted probability:

$$\begin{aligned}\pi_{ij}(X_i) &\equiv \Pr(Y_i = j \mid X_i) \;=\; \Pr(Y_i \leq j \mid X_i) - \Pr(Y_i \leq j-1 \mid X_i) \\ &= \begin{cases} \frac{\exp(\psi_j - X_i^\top \beta)}{1+\exp(\psi_j - X_i^\top \beta)} - \frac{\exp(\psi_{j-1} - X_i^\top \beta)}{1+\exp(\psi_{j-1} - X_i^\top \beta)} & \text{for logit} \\ \Phi\left(\psi_j - X_i^\top \beta\right) - \Phi\left(\psi_{j-1} - X_i^\top \beta\right) & \text{for probit} \end{cases}\end{aligned}$$

- ATE (APE): $\tau_j \;=\; \mathbb{E}\left[\pi_j(T_i = 1, W_i) - \pi_j(T_i = 0, W_i)\right]$

- Estimate β and ψ via MLE, plug the estimates in, replace $\mathbb{E}$ with $\frac{1}{n}\sum$, and compute CI by delta or MC or bootstrap

Calculating Quantities of Interest

- Predicted probability:

$$\begin{aligned}
\pi_{ij}(X_i) &\equiv \Pr(Y_i = j \mid X_i) = \Pr(Y_i \leq j \mid X_i) - \Pr(Y_i \leq j-1 \mid X_i) \\
&= \begin{cases} \frac{\exp(\psi_j - X_i^\top \beta)}{1+\exp(\psi_j - X_i^\top \beta)} - \frac{\exp(\psi_{j-1} - X_i^\top \beta)}{1+\exp(\psi_{j-1} - X_i^\top \beta)} & \text{for logit} \\ \Phi\left(\psi_j - X_i^\top \beta\right) - \Phi\left(\psi_{j-1} - X_i^\top \beta\right) & \text{for probit} \end{cases}
\end{aligned}$$

- ATE (APE): $\tau_j = \mathbb{E}\left[\pi_j(T_i = 1, W_i) - \pi_j(T_i = 0, W_i)\right]$

- Estimate β and ψ via MLE, plug the estimates in, replace $\mathbb{E}$ with $\frac{1}{n}\sum$, and compute CI by delta or MC or bootstrap

- Note that $X_i^\top \beta$ appears both before and after the minus sign in π_{ij}

Calculating Quantities of Interest

- Predicted probability:

$$\begin{aligned} \pi_{ij}(X_i) &\equiv \Pr(Y_i = j \mid X_i) = \Pr(Y_i \leq j \mid X_i) - \Pr(Y_i \leq j-1 \mid X_i) \\ &= \begin{cases} \frac{\exp(\psi_j - X_i^\top \beta)}{1+\exp(\psi_j - X_i^\top \beta)} - \frac{\exp(\psi_{j-1} - X_i^\top \beta)}{1+\exp(\psi_{j-1} - X_i^\top \beta)} & \text{for logit} \\ \Phi\left(\psi_j - X_i^\top \beta\right) - \Phi\left(\psi_{j-1} - X_i^\top \beta\right) & \text{for probit} \end{cases} \end{aligned}$$

- ATE (APE): $\tau_j = \mathbb{E}\left[\pi_j(T_i = 1, W_i) - \pi_j(T_i = 0, W_i)\right]$

- Estimate β and ψ via MLE, plug the estimates in, replace $\mathbb{E}$ with $\frac{1}{n}\sum$, and compute CI by delta or MC or bootstrap

- Note that $X_i^\top \beta$ appears both before and after the minus sign in π_{ij} $\longrightarrow$ Direction of effect of X_i on Y_{ij} is ambiguous (except top and bottom)

Calculating Quantities of Interest

- Predicted probability:

$$\begin{aligned}\pi_{ij}(X_i) &\equiv \Pr(Y_i = j \mid X_i) = \Pr(Y_i \le j \mid X_i) - \Pr(Y_i \le j-1 \mid X_i) \\ &= \begin{cases} \frac{\exp(\psi_j - X_i^\top \beta)}{1+\exp(\psi_j - X_i^\top \beta)} - \frac{\exp(\psi_{j-1} - X_i^\top \beta)}{1+\exp(\psi_{j-1} - X_i^\top \beta)} & \text{for logit} \\ \Phi\left(\psi_j - X_i^\top \beta\right) - \Phi\left(\psi_{j-1} - X_i^\top \beta\right) & \text{for probit} \end{cases}\end{aligned}$$

- ATE (APE): $\tau_j = \mathbb{E}\left[\pi_j(T_i = 1, W_i) - \pi_j(T_i = 0, W_i)\right]$

- Estimate β and ψ via MLE, plug the estimates in, replace $\mathbb{E}$ with $\frac{1}{n}\sum$, and compute CI by delta or MC or bootstrap

- Note that $X_i^\top \beta$ appears both before and after the minus sign in π_{ij}
  $\longrightarrow$ Direction of effect of X_i on Y_{ij} is ambiguous (except top and bottom)
  $\longrightarrow$ Again, calculate quantities of interest, not just coefficients

Example: Immigration and Media Priming

Brader, Valentino and Suhay (2008):

- Y_i: Ordinal response to question about increasing immigration
- T_{1i}, T_{2i}: Media cues (immigrant ethnicity $\times$ story tone)
- W_i: Respondent age and income

Example: Immigration and Media Priming

Brader, Valentino and Suhay (2008):

- Y_i: Ordinal response to question about increasing immigration
- T_{1i}, T_{2i}: Media cues (immigrant ethnicity $\times$ story tone)
- W_i: Respondent age and income

Estimated coefficients:

```
Coefficients:
          Value  s.e.   t
tone       0.27  0.32  0.85
eth       -0.33  0.32 -1.02
ppage      0.01  0.02  1.40
ppincimp   0.00  0.03  0.06
tone:eth   0.90  0.46  2.16

Intercepts:
    Value  s.e.    t
1|2 -1.93  0.58  -3.32
2|3 -0.12  0.55  -0.21
3|4  1.12  0.56   2.01
```

Example: Immigration and Media Priming

Brader, Valentino and Suhay (2008):

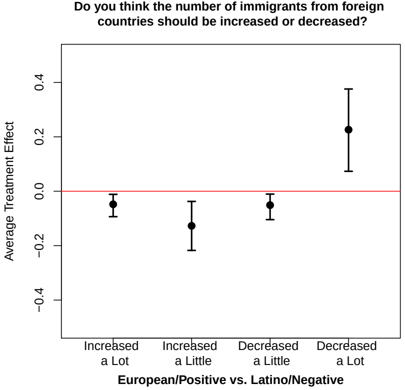
- Y_i: Ordinal response to question about increasing immigration
- T_{1i}, T_{2i}: Media cues (immigrant ethnicity $\times$ story tone)
- W_i: Respondent age and income

Estimated coefficients:

```
Coefficients:
          Value  s.e.   t
tone       0.27  0.32  0.85
eth       -0.33  0.32 -1.02
ppage      0.01  0.02  1.40
ppincimp   0.00  0.03  0.06
tone:eth   0.90  0.46  2.16

Intercepts:
    Value  s.e.    t
1|2 -1.93  0.58  -3.32
2|3 -0.12  0.55  -0.21
3|4  1.12  0.56   2.01
```

ATE:



Example: Peer Bereavement

Andersen, Silver, Koperwas, Stewart and Kirschbaum (2013):

Example: Peer Bereavement

Andersen, Silver, Koperwas, Stewart and Kirschbaum (2013):
Study of response by college students to a sequence of 14 peer deaths in one academic year.

- Y_i: Severity of acute reaction (1-5 scale)
- X_i: gender, number of peers known, media exposure

Example: Peer Bereavement

Andersen, Silver, Koperwas, Stewart and Kirschbaum (2013):
Study of response by college students to a sequence of 14 peer deaths in one academic year.

- Y_i: Severity of acute reaction (1-5 scale)
- X_i: gender, number of peers known, media exposure

	Coef	SE	CI	P-value	RR Strong (95% CI)	RR Extreme (95% CI)
Female	0.89	0.23	(0.44, 1.35)	0	2.71 (1.51, 4.82)	13.69 (2.84, 45.55)
Num. Peers Known	0.54	0.16	(0.23, 0.85)	0	1.43 (1.15, 1.82)	3.53 (1.6, 7.25)
Media Exposure	0.25	0.06	(0.13, 0.36)	0	1.17 (1.08, 1.3)	1.73 (1.31, 2.38)

Example: Peer Bereavement

Andersen, Silver, Koperwas, Stewart and Kirschbaum (2013):
Study of response by college students to a sequence of 14 peer deaths in one academic year.

- Y_i: Severity of acute reaction (1-5 scale)
- X_i: gender, number of peers known, media exposure

	Coef	SE	CI	P-value	RR Strong (95% CI)	RR Extreme (95% CI)
Female	0.89	0.23	(0.44, 1.35)	0	2.71 (1.51, 4.82)	13.69 (2.84, 45.55)
Num. Peers Known	0.54	0.16	(0.23, 0.85)	0	1.43 (1.15, 1.82)	3.53 (1.6, 7.25)
Media Exposure	0.25	0.06	(0.13, 0.36)	0	1.17 (1.08, 1.3)	1.73 (1.31, 2.38)

The Risk Ratio measures the relative probability of being in the outcome category based on different values of the independent variable.

Example: Peer Bereavement

Andersen, Silver, Koperwas, Stewart and Kirschbaum (2013):
Study of response by college students to a sequence of 14 peer deaths in one academic year.

- Y_i: Severity of acute reaction (1-5 scale)
- X_i: gender, number of peers known, media exposure

	Coef	SE	CI	P-value	RR Strong (95% CI)	RR Extreme (95% CI)
Female	0.89	0.23	(0.44, 1.35)	0	2.71 (1.51, 4.82)	13.69 (2.84, 45.55)
Num. Peers Known	0.54	0.16	(0.23, 0.85)	0	1.43 (1.15, 1.82)	3.53 (1.6, 7.25)
Media Exposure	0.25	0.06	(0.13, 0.36)	0	1.17 (1.08, 1.3)	1.73 (1.31, 2.38)

The Risk Ratio measures the relative probability of being in the outcome category based on different values of the independent variable. Thus the RR for the Strong Reaction category for Female can be understood as

$$RR_{\text{Strong}} = \frac{\Pr(\text{Strong Reaction}|\text{Female})}{\Pr(\text{Strong Reaction}|\text{Male})}$$

Example: Peer Bereavement

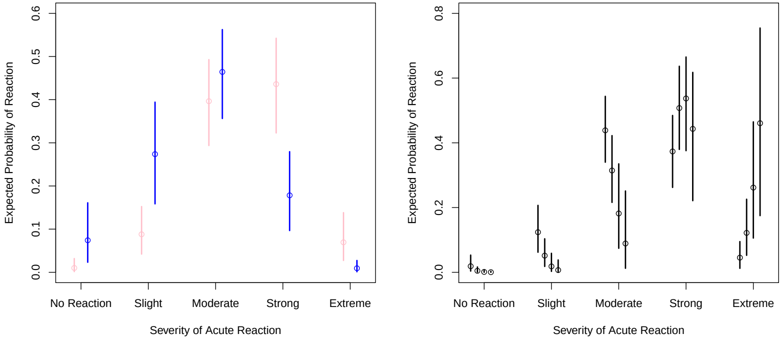


Figure: This plot shows the expected probabilities of being in each category of reaction given gender (left) and knowing 1 to 4 people (right) with 95% confidence intervals.

Ordered Categorical Conclusions

- Straightforward to derive from latent variable representation

Ordered Categorical Conclusions

- Straightforward to derive from latent variable representation
- Ordered probit is often easier to work with,

Ordered Categorical Conclusions

- Straightforward to derive from latent variable representation
- Ordered probit is often easier to work with,

Ordered Categorical Conclusions

- Straightforward to derive from latent variable representation
- Ordered probit is often easier to work with, ordered logit has a nice interpretation as a proportional odds model

$$\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \lambda_j \exp(x_i \beta)$$

where γ_{ij} is the cumulative probability and λ_j is the baseline odds.

Ordered Categorical Conclusions

- Straightforward to derive from latent variable representation
- Ordered probit is often easier to work with, ordered logit has a nice interpretation as a proportional odds model

$$\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \lambda_j \exp(x_i \beta)$$

where γ_{ij} is the cumulative probability and λ_j is the baseline odds. Covariates raise or lower the odds of a response in category j or below.

Ordered Categorical Conclusions

- Straightforward to derive from latent variable representation
- Ordered probit is often easier to work with, ordered logit has a nice interpretation as a proportional odds model

$$\log \frac{\gamma_{ij}}{1-\gamma_{ij}} = \lambda_j \exp(x_i \beta)$$

  where γ_{ij} is the cumulative probability and λ_j is the baseline odds. Covariates raise or lower the odds of a response in category j or below.
- Visualization and appropriate quantities of interest can be tricky. Let the substance guide you.

Multinomial Logit

Multinomial Logit

Multinomial Logit

- Sometimes we encounter unordered categories (choose a Ph.D. Sociology, Politics, Psychology, Statistics).

Multinomial Logit

- Sometimes we encounter unordered categories (choose a Ph.D. Sociology, Politics, Psychology, Statistics).
- We can generalize the logit model to two choices to get the multinomial logit model

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid X_i) \;=\; \frac{\exp(X_i^\top \beta_j)}{\sum_{k=1}^{J} \exp(X_i^\top \beta_k)},$$

where X_i = individual-specific characteristics of unit i

Multinomial Logit

- Sometimes we encounter unordered categories (choose a Ph.D. Sociology, Politics, Psychology, Statistics).
- We can generalize the logit model to two choices to get the multinomial logit model

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid X_i) \;=\; \frac{\exp(X_i^\top \beta_j)}{\sum_{k=1}^{J} \exp(X_i^\top \beta_k)},$$

  where X_i = individual-specific characteristics of unit i
- category-specific set of coefficients (one category omitted for identification)

Multinomial Logit

- Sometimes we encounter unordered categories (choose a Ph.D. Sociology, Politics, Psychology, Statistics).
- We can generalize the logit model to two choices to get the multinomial logit model

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid X_i) \;=\; \frac{\exp(X_i^\top \beta_j)}{\sum_{k=1}^{J} \exp(X_i^\top \beta_k)},$$

  where X_i = individual-specific characteristics of unit i
- category-specific set of coefficients (one category omitted for identification)
- Multinomial logit also has a latent variable interpretation: make choice with greatest utility Y_{ij}^*. When the stochastic component on the utility is $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim}$ type I extreme value distribution, multinomial logit is implied.

Multinomial Logit

- Sometimes we encounter unordered categories (choose a Ph.D. Sociology, Politics, Psychology, Statistics).
- We can generalize the logit model to two choices to get the multinomial logit model

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid X_i) \;=\; \frac{\exp(X_i^\top \beta_j)}{\sum_{k=1}^{J} \exp(X_i^\top \beta_k)},$$

  where X_i = individual-specific characteristics of unit i
- category-specific set of coefficients (one category omitted for identification)
- Multinomial logit also has a latent variable interpretation: make choice with greatest utility Y_{ij}^*. When the stochastic component on the utility is $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim}$ type I extreme value distribution, multinomial logit is implied.
- Coefficients are relative to a baseline category- so again we want to compute quantities of interest for interpretation.

Independence of Irrelevant Alternatives

Independence of Irrelevant Alternatives

Independence of Irrelevant Alternatives

- Multinomial logit assumes iid errors in the latent utility model, this implies that unobserved factors affecting Y_{ij}^* are unrelated to those affecting Y_{ik}^*.

Independence of Irrelevant Alternatives

- Multinomial logit assumes iid errors in the latent utility model, this implies that unobserved factors affecting Y^*_{ij} are unrelated to those affecting Y^*_{ik}.
- MNL makes the Independence of Irrelevant Alternatives (IIA) assumption:

$$\frac{\Pr(\text{Choose } j \mid j \text{ or } k)}{\Pr(\text{Choose } k \mid j \text{ or } k)} = \frac{\Pr(\text{Choose } j \mid j \text{ or } k \text{ or } l)}{\Pr(\text{Choose } k \mid j \text{ or } k \text{ or } l)} \quad \text{for any } l \in \{1, ...J\}$$

Independence of Irrelevant Alternatives

- Multinomial logit assumes iid errors in the latent utility model, this implies that unobserved factors affecting Y^*_{ij} are unrelated to those affecting Y^*_{ik}.
- MNL makes the Independence of Irrelevant Alternatives (IIA) assumption:

$$\frac{\Pr(\text{Choose } j \mid j \text{ or } k)}{\Pr(\text{Choose } k \mid j \text{ or } k)} = \frac{\Pr(\text{Choose } j \mid j \text{ or } k \text{ or } l)}{\Pr(\text{Choose } k \mid j \text{ or } k \text{ or } l)} \quad \text{for any } l \in \{1, \ldots J\}$$

- A classical example of IIA violation: the red bus-blue bus problem

Independence of Irrelevant Alternatives

- Multinomial logit assumes iid errors in the latent utility model, this implies that unobserved factors affecting Y_{ij}^* are unrelated to those affecting Y_{ik}^*.
- MNL makes the Independence of Irrelevant Alternatives (IIA) assumption:

$$\frac{\Pr(\text{Choose } j \mid j \text{ or } k)}{\Pr(\text{Choose } k \mid j \text{ or } k)} = \frac{\Pr(\text{Choose } j \mid j \text{ or } k \text{ or } l)}{\Pr(\text{Choose } k \mid j \text{ or } k \text{ or } l)} \quad \text{for any } l \in \{1, \ldots J\}$$

- A classical example of IIA violation: the red bus-blue bus problem
- That is, the multinomial choice reduces to a series of independent pairwise comparisons

Relaxing IIA with Multinomial Probit

Relaxing IIA with Multinomial Probit

- To relax IIA, we need to allow the stochastic component of the utility ϵ_{ij} to be correlated across choices j for each voter.

Relaxing IIA with Multinomial Probit

- To relax IIA, we need to allow the stochastic component of the utility ϵ_{ij} to be correlated across choices j for each voter.
- Multinomial probit model (MNP):

$$Y_i^* \;=\; X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \;\cdots\; Y_{iJ}^*]^\top \\ X_i = [X_{i1} \;\cdots\; X_{iJ}]^\top \end{cases}$$

Relaxing IIA with Multinomial Probit

- To relax IIA, we need to allow the stochastic component of the utility ϵ_{ij} to be correlated across choices j for each voter.
- Multinomial probit model (MNP):

$$Y_i^* \; = \; X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \; \cdots \; Y_{iJ}^*]^\top \\ X_i = [X_{i1} \; \cdots \; X_{iJ}]^\top \end{cases}$$

- Restrictions on level and scale of Y_i^* for identification

Relaxing IIA with Multinomial Probit

- To relax IIA, we need to allow the stochastic component of the utility ϵ_{ij} to be correlated across choices j for each voter.
- Multinomial probit model (MNP):

$$Y_i^* \; = \; X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \; \cdots \; Y_{iJ}^*]^\top \\ X_i = [X_{i1} \; \cdots \; X_{iJ}]^\top \end{cases}$$

- Restrictions on level and scale of Y_i^* for identification
- Computation is difficult because integral is intractable

Relaxing IIA with Multinomial Probit

- To relax IIA, we need to allow the stochastic component of the utility ϵ_{ij} to be correlated across choices j for each voter.
- Multinomial probit model (MNP):

$$Y_i^* = X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \ \cdots \ Y_{iJ}^*]^\top \\ X_i = [X_{i1} \ \cdots \ X_{iJ}]^\top \end{cases}$$

- Restrictions on level and scale of Y_i^* for identification
- Computation is difficult because integral is intractable
- Moreover, # of parameters in Σ_J increases as J gets large, but data contain little information about Σ_J:

J	3	4	5	6	7
# of elements in Σ_J	6	10	15	21	28
# of parameters identified	2	5	9	14	20

The Poisson Distribution

It's a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time.

Examples:

The Poisson Distribution

It's a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time.

Examples:

1. number of terrorist attacks in a given year

The Poisson Distribution

It's a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time.

Examples:

1. number of terrorist attacks in a given year
2. number of publications by a Professor in a career

The Poisson Distribution

It's a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time.
Examples:

1. number of terrorist attacks in a given year
2. number of publications by a Professor in a career
3. number of days absent from school for High School Sophomores

The Poisson Distribution

It's a discrete probability distribution which gives the probability that some number of events will occur in a fixed period of time.
Examples:

1. number of terrorist attacks in a given year
2. number of publications by a Professor in a career
3. number of days absent from school for High School Sophomores
4. logo for the Stata Press:

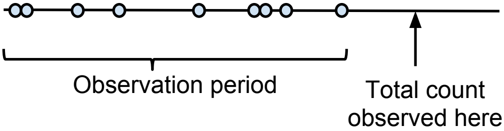
Poisson distribution's first principles:

Poisson distribution's first principles:

1. Begin with an observation period and count point:

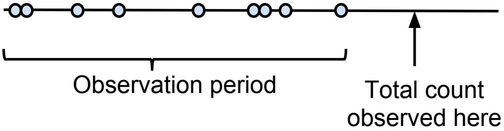
Poisson distribution's first principles:

1. Begin with an observation period and count point:



Poisson distribution's first principles:

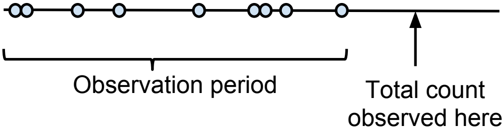
1. Begin with an observation period and count point:



2. Assumptions are about: events occurring between start and count observation. The process of event generation is not observed.

Poisson distribution's first principles:

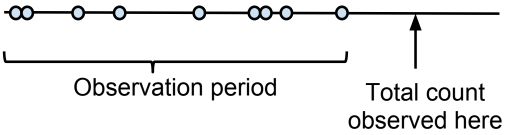
1. Begin with an observation period and count point:



2. Assumptions are about: events occurring between start and count observation. The process of event generation is not observed.
3. 0 events occur at the start of the period

Poisson distribution's first principles:

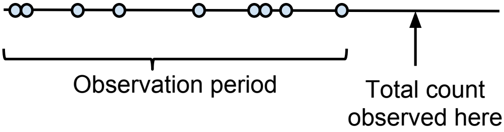
1. Begin with an observation period and count point:



2. Assumptions are about: events occurring between start and count observation. The process of event generation is not observed.
3. 0 events occur at the start of the period
4. Observe only: number of events at end of the period

Poisson distribution's first principles:

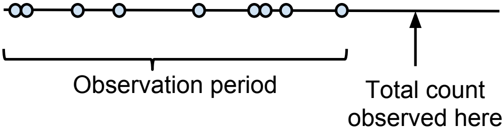
1. Begin with an observation period and count point:



2. Assumptions are about: events occurring between start and count observation. The process of event generation is not observed.
3. 0 events occur at the start of the period
4. Observe only: number of events at end of the period
5. No 2 events can occur at the same time

Poisson distribution's first principles:

1. Begin with an observation period and count point:



2. Assumptions are about: events occurring between start and count observation. The process of event generation is not observed.
3. 0 events occur at the start of the period
4. Observe only: number of events at end of the period
5. No 2 events can occur at the same time
6. $\Pr(\text{event at time } t \mid \text{all events up to time } t-1)$ is constant for all t.

The Poisson Distribution

The Poisson Distribution

Here is the probability density function (PDF) for a random variable Y that is distributed $\text{Pois}(\lambda)$:

$$\Pr(Y = y) = \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda}.$$

The Poisson Distribution

Here is the probability density function (PDF) for a random variable Y that is distributed Pois(λ):

$$\Pr(Y = y) = \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda}.$$

-suppose $Y \sim Pois(3)$. What's $\Pr(Y = 4)$?

$$\Pr(Y = 4) = \frac{3^4}{4!}\mathrm{e}^{-3} = 0.168.$$

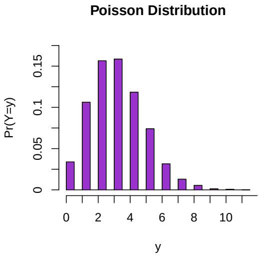
The Poisson Distribution

Here is the probability density function (PDF) for a random variable Y that is distributed Pois(λ):

$$\Pr(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}.$$

-suppose $Y \sim Pois(3)$. What's $\Pr(Y = 4)$?

$$\Pr(Y = 4) = \frac{3^4}{4!} e^{-3} = 0.168.$$



The Poisson Distribution

(λ):

$$\Pr(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}.$$

The Poisson Distribution

(λ):

$$\Pr(Y = y) = \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda}.$$

Using a little bit of geometric series trickery, it isn't too hard to show that $\mathrm{E}[Y] = \sum_{y=0}^{\infty} y \cdot \frac{\lambda^y}{y!}\mathrm{e}^{-\lambda} = \lambda$.

The Poisson Distribution

(λ):

$$\Pr(Y = y) = \frac{\lambda^y}{y!} e^{-\lambda}.$$

Using a little bit of geometric series trickery, it isn't too hard to show that $\mathrm{E}[Y] = \sum_{y=0}^{\infty} y \cdot \frac{\lambda^y}{y!} e^{-\lambda} = \lambda$.

It also turns out that $\mathrm{Var}(Y) = \lambda$, a feature of the model we will discuss later on.

The Poisson Distribution

The Poisson Distribution

Poisson data arises when there is some discrete event which occurs (possibly multiple times) at a constant rate for some fixed time period.

The Poisson Distribution

Poisson data arises when there is some discrete event which occurs (possibly multiple times) at a constant rate for some fixed time period.

This constant rate assumption could be restated: the probability of an event occurring at any moment is independent of whether an event has occurred at any other moment.

The Poisson Distribution

Poisson data arises when there is some discrete event which occurs (possibly multiple times) at a constant rate for some fixed time period.

This constant rate assumption could be restated: the probability of an event occurring at any moment is independent of whether an event has occurred at any other moment.

Derivation of the distribution has some other technical first principles, but the above is the most important.

Connections to Distributions We Have Seen

Connections to Distributions We Have Seen

- Take Binom(n, p) and let $n \rightarrow \infty$ and $p \rightarrow 0$ holding $np = \mu$ constant

Connections to Distributions We Have Seen

- Take Binom(n, p) and let $n \to \infty$ and $p \to 0$ holding $np = \mu$ constant
- If the number of arrivals in the time interval $[0, t]$ follows a Poisson(λt) then the wait times are distributed Exponential with mean $1/\lambda$.

Connections to Distributions We Have Seen

- Take Binom(n, p) and let $n \to \infty$ and $p \to 0$ holding $np = \mu$ constant
- If the number of arrivals in the time interval $[0, t]$ follows a Poisson(λt) then the wait times are distributed Exponential with mean $1/\lambda$.
- For $Y_j|(X = k)$ ~Multinom(k, p_j) then each Y_j ~Pois(λp_j).

Connections to Distributions We Have Seen

- Take Binom(n, p) and let $n \to \infty$ and $p \to 0$ holding $np = \mu$ constant
- If the number of arrivals in the time interval $[0, t]$ follows a Poisson(λt) then the wait times are distributed Exponential with mean $1/\lambda$.
- For $Y_j | (X = k) \sim$Multinom(k, p_j) then each $Y_j \sim$Pois(λp_j).
- If $X_i \sim$Pois(λ_i) for $i = 1 \dots n$ independent then $Y = \sum_{i=1}^{n} X_i \sim \text{Pois}(\sum_{i=1}^{n} \lambda_i)$

The Poisson regression model:

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$

The Poisson regression model:

$$
\begin{aligned}
Y_i &\sim \text{Poisson}(y_i|\lambda_i) \\
\lambda_i &= \exp(x_i\beta)
\end{aligned}
$$

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\mathbb{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\mathbb{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The log-likelihood:

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\mathbb{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The log-likelihood:

$$\ln L(\beta|y) = \sum_{i=1}^{n} \{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\}$$

The Poisson regression model:

$$\begin{aligned} Y_i &\sim \text{Poisson}(y_i|\lambda_i) \\ \lambda_i &= \exp(x_i\beta) \end{aligned}$$

and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\mathbb{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The log-likelihood:

$$\begin{aligned} \ln L(\beta|y) &= \sum_{i=1}^{n} \{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\} \\ &= \sum_{i=1}^{n} \{(x_i\beta)y_i - \exp(x_i\beta) - \ln y_i!\} \end{aligned}$$

The Poisson regression model:

$$Y_i \sim \text{Poisson}(y_i|\lambda_i)$$
$$\lambda_i = \exp(x_i\beta)$$

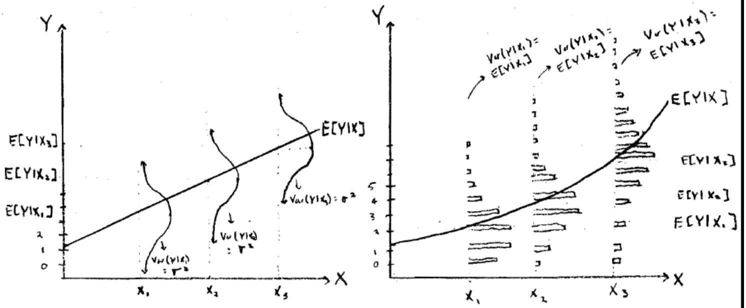
and, as usual, Y_i and Y_j are independent $\forall\ i \neq j$, conditional on X.
The probability density of all the data:

$$\mathbb{P}(y|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}$$

The log-likelihood:

$$\begin{aligned}
\ln L(\beta|y) &= \sum_{i=1}^{n} \{y_i \ln(\lambda_i) - \lambda_i - \ln(y_i!)\} \\
&= \sum_{i=1}^{n} \{(x_i\beta)y_i - \exp(x_i\beta) - \ln y_i!\} \\
&\doteq \sum_{i=1}^{n} \{(x_i\beta)y_i - \exp(x_i\beta)\}
\end{aligned}$$

Comparing with the Linear Model



Example: Civil Conflict in Northern Ireland

Background: a conflict largely along religious lines about the status of Northern Ireland within the United Kingdom, and the division of resources and political power between Northern Ireland's Protestant (mainly Unionist) and Catholic (mainly Republican) communities.

Example: Civil Conflict in Northern Ireland

Background: a conflict largely along religious lines about the status of Northern Ireland within the United Kingdom, and the division of resources and political power between Northern Ireland's Protestant (mainly Unionist) and Catholic (mainly Republican) communities.

The data: the number of Republican deaths for every month from 1969, the beginning of sustained violence, to 2001 (at which point, most organized violence had subsided). Also, the unemployment rates in the two main religious communities.

Example: Civil Conflict in Northern Ireland

Example: Civil Conflict in Northern Ireland

The model: Let $Y_i = \#$ of Republican deaths in a month. Our sole predictor for the moment will be: $U_C =$ the unemployment rate among Northern Ireland's Catholics.

Example: Civil Conflict in Northern Ireland

The model: Let $Y_i = \#$ of Republican deaths in a month. Our sole predictor for the moment will be: $U_C =$ the unemployment rate among Northern Ireland's Catholics.

Our model is then:

$$Y_i \sim Pois(\lambda_i)$$

and

$$\lambda_i = E[Y_i | U_i^C] = exp(\beta_0 + \beta_1 * U_i^C).$$

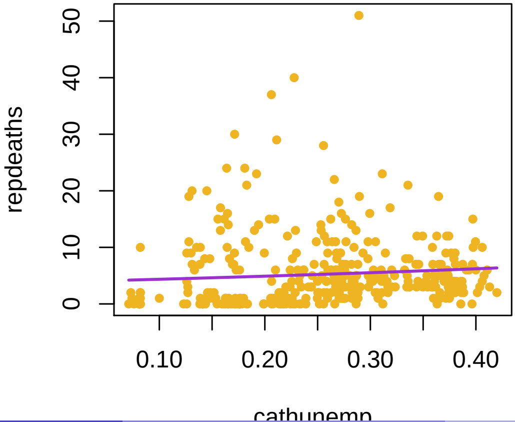
Estimate (just as we have all along!)

```
mod <- glm(repdeaths ~ cathunemp,
         data = troubles, family = poisson(link="log"))

> summary(mod)$coefficients
            Estimate Std. Error  z value     Pr(>|z|)
(Intercept) 1.295875  0.1805327 7.178064 7.070547e-13
cathunemp   1.406498  0.6689819 2.102445 3.551432e-02
```

Our fitted model

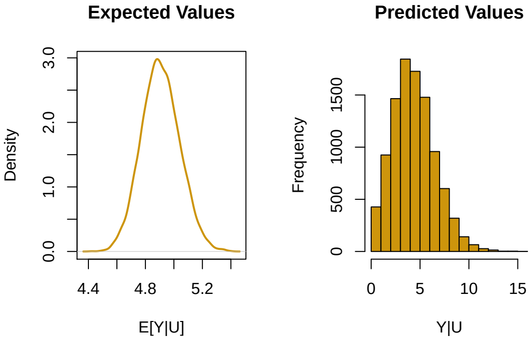
$$\lambda_i = E[Y_i | U_i^C] = exp(1.296 + 1.407 * U_i^C).$$



Some fitted and predicted values

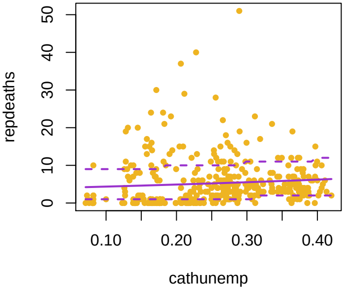
Suppose U_C is equal to .2.

```
mod.coef <- coef(mod); mod.vcov <- vcov(mod)
beta.draws <- mvrnorm(10000, mod.coef, mod.vcov)
lambda.draws <- exp(beta.draws[,1] + .2*beta.draws[,2])
outcome.draws <- rpois(10000, lambda.draws)
```



Overdispersion

36% of observations lie outside the 2.5% or 97.5% quantile of the Poisson distribution that we are alleging generated them.



Overdispersion in Poisson Model

- The Poisson model assumes $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i)$

Overdispersion in Poisson Model

- The Poisson model assumes $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i)$
- But for many count data, $\mathbb{E}(Y_i \mid X_i) < \mathbb{V}(Y_i \mid X_i)$

Overdispersion in Poisson Model

- The Poisson model assumes $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i)$
- But for many count data, $\mathbb{E}(Y_i \mid X_i) < \mathbb{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error

Overdispersion in Poisson Model

- The Poisson model assumes $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i)$
- But for many count data, $\mathbb{E}(Y_i \mid X_i) < \mathbb{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare

Overdispersion in Poisson Model

- The Poisson model assumes $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i)$
- But for many count data, $\mathbb{E}(Y_i \mid X_i) < \mathbb{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare
- One solution to this is to modify the Poisson model by assuming:

$$\mathbb{E}(Y_i \mid X_i) = \mu_i = \exp(X_i^\top \beta) \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = V_i = \phi \mu_i$$

Overdispersion in Poisson Model

- The Poisson model assumes $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i)$
- But for many count data, $\mathbb{E}(Y_i \mid X_i) < \mathbb{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare
- One solution to this is to modify the Poisson model by assuming:

$$\mathbb{E}(Y_i \mid X_i) = \mu_i = \exp(X_i^\top \beta) \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = V_i = \phi\mu_i$$

- This is called the overdispersed Poisson regression model

Overdispersion in Poisson Model

- The Poisson model assumes $\mathbb{E}(Y_i \mid X_i) = \mathbb{V}(Y_i \mid X_i)$
- But for many count data, $\mathbb{E}(Y_i \mid X_i) < \mathbb{V}(Y_i \mid X_i)$
- Potential sources of overdispersion:
  1. unobserved heterogeneity
  2. clustering
  3. contagion or diffusion
  4. (classical) measurement error
- Underdispersion could occur, but rare
- One solution to this is to modify the Poisson model by assuming:

$$\mathbb{E}(Y_i \mid X_i) = \mu_i = \exp(X_i^\top \beta) \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = V_i = \phi \mu_i$$

- This is called the overdispersed Poisson regression model
- When $\phi > 1$, this corresponds to a type of the negative binomial regression model (more on this later)

Derivation as a Gamma-Poisson Mixture

Derivation as a Gamma-Poisson Mixture

Here's the new stochastic component:

Derivation as a Gamma-Poisson Mixture

Here's the new stochastic component:

$$\begin{aligned} Y_i|\varsigma_i &\sim Poisson(\varsigma_i \lambda_i) \\ \varsigma_i &\sim \frac{1}{\theta} Gamma(\theta) \end{aligned}$$

Derivation as a Gamma-Poisson Mixture

Here's the new stochastic component:

$$\begin{aligned} Y_i|\varsigma_i &\sim Poisson(\varsigma_i\lambda_i) \\ \varsigma_i &\sim \frac{1}{\theta}Gamma(\theta) \end{aligned}$$

Note that Gamma(θ) implicitly has location parameter 1, so its mean is θ.

Derivation as a Gamma-Poisson Mixture

Here's the new stochastic component:

$$\begin{aligned} Y_i|\varsigma_i &\sim Poisson(\varsigma_i\lambda_i) \\ \varsigma_i &\sim \frac{1}{\theta}Gamma(\theta) \end{aligned}$$

Note that Gamma(θ) implicitly has location parameter 1, so its mean is θ. This means that $\frac{1}{\theta}$Gamma(θ) has mean 1, and so Poisson($\varsigma_i\lambda_i$) has mean λ_i.

Derivation as a Gamma-Poisson Mixture

Using a similar approach to that described in UPM pgs. 51-52 we can derive the marginal distribution of Y as

$$Y_i \sim Negbin(\lambda_i, \theta)$$

Derivation as a Gamma-Poisson Mixture

Using a similar approach to that described in UPM pgs. 51-52 we can derive the marginal distribution of Y as

$$Y_i \sim Negbin(\lambda_i, \theta)$$

where

$$f_{nb}(y_i|\lambda_i, \theta) = \frac{\Gamma(\theta + y_i)}{y_i!\Gamma(\theta)} \frac{\lambda_i^{y_i} \theta^\theta}{(\lambda_i + \theta)^{\theta + y_i}}$$

Derivation as a Gamma-Poisson Mixture

Using a similar approach to that described in UPM pgs. 51-52 we can derive the marginal distribution of Y as

$$Y_i \sim Negbin(\lambda_i, \theta)$$

where

$$f_{nb}(y_i|\lambda_i, \theta) = \frac{\Gamma(\theta + y_i)}{y_i!\Gamma(\theta)} \frac{\lambda_i^{y_i}\theta^\theta}{(\lambda_i + \theta)^{\theta + y_i}}$$

Notes:

1. $\mathrm{E}[Y_i] = \lambda_i$ and $\mathrm{Var}(Y_i) = \lambda_i + \frac{\lambda_i^2}{\theta}$. What values of θ would be evidence <u>against</u> overdispersion?

Derivation as a Gamma-Poisson Mixture

Using a similar approach to that described in UPM pgs. 51-52 we can derive the marginal distribution of Y as

$$Y_i \sim Negbin(\lambda_i, \theta)$$

where

$$f_{nb}(y_i|\lambda_i, \theta) = \frac{\Gamma(\theta + y_i)}{y_i!\Gamma(\theta)} \frac{\lambda_i^{y_i}\theta^{\theta}}{(\lambda_i + \theta)^{\theta + y_i}}$$

Notes:

1. $\mathrm{E}[Y_i] = \lambda_i$ and $\mathrm{Var}(Y_i) = \lambda_i + \frac{\lambda_i^2}{\theta}$. What values of θ would be evidence <u>against</u> overdispersion?
2. we still have the same old systematic component: $\lambda_i = \exp(X_i\beta)$.

Estimates

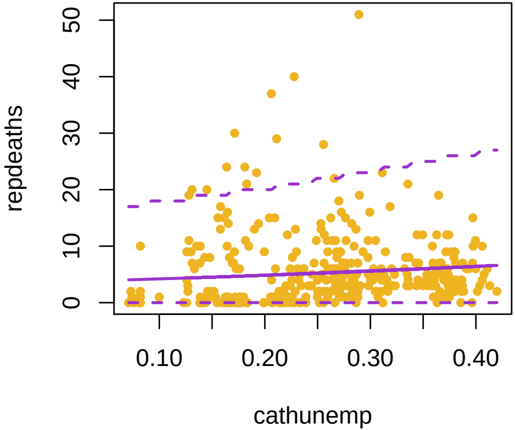
```
mod <- zelig(repdeaths ~ cathunemp, data = troubles,
             model = "negbin")
summary(mod)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.2959     0.1805   7.178 7.07e-13 ***
cathunemp     1.4065     0.6690   2.102   0.0355 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

...
              Theta:  0.8551
          Std. Err.:  0.0754
```

Overdispersion Handled!

5.68% of observations lie at or above the 95% quantile of the Negative Binomial distribution that we are alleging generated them.



Binomial Regression Model

- Sometimes count data have a known upper bound M_i

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i
  - # of children who drop out of high school in a family with M_i children

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - \# of votes for a third party candidate in precinct with population M_i
  - \# of children who drop out of high school in a family with M_i children
- If M_i "trials" are all independent, we have the binomial distribution:

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i
  - # of children who drop out of high school in a family with M_i children
- If M_i "trials" are all independent, we have the binomial distribution:

$$p(Y_i \mid M_i, \pi_i) \;=\; \binom{M_i}{Y_i} \pi_i^{Y_i} (1 - \pi_i)^{M_i - Y_i}$$

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i
  - # of children who drop out of high school in a family with M_i children
- If M_i "trials" are all independent, we have the binomial distribution:

$$p(Y_i \mid M_i, \pi_i) = \binom{M_i}{Y_i} \pi_i^{Y_i} (1 - \pi_i)^{M_i - Y_i}$$

- An exponential family with $\mathbb{E}(Y_i) = M_i \pi_i$ and $\mathbb{V}(Y_i) = M_i \pi_i (1 - \pi_i)$

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i
  - # of children who drop out of high school in a family with M_i children
- If M_i "trials" are all independent, we have the binomial distribution:

$$p(Y_i \mid M_i, \pi_i) \;=\; \binom{M_i}{Y_i} \pi_i^{Y_i}(1-\pi_i)^{M_i - Y_i}$$

- An exponential family with $\mathbb{E}(Y_i) = M_i\pi_i$ and $\mathbb{V}(Y_i) = M_i\pi_i(1-\pi_i)$
- We can thus consider a GLM, the binomial regression model,

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i
  - # of children who drop out of high school in a family with M_i children
- If M_i "trials" are all independent, we have the binomial distribution:

$$p(Y_i \mid M_i, \pi_i) = \binom{M_i}{Y_i} \pi_i^{Y_i} (1 - \pi_i)^{M_i - Y_i}$$

- An exponential family with $\mathbb{E}(Y_i) = M_i \pi_i$ and $\mathbb{V}(Y_i) = M_i \pi_i (1 - \pi_i)$
- We can thus consider a GLM, the binomial regression model, by setting $\pi_i = g^{-1}(X_i^\top \beta)$

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i
  - # of children who drop out of high school in a family with M_i children
- If M_i "trials" are all independent, we have the binomial distribution:

$$p(Y_i \mid M_i, \pi_i) \;=\; \binom{M_i}{Y_i} \pi_i^{Y_i} (1 - \pi_i)^{M_i - Y_i}$$

- An exponential family with $\mathbb{E}(Y_i) = M_i \pi_i$ and $\mathbb{V}(Y_i) = M_i \pi_i (1 - \pi_i)$
- We can thus consider a GLM, the binomial regression model, by setting $\pi_i = g^{-1}(X_i^\top \beta)$
- Common links: logit (canonical), probit, cloglog

Binomial Regression Model

- Sometimes count data have a known upper bound M_i
- Examples:
  - # of votes for a third party candidate in precinct with population M_i
  - # of children who drop out of high school in a family with M_i children
- If M_i "trials" are all independent, we have the binomial distribution:

$$p(Y_i \mid M_i, \pi_i) \; = \; \binom{M_i}{Y_i} \pi_i^{Y_i} (1 - \pi_i)^{M_i - Y_i}$$

- An exponential family with $\mathbb{E}(Y_i) = M_i \pi_i$ and $\mathbb{V}(Y_i) = M_i \pi_i (1 - \pi_i)$
- We can thus consider a GLM, the binomial regression model, by setting $\pi_i = g^{-1}(X_i^\top \beta)$
- Common links: logit (canonical), probit, cloglog
- Note that if $M_i = 1$ for all i, this reduces to a binary outcome model

Estimation and Overdispersion

- The log-likelihood:

$$\ell(\beta \mid X_i) \;=\; \sum_{i=1}^{N} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + M_i \log(1-\pi_i) + \log\binom{M_i}{Y_i} \right\}$$

Estimation and Overdispersion

- The log-likelihood:

$$\ell(\beta \mid X_i) \; = \; \sum_{i=1}^{N} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + M_i \log(1-\pi_i) + \log\binom{M_i}{Y_i} \right\}$$

- Use the standard MLE/GLM machinary to estimate β and calculate quantities of interest

Estimation and Overdispersion

- The log-likelihood:

$$\ell(\beta \mid X_i) \;=\; \sum_{i=1}^{N} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + M_i \log(1-\pi_i) + \log\binom{M_i}{Y_i} \right\}$$

- Use the standard MLE/GLM machinary to estimate β and calculate quantities of interest

- Data are often overdispersed due to dependence between trials

Estimation and Overdispersion

- The log-likelihood:

$$\ell(\beta \mid X_i) \;=\; \sum_{i=1}^{N}\left\{Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + M_i \log(1-\pi_i) + \log\binom{M_i}{Y_i}\right\}$$

- Use the standard MLE/GLM machinary to estimate β and calculate quantities of interest

- Data are often overdispersed due to dependence between trials

- Modify the variance function by including a dispersion parameter:

Estimation and Overdispersion

- The log-likelihood:

$$\ell(\beta \mid X_i) \;=\; \sum_{i=1}^{N} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + M_i \log(1-\pi_i) + \log\binom{M_i}{Y_i} \right\}$$

- Use the standard MLE/GLM machinary to estimate β and calculate quantities of interest
- Data are often overdispersed due to dependence between trials
- Modify the variance function by including a dispersion parameter:

$$\mathbb{E}(Y_i \mid X_i) = M_i\pi_i \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = \phi M_i\pi_i(1-\pi_i)$$

Estimation and Overdispersion

- The log-likelihood:

$$\ell(\beta \mid X_i) \;=\; \sum_{i=1}^{N} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + M_i \log(1-\pi_i) + \log\binom{M_i}{Y_i} \right\}$$

- Use the standard MLE/GLM machinary to estimate β and calculate quantities of interest

- Data are often overdispersed due to dependence between trials
- Modify the variance function by including a dispersion parameter:

$$\mathbb{E}(Y_i \mid X_i) = M_i\pi_i \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = \phi M_i\pi_i(1-\pi_i)$$

- Estimate β and ϕ via QMLE

Estimation and Overdispersion

- The log-likelihood:

$$\ell(\beta \mid X_i) \;=\; \sum_{i=1}^{N} \left\{ Y_i \log\left(\frac{\pi_i}{1-\pi_i}\right) + M_i \log(1-\pi_i) + \log \binom{M_i}{Y_i} \right\}$$

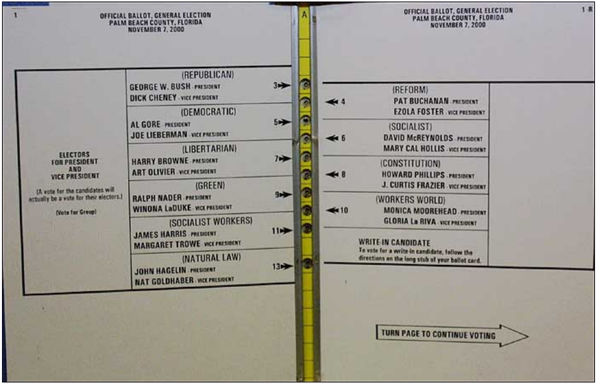
- Use the standard MLE/GLM machinary to estimate β and calculate quantities of interest
- Data are often overdispersed due to dependence between trials
- Modify the variance function by including a dispersion parameter:

$$\mathbb{E}(Y_i \mid X_i) = M_i\pi_i \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = \phi M_i\pi_i(1-\pi_i)$$

- Estimate β and ϕ via QMLE
- This is a GLM, so we have the same robustness properties as the Poisson case

Example: Butterfly Ballot in 2000 Presidential Election

Wand et al. (2001): Did the butterfly ballot give the election to Bush?



- Y_i: Number of votes cast for Buchanan in county i
- X_i: Past Republican & third-party vote shares, demographic covariates
- Wand et al. examine residuals to see how abberant the vote share was in Palm Beach

Fitting GLMs in R

Family	Canonical Link (Default)	Variance	Model
`gaussian`	`identity`	$\phi\ (= \sigma^2)$	normal linear
`binomial`	`logit`	$\mu(1-\mu)$	logit, probit, binomial
`poisson`	`log`	μ	Poisson
`quasibinomial`	`logit`	$\phi\mu(1-\mu)$	overdispersed binomial
`quasipoisson`	`log`	$\phi\mu$	overdispersed Poisson

- Other choices not covered in this course: `Gamma`, `inverse.gaussian`

Fitting GLMs in R

Family	Canonical Link (Default)	Variance	Model
`gaussian`	`identity`	$\phi\ (=\sigma^2)$	normal linear
`binomial`	`logit`	$\mu(1-\mu)$	logit, probit, binomial
`poisson`	`log`	μ	Poisson
`quasibinomial`	`logit`	$\phi\mu(1-\mu)$	overdispersed binomial
`quasipoisson`	`log`	$\phi\mu$	overdispersed Poisson

- Other choices not covered in this course: `Gamma`, `inverse.gaussian`
- You can roll your own GLM using the `quasi` family

Fitting GLMs in R

Family	Canonical Link (Default)	Variance	Model
`gaussian`	`identity`	$\phi\ (= \sigma^2)$	normal linear
`binomial`	`logit`	$\mu(1-\mu)$	logit, probit, binomial
`poisson`	`log`	μ	Poisson
`quasibinomial`	`logit`	$\phi\mu(1-\mu)$	overdispersed binomial
`quasipoisson`	`log`	$\phi\mu$	overdispersed Poisson

- Other choices not covered in this course: `Gamma`, `inverse.gaussian`
- You can roll your own GLM using the `quasi` family
- The negative binomial regression (NB2) can be fitted via the `glm.nb` function in `MASS`

Other Models

Note that there are many other count models for different types of situations:

Other Models

Note that there are many other count models for different types of situations:

- Generalized Event Count (GEC) Model

Other Models

Note that there are many other count models for different types of situations:

- Generalized Event Count (GEC) Model
- Zero-Inflated Poisson

Other Models

Note that there are many other count models for different types of situations:

- Generalized Event Count (GEC) Model
- Zero-Inflated Poisson
- Zero-Inflated Negative Binomial

Other Models

Note that there are many other count models for different types of situations:

- Generalized Event Count (GEC) Model
- Zero-Inflated Poisson
- Zero-Inflated Negative Binomial
- Zero-Truncated Models

Other Models

Note that there are many other count models for different types of situations:

- Generalized Event Count (GEC) Model
- Zero-Inflated Poisson
- Zero-Inflated Negative Binomial
- Zero-Truncated Models
- Hurdle Models

What are duration models used for?

- Survival models = duration models = event history models

What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)

What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)
- Used in biostatistics and engineering: i.e. how long until a patient dies

What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)
- Used in biostatistics and engineering: i.e. how long until a patient dies
- Models the relationship between duration and covariates (how does an increase in X affect the duration Y)

What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)
- Used in biostatistics and engineering: i.e. how long until a patient dies
- Models the relationship between duration and covariates (how does an increase in X affect the duration Y)
- In social science, used in questions such as how long a coalition government lasts,

What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)
- Used in biostatistics and engineering: i.e. how long until a patient dies
- Models the relationship between duration and covariates (how does an increase in X affect the duration Y)
- In social science, used in questions such as how long a coalition government lasts, how long until someone gets a job,

What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)
- Used in biostatistics and engineering: i.e. how long until a patient dies
- Models the relationship between duration and covariates (how does an increase in X affect the duration Y)
- In social science, used in questions such as how long a coalition government lasts, how long until someone gets a job, how a program extends life expectancy

What are duration models used for?

- Survival models = duration models = event history models
- Dependent variable Y is the duration of time that observations spend in some state before experiencing an event (aka failure, death)
- Used in biostatistics and engineering: i.e. how long until a patient dies
- Models the relationship between duration and covariates (how does an increase in X affect the duration Y)
- In social science, used in questions such as how long a coalition government lasts, how long until someone gets a job, how a program extends life expectancy
- Observations should be measured in the same (temporal) units, i.e. don't have some units' duration measured in days and others in months

Why not just use OLS?

Three reasons:

Why not just use OLS?

Three reasons:

1. The normal linear model assumes Y is Normal but duration dependent variables are always positive (number of years, etc.)

Why not just use OLS?

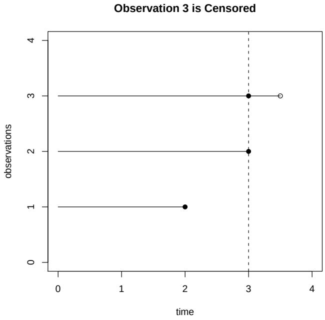
Three reasons:

1. The normal linear model assumes Y is Normal but duration dependent variables are always positive (number of years, etc.)
2. Duration models can handle censoring

Why not just use OLS?

Three reasons:

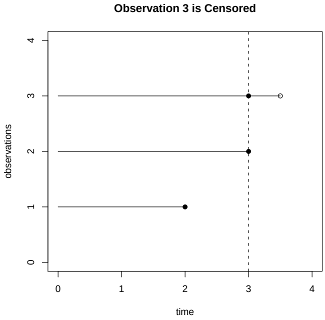
1. The normal linear model assumes Y is Normal but duration dependent variables are always positive (number of years, etc.)
2. Duration models can handle censoring



Why not just use OLS?

Three reasons:

1. The normal linear model assumes Y is Normal but duration dependent variables are always positive (number of years, etc.)
2. Duration models can handle censoring



Observation 3 is censored in that it has not experienced the event at the time we stop collecting data, so we don't know its true duration

Why not use OLS?

3. Duration models can handle time-varying covariates

Why not use OLS?

3. Duration models can handle time-varying covariates
   - If Y is duration of a regime, GDP may change during the duration of the regime

Why not use OLS?

3. Duration models can handle time-varying covariates
   - If Y is duration of a regime, GDP may change during the duration of the regime
   - OLS cannot handle multiple values of GDP per observation

Why not use OLS?

3. Duration models can handle time-varying covariates
   - If Y is duration of a regime, GDP may change during the duration of the regime
   - OLS cannot handle multiple values of GDP per observation
   - You can set up data in a special way with duration models such that you can accommodate time-varying covariates

Duration/Survival Model Jargon

Duration/Survival Model Jargon

Let T denote a continuous positive random variable representing the duration/survival times ($T = Y$)

Duration/Survival Model Jargon

Let T denote a continuous positive random variable representing the duration/survival times ($T = Y$)

T has a probability density function $f(t)$

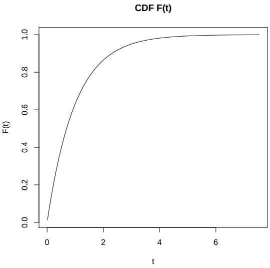
Duration/Survival Model Jargon

Let T denote a continuous positive random variable representing the duration/survival times ($T = Y$)

T has a probability density function $f(t)$

Duration/Survival Model Jargon

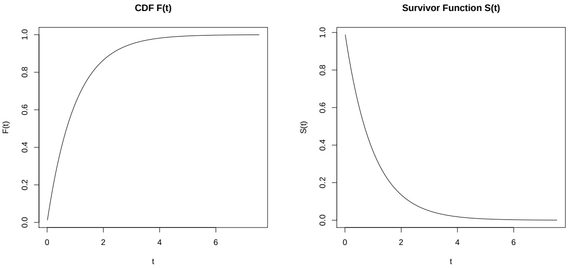
F(t): the CDF of $f(t)$, $\int_0^t f(u)du = P(T \leq t)$, which is the probability of an event occurring before (or at exactly) time t



Duration/Survival Model Jargon

F(t): the CDF of $f(t)$, $\int_0^t f(u)du = P(T \leq t)$, which is the probability of an event occurring before (or at exactly) time t

Survivor function: The probability of surviving (i.e. no event occuring) until at least time t: $S(t) = 1 - F(t) = P(T > t)$

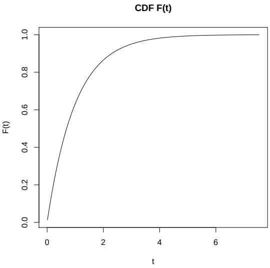


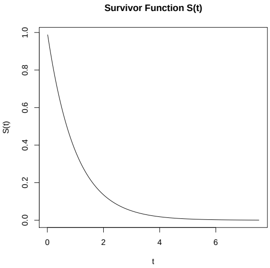
Duration/Survival Model Jargon

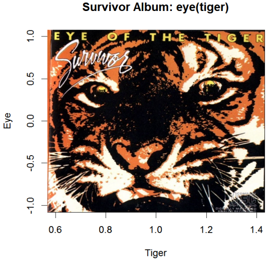
F(t): the CDF of $f(t)$, $\int_0^t f(u)du = P(T \leq t)$, which is the probability of an event occurring before (or at exactly) time t

Survivor function: The probability of surviving (i.e. no event occuring) until at least time t: $S(t) = 1 - F(t) = P(T > t)$

Eye of the Tiger: 1982 album by the band Survivor, which reached number 2 on the US Billboard 200 chart.







Duration/Survival Model Jargon

Duration/Survival Model Jargon

Hazard rate (or hazard function): $h(t)$ is roughly the probability of an event at time t given survival up to time t

Duration/Survival Model Jargon

Hazard rate (or hazard function): $h(t)$ is roughly the probability of an event at time t given survival up to time t

$$h(t) = P(t \leq T < t + \tau | T \geq t)$$

Duration/Survival Model Jargon

Hazard rate (or hazard function): $h(t)$ is roughly the probability of an event at time t given survival up to time t

$$
\begin{aligned}
h(t) &= P(t \leq T < t + \tau | T \geq t) \\
&= P(\text{event at } t | \text{survival up to } t)
\end{aligned}
$$

Duration/Survival Model Jargon

Hazard rate (or hazard function): $h(t)$ is roughly the probability of an event at time t given survival up to time t

$$\begin{aligned} h(t) &= P(t \leq T < t + \tau | T \geq t) \\ &= P(\text{event at } t | \text{survival up to } t) \\ &= \frac{P(\text{survival up to } t | \text{event at } t) P(\text{event at } t)}{P(\text{survival up to } t)} \end{aligned}$$

Duration/Survival Model Jargon

Hazard rate (or hazard function): $h(t)$ is roughly the probability of an event at time t given survival up to time t

$$
\begin{aligned}
h(t) &= P(t \leq T < t + \tau | T \geq t) \\
&= P(\text{event at } t | \text{survival up to } t) \\
&= \frac{P(\text{survival up to } t | \text{event at } t) P(\text{event at } t)}{P(\text{survival up to } t)} \\
&= \frac{P(\text{event at } t)}{P(\text{survival up to } t)}
\end{aligned}
$$

Duration/Survival Model Jargon

Hazard rate (or hazard function): $h(t)$ is roughly the probability of an event at time t given survival up to time t

$$\begin{aligned} h(t) &= P(t \leq T < t + \tau | T \geq t) \\ &= P(\text{event at } t | \text{survival up to } t) \\ &= \frac{P(\text{survival up to } t | \text{event at } t) P(\text{event at } t)}{P(\text{survival up to } t)} \\ &= \frac{P(\text{event at } t)}{P(\text{survival up to } t)} \\ &= \frac{f(t)}{S(t)} \end{aligned}$$

Relating the Density, Survival, and Hazard Functions

$$h(t) = \frac{f(t)}{S(t)}$$

implies

$$\underbrace{f(t)}_{\text{density function}} = \underbrace{h(t)}_{\text{hazard function}} \cdot \underbrace{S(t)}_{\text{survival function}}$$

Modeling with Covariates

Modeling with Covariates

We can model the mean of the duration times as a function of covariates via a link function $g(\cdot)$

Modeling with Covariates

We can model the mean of the duration times as a function of covariates via a link function $g(\cdot)$

$$g(E[T_i]) = X_i\beta$$

Modeling with Covariates

We can model the mean of the duration times as a function of covariates via a link function $g(\cdot)$

$$g(E[T_i]) = X_i\beta$$

and estimate β via maximum likelihood.

How to estimate parametric survival models

How to estimate parametric survival models

They might seem fancy and complicated, but we estimate these models the same as every other model!

How to estimate parametric survival models

They might seem fancy and complicated, but we estimate these models the same as every other model!

1. Make an assumption that T_i follows a specific distribution $f(t)$ (i.e. choose the stochastic component).

How to estimate parametric survival models

They might seem fancy and complicated, but we estimate these models the same as every other model!

1. Make an assumption that T_i follows a specific distribution $f(t)$ (i.e. choose the stochastic component).
2. Model the hazard rate with covariates (i.e. specify the systematic component).

How to estimate parametric survival models

They might seem fancy and complicated, but we estimate these models the same as every other model!

1. Make an assumption that T_i follows a specific distribution $f(t)$ (i.e. choose the stochastic component).
2. Model the hazard rate with covariates (i.e. specify the systematic component).
3. Estimate via maximum likelihood.

How to estimate parametric survival models

They might seem fancy and complicated, but we estimate these models the same as every other model!

1. Make an assumption that T_i follows a specific distribution $f(t)$ (i.e. choose the stochastic component).
2. Model the hazard rate with covariates (i.e. specify the systematic component).
3. Estimate via maximum likelihood.
4. Interpret quantities of interest (hazard ratios, expected survival times).

What's Special About Survival Models?

What's Special About Survival Models?

Censoring:

... it makes modeling a little tricky.

What's Special About Survival Models?

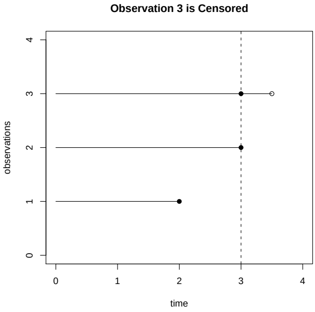
Censoring:

... it makes modeling a little tricky.

But not too tricky

What's Special About Survival Models?

Censoring:

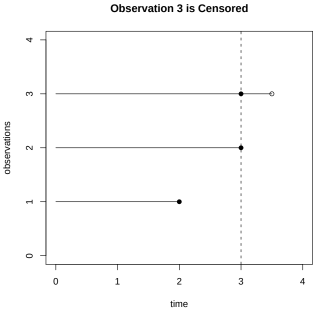


... it makes modeling a little tricky.

But not too tricky

What's Special About Survival Models?

Censoring:



... it makes modeling a little tricky. But not too tricky

Observation 3 is censored because it had not experienced the event when we collected the data, so we don't know its true duration.

Censoring

Censoring

Observations that are censored give us information about how long they survive.

Censoring

Observations that are censored give us information about how long they survive.

For censored observations, we know that they survived at least until some observed time, t^c, and that the true duration, t is greater than or equal to t^c.

Censoring

Observations that are censored give us information about how long they survive.

For censored observations, we know that they survived at least until some observed time, t^c, and that the true duration, t is greater than or equal to t^c.

For each observation, let's create a censoring indicator variable, c_i, such that

$$c_i = \begin{cases} 1 & \text{if not censored} \\ 0 & \text{if censored} \end{cases}$$

Censoring

We can incorporate the information from the censored observations into the likelihood function.

Censoring

We can incorporate the information from the censored observations into the likelihood function.

$$\mathcal{L} = \prod_{i=1}^{n} \underbrace{[f(t_i)]^{c_i}}_{\text{uncensored}} \underbrace{[P(T_i \geq t_i^c)]^{1-c_i}}_{\text{censored}}$$

Censoring

We can incorporate the information from the censored observations into the likelihood function.

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \underbrace{[f(t_i)]^{c_i}}_{\text{uncensored}} \underbrace{[P(T_i \geq t_i^c)]^{1-c_i}}_{\text{censored}} \\
&= \prod_{i=1}^{n} [f(t_i)]^{c_i} \, [1 - F(t_i)]^{1-c_i}
\end{aligned}$$

Censoring

We can incorporate the information from the censored observations into the likelihood function.

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \underbrace{[f(t_i)]^{c_i}}_{\text{uncensored}} \underbrace{[P(T_i \geq t_i^c)]^{1-c_i}}_{\text{censored}} \\
&= \prod_{i=1}^{n} [f(t_i)]^{c_i} \; [1 - F(t_i)]^{1-c_i} \\
&= \prod_{i=1}^{n} [f(t_i)]^{c_i} \; [S(t_i)]^{1-c_i}
\end{aligned}$$

Censoring

We can incorporate the information from the censored observations into the likelihood function.

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \underbrace{[f(t_i)]^{c_i}}_{\text{uncensored}} \underbrace{[P(T_i \geq t_i^c)]^{1-c_i}}_{\text{censored}} \\
&= \prod_{i=1}^{n} [f(t_i)]^{c_i} \; [1 - F(t_i)]^{1-c_i} \\
&= \prod_{i=1}^{n} [f(t_i)]^{c_i} \; [S(t_i)]^{1-c_i}
\end{aligned}
$$

So uncensored observations contribute to the density function and censored observations contribute to the survivor function in the likelihood.

The Poisson Process

The Poisson Process

- Popular example of stochastic process

The Poisson Process

- Popular example of stochastic process
- Principles of Poisson process:

The Poisson Process

- Popular example of stochastic process
- Principles of Poisson process:
  - Independent increments: number of events occurring in two disjoint intervals is independent

The Poisson Process

- Popular example of stochastic process
- Principles of Poisson process:
    - Independent increments: number of events occurring in two disjoint intervals is independent
    - Stationary increments: probability distribution of number of occurrences depends only on the time length of interval (because of common rate)

The Poisson Process

- Popular example of stochastic process
- Principles of Poisson process:
  - Independent increments: number of events occurring in two disjoint intervals is independent
  - Stationary increments: probability distribution of number of occurrences depends only on the time length of interval (because of common rate)
- Events occur at rate λ (expected occurrences per unit of time)

The Poisson Process

- Popular example of stochastic process
- Principles of Poisson process:
  - Independent increments: number of events occurring in two disjoint intervals is independent
  - Stationary increments: probability distribution of number of occurrences depends only on the time length of interval (because of common rate)
- Events occur at rate λ (expected occurrences per unit of time)
- $N_\tau =$ number of arrivals in time period of length τ
  - $N_\tau \sim \text{Poisson}(\lambda\tau)$

The Poisson Process

The Poisson Process

- Exponential distribution measures the times between events in a Poisson process

The Poisson Process

- Exponential distribution measures the times between events in a Poisson process
- $T =$ time to wait until next event in a Poisson process with rate λ

The Poisson Process

- Exponential distribution measures the times between events in a Poisson process
- T = time to wait until next event in a Poisson process with rate λ
- $T \sim \mathrm{Expo}(\lambda)$

The Poisson Process

- Exponential distribution measures the times between events in a Poisson process
- T = time to wait until next event in a Poisson process with rate λ
- $T \sim \text{Expo}(\lambda)$
- Memorylessness property: how much you have waited already is irrelevant

$$P(T > t + k | T > t) = P(T > k)$$

$$P(T > 3 + 5 | T > 3) = P(T > 5)$$

Two Possible Parameterizations of the Exponential Model

Two Possible Parameterizations of the Exponential Model

- $\lambda_i > 0$ is the **rate** parameter

$$T_i \sim \text{Exponential}(\lambda_i)$$

Two Possible Parameterizations of the Exponential Model

- $\lambda_i > 0$ is the **rate** parameter

$$T_i \sim \text{Exponential}(\lambda_i)$$

$$f(t_i) = \lambda_i e^{-\lambda_i t_i}$$

Two Possible Parameterizations of the Exponential Model

- $\lambda_i > 0$ is the **rate** parameter

$$T_i \sim \text{Exponential}(\lambda_i)$$

$$f(t_i) = \lambda_i e^{-\lambda_i t_i}$$

$$E(T_i) = \frac{1}{\lambda_i}$$

Two Possible Parameterizations of the Exponential Model

- $\lambda_i > 0$ is the **rate** parameter

$$T_i \sim \text{Exponential}(\lambda_i)$$

$$f(t_i) = \lambda_i e^{-\lambda_i t_i}$$

$$E(T_i) = \frac{1}{\lambda_i}$$

- $\theta_i > 0$ is **scale** parameter $(\theta_i = \frac{1}{\lambda_i})$

$$T_i \sim \text{Exponential}(\theta_i)$$

Two Possible Parameterizations of the Exponential Model

- $\lambda_i > 0$ is the **rate** parameter

$$T_i \sim \text{Exponential}(\lambda_i)$$

$$f(t_i) = \lambda_i e^{-\lambda_i t_i}$$

$$E(T_i) = \frac{1}{\lambda_i}$$

- $\theta_i > 0$ is **scale** parameter ($\theta_i = \frac{1}{\lambda_i}$)

$$T_i \sim \text{Exponential}(\theta_i)$$

$$f(t_i) = \frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}$$

Two Possible Parameterizations of the Exponential Model

- $\lambda_i > 0$ is the **rate** parameter

$$T_i \sim \text{Exponential}(\lambda_i)$$

$$f(t_i) = \lambda_i e^{-\lambda_i t_i}$$

$$E(T_i) = \frac{1}{\lambda_i}$$

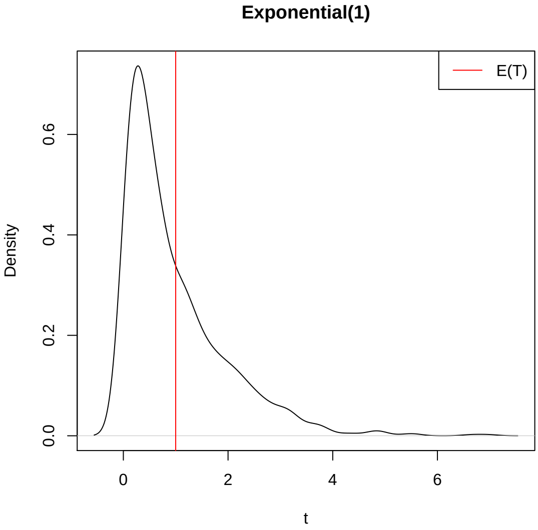
- $\theta_i > 0$ is **scale** parameter ($\theta_i = \frac{1}{\lambda_i}$)

$$T_i \sim \text{Exponential}(\theta_i)$$

$$f(t_i) = \frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}$$

$$E(T_i) = \theta_i$$

The Exponential Model



Link Functions

Link Functions

- If you use a rate parameterization with λ_i:

Link Functions

- If you use a rate parameterization with λ_i:

$$E(T_i) = \frac{1}{\lambda_i} = \frac{1}{\exp(x_i\beta)}$$

Link Functions

- If you use a rate parameterization with λ_i:

$$E(T_i) = \frac{1}{\lambda_i} = \frac{1}{\exp(x_i\beta)}$$

Positive β implies that expected duration time decreases as x increases.

Link Functions

- If you use a rate parameterization with λ_i:

$$E(T_i) = \frac{1}{\lambda_i} = \frac{1}{\exp(x_i \beta)}$$

  Positive β implies that expected duration time decreases as x increases.

- If you use a scale parameterization with θ_i

Link Functions

- If you use a rate parameterization with λ_i:

$$E(T_i) = \frac{1}{\lambda_i} = \frac{1}{\exp(x_i\beta)}$$

  Positive β implies that expected duration time decreases as x increases.

- If you use a scale parameterization with θ_i

$$E(T_i) = \theta_i = \exp(x_i\beta)$$

Link Functions

- If you use a rate parameterization with λ_i:

$$E(T_i) = \frac{1}{\lambda_i} = \frac{1}{\exp(x_i\beta)}$$

Positive β implies that expected duration time decreases as x increases.

- If you use a scale parameterization with θ_i

$$E(T_i) = \theta_i = \exp(x_i\beta)$$

Positive β implies that expected duration time increases as x increases.

Hazard Function for Rate Parametrization

For $T_i \sim \text{Exponential}(\lambda_i)$:

Hazard Function for Rate Parametrization

For $T_i \sim \text{Exponential}(\lambda_i)$:

$$f(t) = \lambda_i e^{-\lambda_i t}$$

Hazard Function for Rate Parametrization

For $T_i \sim \mathrm{Exponential}(\lambda_i)$:

$$
\begin{aligned}
f(t) &= \lambda_i e^{-\lambda_i t} \\
S(t) &= 1 - F(t)
\end{aligned}
$$

Hazard Function for Rate Parametrization

For $T_i \sim \text{Exponential}(\lambda_i)$:

$$
\begin{aligned}
f(t) &= \lambda_i e^{-\lambda_i t} \\
\\
S(t) &= 1 - F(t) \\
&= 1 - (1 - e^{-\lambda t})
\end{aligned}
$$

Hazard Function for Rate Parametrization

For $T_i \sim \text{Exponential}(\lambda_i)$:

$$f(t) = \lambda_i e^{-\lambda_i t}$$

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda_i t} \end{aligned}$$

Hazard Function for Rate Parametrization

For $T_i \sim \mathrm{Exponential}(\lambda_i)$:

$$f(t) = \lambda_i e^{-\lambda_i t}$$

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda_i t} \end{aligned}$$

$$h(t) = \frac{f(t)}{S(t)}$$

Hazard Function for Rate Parametrization

For $T_i \sim \mathrm{Exponential}(\lambda_i)$:

$$f(t) = \lambda_i e^{-\lambda_i t}$$

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda_i t} \end{aligned}$$

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\lambda_i e^{-\lambda_i t}}{e^{-\lambda_i t}} \end{aligned}$$

Hazard Function for Rate Parametrization

For $T_i \sim \text{Exponential}(\lambda_i)$:

$$f(t) = \lambda_i e^{-\lambda_i t}$$

$$\begin{aligned} S(t) &= 1 - F(t) \\ &= 1 - (1 - e^{-\lambda t}) \\ &= e^{-\lambda_i t} \end{aligned}$$

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\lambda_i e^{-\lambda_i t}}{e^{-\lambda_i t}} \\ &= \lambda_i \end{aligned}$$

Hazard Function for Scale Parametrization

For $T_i \sim \text{Exponential}(\theta_i)$:

Hazard Function for Scale Parametrization

For $T_i \sim \mathrm{Exponential}(\theta_i)$:

$$f(t) = \frac{1}{\theta_i}\exp[-\frac{t}{\theta_i}]$$

Hazard Function for Scale Parametrization

For $T_i \sim \text{Exponential}(\theta_i)$:

$$
\begin{aligned}
f(t) &= \frac{1}{\theta_i}\exp[-\frac{t}{\theta_i}] \\
S(t) &= 1 - F(t)
\end{aligned}
$$

Hazard Function for Scale Parametrization

For $T_i \sim \mathrm{Exponential}(\theta_i)$:

$$\begin{aligned} f(t) &= \frac{1}{\theta_i} \exp[-\frac{t}{\theta_i}] \\ S(t) &= 1 - F(t) \\ &= 1 - (1 - \exp[-\frac{t}{\theta_i}]) \end{aligned}$$

Hazard Function for Scale Parametrization

For $T_i \sim \text{Exponential}(\theta_i)$:

$$
\begin{aligned}
f(t) &= \frac{1}{\theta_i} \exp[-\frac{t}{\theta_i}] \\
S(t) &= 1 - F(t) \\
&= 1 - (1 - \exp[-\frac{t}{\theta_i}]) \\
&= \exp[-\frac{t}{\theta_i}]
\end{aligned}
$$

Hazard Function for Scale Parametrization

For $T_i \sim \mathrm{Exponential}(\theta_i)$:

$$
\begin{aligned}
f(t) &= \frac{1}{\theta_i}\exp[-\frac{t}{\theta_i}] \\
S(t) &= 1 - F(t) \\
&= 1 - (1 - \exp[-\frac{t}{\theta_i}]) \\
&= \exp[-\frac{t}{\theta_i}] \\
h(t) &= \frac{f(t)}{S(t)}
\end{aligned}
$$

Hazard Function for Scale Parametrization

For $T_i \sim \text{Exponential}(\theta_i)$:

$$
\begin{aligned}
f(t) &= \frac{1}{\theta_i}\exp[-\frac{t}{\theta_i}] \\
\\
S(t) &= 1 - F(t) \\
&= 1 - (1 - \exp[-\frac{t}{\theta_i}]) \\
&= \exp[-\frac{t}{\theta_i}] \\
\\
h(t) &= \frac{f(t)}{S(t)} \\
&= \frac{\frac{1}{\theta_i}\exp[-\frac{t}{\theta_i}]}{\exp[-\frac{t}{\theta_i}]}
\end{aligned}
$$

Hazard Function for Scale Parametrization

For $T_i \sim \text{Exponential}(\theta_i)$:

$$\begin{aligned} f(t) &= \frac{1}{\theta_i} \exp[-\frac{t}{\theta_i}] \\ S(t) &= 1 - F(t) \\ &= 1 - (1 - \exp[-\frac{t}{\theta_i}]) \\ &= \exp[-\frac{t}{\theta_i}] \\ h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{\frac{1}{\theta_i}\exp[-\frac{t}{\theta_i}]}{\exp[-\frac{t}{\theta_i}]} = \frac{1}{\theta_i} \end{aligned}$$

Let's work with the scale parametrization

- Note that $h(t) = \frac{1}{\theta_i}$, which does not depend on t!

Let's work with the scale parametrization

- Note that $h(t) = \frac{1}{\theta_i}$, which does not depend on t!
  - The exponential model thus assume a **flat hazard**: Every unit / individual has their own hazard rate, but it does not change over time

Let's work with the scale parametrization

- Note that $h(t) = \frac{1}{\theta_i}$, which does not depend on t!
    - The exponential model thus assume a **flat hazard**: Every unit / individual has their own hazard rate, but it does not change over time
    - Connected to **memorylessness property** of the exponential distribution

Let's work with the scale parametrization

- Note that $h(t) = \frac{1}{\theta_i}$, which does not depend on t!
  - The exponential model thus assume a **flat hazard**: Every unit / individual has their own hazard rate, but it does not change over time
  - Connected to **memorylessness property** of the exponential distribution

Modeling $h(t)$ with covariates:

Let's work with the scale parametrization

- Note that $h(t) = \frac{1}{\theta_i}$, which does not depend on t!
  - The exponential model thus assume a **flat hazard**: Every unit / individual has their own hazard rate, but it does not change over time
  - Connected to **memorylessness property** of the exponential distribution

Modeling $h(t)$ with covariates:

$$h(t) = \frac{1}{\theta_i} = \exp[-x_i\beta]$$

Let's work with the scale parametrization

- Note that $h(t) = \frac{1}{\theta_i}$, which does not depend on t!
  - The exponential model thus assume a **flat hazard**: Every unit / individual has their own hazard rate, but it does not change over time
  - Connected to **memorylessness property** of the exponential distribution

Modeling $h(t)$ with covariates:

$$h(t) = \frac{1}{\theta_i} = \exp[-x_i\beta]$$

Positive β implies that hazard decreases and average survival time increases as x increases.

Estimation via ML:

Estimation via ML:

$$\mathcal{L} = \prod_{i=1}^{n} [f(t_i)]^{1-c_i} [1 - F(t_i)]^{c_i}$$

Estimation via ML:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \left[f(t_i)\right]^{1-c_i} \left[1 - F(t_i)\right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}\right]^{1-c_i} \left[e^{-\frac{t_i}{\theta_i}}\right]^{c_i}
\end{aligned}
$$

Estimation via ML:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [f(t_i)]^{1-c_i} [1 - F(t_i)]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}\right]^{1-c_i} \left[e^{-\frac{t_i}{\theta_i}}\right]^{c_i} \\
\ell &= \sum_{i=1}^{n} (1 - c_i)(\ln \frac{1}{\theta_i} - \frac{t_i}{\theta_i}) + c_i(-\frac{t_i}{\theta_i})
\end{aligned}
$$

Estimation via ML:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [f(t_i)]^{1-c_i} [1 - F(t_i)]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}\right]^{1-c_i} \left[e^{-\frac{t_i}{\theta_i}}\right]^{c_i} \\
\ell &= \sum_{i=1}^{n} (1 - c_i)(\ln \frac{1}{\theta_i} - \frac{t_i}{\theta_i}) + c_i(-\frac{t_i}{\theta_i}) \\
&= \sum_{i=1}^{n} (1 - c_i)(\ln e^{-\mathbf{x}_i \beta} - e^{-\mathbf{x}_i \beta} t_i) + c_i(-e^{-\mathbf{x}_i \beta} t_i)
\end{aligned}
$$

Estimation via ML:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \left[f(t_i)\right]^{1-c_i} \left[1 - F(t_i)\right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}\right]^{1-c_i} \left[e^{-\frac{t_i}{\theta_i}}\right]^{c_i} \\
\ell &= \sum_{i=1}^{n} (1 - c_i)(\ln \frac{1}{\theta_i} - \frac{t_i}{\theta_i}) + c_i(-\frac{t_i}{\theta_i}) \\
&= \sum_{i=1}^{n} (1 - c_i)(\ln e^{-\mathbf{x}_i \beta} - e^{-\mathbf{x}_i \beta} t_i) + c_i(-e^{-\mathbf{x}_i \beta} t_i) \\
&= \sum_{i=1}^{n} (1 - c_i)(-\mathbf{x}_i \beta - e^{-\mathbf{x}_i \beta} t_i) - c_i(e^{-\mathbf{x}_i \beta} t_i)
\end{aligned}
$$

Estimation via ML:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [f(t_i)]^{1-c_i} [1 - F(t_i)]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{1}{\theta_i} e^{-\frac{t_i}{\theta_i}}\right]^{1-c_i} \left[e^{-\frac{t_i}{\theta_i}}\right]^{c_i} \\
\ell &= \sum_{i=1}^{n} (1 - c_i)(\ln \frac{1}{\theta_i} - \frac{t_i}{\theta_i}) + c_i(-\frac{t_i}{\theta_i}) \\
&= \sum_{i=1}^{n} (1 - c_i)(\ln e^{-\mathbf{x}_i\beta} - e^{-\mathbf{x}_i\beta} t_i) + c_i(-e^{-\mathbf{x}_i\beta} t_i) \\
&= \sum_{i=1}^{n} (1 - c_i)(-\mathbf{x}_i\beta - e^{-\mathbf{x}_i\beta} t_i) - c_i(e^{-\mathbf{x}_i\beta} t_i) \\
&= \sum_{i=1}^{n} (1 - c_i)(-\mathbf{x}_i\beta) - e^{-\mathbf{x}_i\beta} t_i
\end{aligned}
$$

Quantities of interest

Quantities of interest

If our outcome variable is how long a parliamentary government lasts, and we're interested in the effect of majority versus minority governments. We could calculate:

Quantities of interest

If our outcome variable is how long a parliamentary government lasts, and we're interested in the effect of majority versus minority governments. We could calculate:

- Find the hazard ratio of majority to minority governments

Quantities of interest

If our outcome variable is how long a parliamentary government lasts, and we're interested in the effect of majority versus minority governments. We could calculate:

- Find the hazard ratio of majority to minority governments
- Expected survival time for majority and minority governments

Quantities of interest

If our outcome variable is how long a parliamentary government lasts, and we're interested in the effect of majority versus minority governments. We could calculate:

- Find the hazard ratio of majority to minority governments
- Expected survival time for majority and minority governments
- Predicted survival times for majority and minority governments

Quantities of interest

If our outcome variable is how long a parliamentary government lasts, and we're interested in the effect of majority versus minority governments. We could calculate:

- Find the hazard ratio of majority to minority governments
- Expected survival time for majority and minority governments
- Predicted survival times for majority and minority governments
- First differences in expected survival times between majority and minority governments

Hazard Ratios

Hazard Ratios

$$\text{HR} = \frac{h(t|\mathbf{x}_{\text{maj}})}{h(t|\mathbf{x}_{\text{min}})}$$

Hazard Ratios

$$
\begin{aligned}
\mathrm{HR} &= \frac{h(t|\mathbf{x}_{\mathrm{maj}})}{h(t|\mathbf{x}_{\mathrm{min}})} \\
&= \frac{e^{-\mathbf{x}_{\mathrm{maj}}\beta}}{e^{-\mathbf{x}_{\mathrm{min}}\beta}}
\end{aligned}
$$

Hazard Ratios

$$
\begin{aligned}
\mathrm{HR} &= \frac{h(t|\mathbf{x}_{\mathrm{maj}})}{h(t|\mathbf{x}_{\mathrm{min}})} \\
&= \frac{e^{-\mathbf{x}_{\mathrm{maj}}\beta}}{e^{-\mathbf{x}_{\mathrm{min}}\beta}} \\
&= \frac{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{maj}}\beta_4} e^{-x_5\beta_5}}{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{min}}\beta_4} e^{-x_5\beta_5}}
\end{aligned}
$$

Hazard Ratios

$$\begin{aligned}
\mathrm{HR} &= \frac{h(t|\mathbf{x}_{\mathrm{maj}})}{h(t|\mathbf{x}_{\mathrm{min}})} \\
&= \frac{e^{-\mathbf{x}_{\mathrm{maj}}\beta}}{e^{-\mathbf{x}_{\mathrm{min}}\beta}} \\
&= \frac{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{maj}}\beta_4} e^{-x_5\beta_5}}{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{min}}\beta_4} e^{-x_5\beta_5}} \\
&= \frac{e^{-x_{\mathrm{maj}}\beta_4}}{e^{-x_{\mathrm{min}}\beta_4}}
\end{aligned}$$

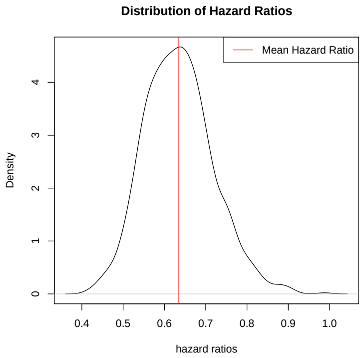
Hazard Ratios

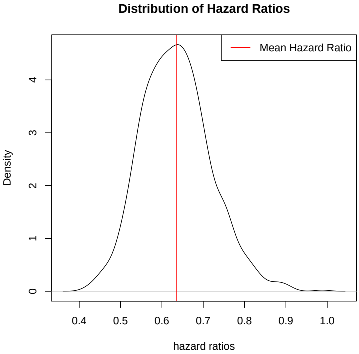
$$
\begin{aligned}
\mathrm{HR} &= \frac{h(t|\mathbf{x}_{\mathrm{maj}})}{h(t|\mathbf{x}_{\mathrm{min}})} \\
&= \frac{e^{-\mathbf{x}_{\mathrm{maj}}\beta}}{e^{-\mathbf{x}_{\mathrm{min}}\beta}} \\
&= \frac{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{maj}}\beta_4} e^{-x_5\beta_5}}{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{min}}\beta_4} e^{-x_5\beta_5}} \\
&= \frac{e^{-x_{\mathrm{maj}}\beta_4}}{e^{-x_{\mathrm{min}}\beta_4}} \\
&= e^{-\beta_4}
\end{aligned}
$$

Hazard Ratios

$$
\begin{aligned}
\mathrm{HR} &= \frac{h(t|\mathbf{x}_{\mathrm{maj}})}{h(t|\mathbf{x}_{\mathrm{min}})} \\
&= \frac{e^{-\mathbf{x}_{\mathrm{maj}}\beta}}{e^{-\mathbf{x}_{\mathrm{min}}\beta}} \\
&= \frac{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{maj}}\beta_4} e^{-x_5\beta_5}}{e^{-\beta_0} e^{-x_1\beta_1} e^{-x_2\beta_2} e^{-x_3\beta_3} e^{-x_{\mathrm{min}}\beta_4} e^{-x_5\beta_5}} \\
&= \frac{e^{-x_{\mathrm{maj}}\beta_4}}{e^{-x_{\mathrm{min}}\beta_4}} \\
&= e^{-\beta_4}
\end{aligned}
$$

Hazard ratio greater than 1 would imply that majority governments fall faster (shorter survival time) than minority governments.


Distribution of Hazard Ratios
Mean Hazard Ratio
Density
0
1
2
3
4
0.4
0.5
0.6
0.7
0.8
0.9
1.0
hazard ratios



Majority governments survive longer than minority governments.

Expected (average) Survival Time

Expected (average) Survival Time

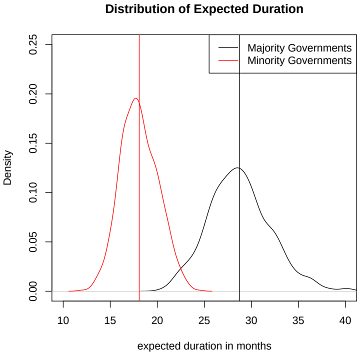
$$E(T|\mathbf{x}_i) = \theta_i$$

Expected (average) Survival Time

$$\begin{aligned} E(T|\mathbf{x}_i) &= \theta_i \\ &= \exp[\mathbf{x}_i\beta] \end{aligned}$$

Expected (average) Survival Time

$$
\begin{aligned}
E(T|\mathbf{x}_i) &= \theta_i \\
&= \exp[\mathbf{x}_i\beta]
\end{aligned}
$$



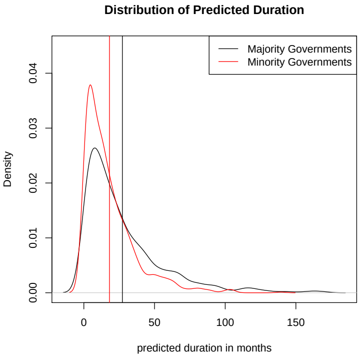
Predicted Survival Time

Predicted Survival Time

Draw predicted values from the exponential distribution.

Predicted Survival Time

Draw predicted values from the exponential distribution.



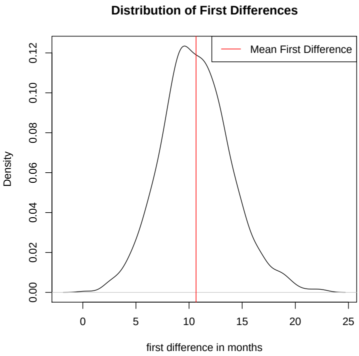
First Differences

First Differences

$$E(T|\mathbf{x}_{\text{maj}}) - E(T|\mathbf{x}_{\text{min}})$$

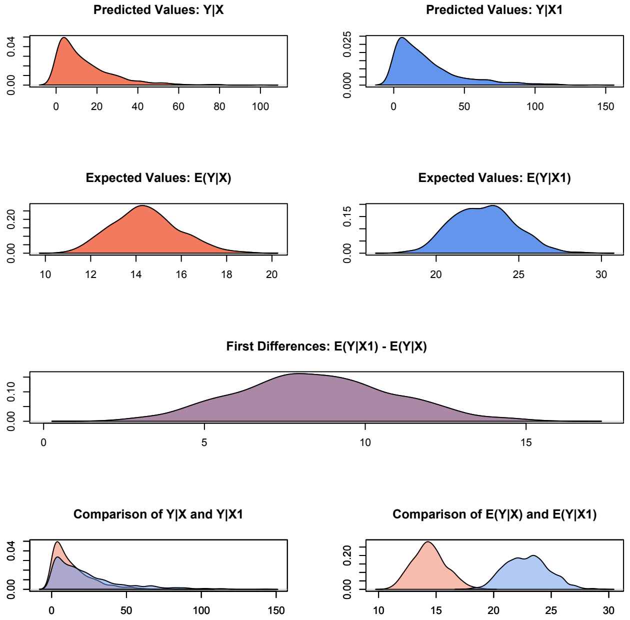
First Differences

$$E(T|\mathbf{x}_{\text{maj}}) - E(T|\mathbf{x}_{\text{min}})$$



Quantities of Interest in Zelig

```
x.min <- setx(z.out,numst2=0)
x.maj <- setx(z.out,numst2=1)
s.out <- sim(z.out, x=x.min,x1=x.maj)
summary(s.out)
plot(s.out)
```


Predicted Values: Y|X
0.00
0.04
0
20
40
60
80
100
Predicted Values: Y|X1
0.000
0.025
0
50
100
150
Expected Values: E(Y|X)
0.00
0.20
10
12
14
16
18
20
Expected Values: E(Y|X1)
0.00
0.15
20
25
30
First Differences: E(Y|X1) - E(Y|X)
0.00
0.10
0
5
10
15
Comparison of Y|X and Y|X1
0.00
0.04
0
50
100
150
Comparison of E(Y|X) and E(Y|X1)
0.00
0.20
10
15
20
25
30

The exponential model is nice and simple, but the assumption of a flat hazard may be too restrictive.

The exponential model is nice and simple, but the assumption of a flat hazard may be too restrictive.

What if we want to loosen that restriction by assuming a monotonic hazard?

The exponential model is nice and simple, but the assumption of a flat hazard may be too restrictive.

What if we want to loosen that restriction by assuming a monotonic hazard?

We can use the Weibull model.

The Weibull Model

Similar to how we generalized the Poisson into a Negative Binomial by adding a parameter, we can do the same with the Exponential by turning it into a Weibull:

The Weibull Model

Similar to how we generalized the Poisson into a Negative Binomial by adding a parameter, we can do the same with the Exponential by turning it into a Weibull:

$$T_i \sim \text{Weibull}(\theta_i, \alpha)$$

The Weibull Model

Similar to how we generalized the Poisson into a Negative Binomial by adding a parameter, we can do the same with the Exponential by turning it into a Weibull:

$$\begin{aligned} T_i &\sim \text{Weibull}(\theta_i, \alpha) \\ E(T_i) &= \theta_i \Gamma\left(1 + \frac{1}{\alpha}\right) \end{aligned}$$

The Weibull Model

Similar to how we generalized the Poisson into a Negative Binomial by adding a parameter, we can do the same with the Exponential by turning it into a Weibull:

$$\begin{aligned} T_i &\sim \text{Weibull}(\theta_i, \alpha) \\ E(T_i) &= \theta_i \Gamma\left(1 + \frac{1}{\alpha}\right) \end{aligned}$$

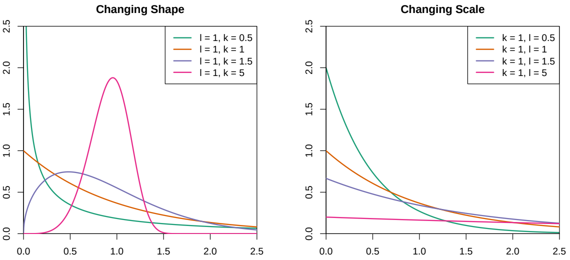
$\theta_i > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter.

The Weibull Model

Similar to how we generalized the Poisson into a Negative Binomial by adding a parameter, we can do the same with the Exponential by turning it into a Weibull:

$$\begin{aligned} T_i &\sim \text{Weibull}(\theta_i, \alpha) \\ E(T_i) &= \theta_i \Gamma\left(1 + \frac{1}{\alpha}\right) \end{aligned}$$

$\theta_i > 0$ is the scale parameter and $\alpha > 0$ is the shape parameter.



The Weibull Model

$$f(t_i) = \left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right]$$

The Weibull Model

$$f(t_i) \;=\; \left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right]$$

Model θ_i with covariates in the systematic component:

$$\theta_i = \exp(x_i \beta)$$

The Weibull Model

$$f(t_i) \;=\; \left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right]$$

Model θ_i with covariates in the systematic component:

$$\theta_i = \exp(x_i\beta)$$

Positive β implies that expected duration time increases as x increases.

$$f(t_i) = \left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right]$$

$$
\begin{aligned}
f(t_i) &= \left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right] \\
S(t_i) &= 1 - F(t_i)
\end{aligned}
$$

$$
\begin{aligned}
f(t_i) &= \left(\frac{\alpha}{\theta_i^\alpha}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^\alpha\right] \\
S(t_i) &= 1 - F(t_i) \\
&= 1 - (1 - e^{-(t_i/\theta_i)^\alpha}) \\
&= e^{-(t_i/\theta_i)^\alpha}
\end{aligned}
$$

$$
\begin{aligned}
f(t_i) &= \left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right]
\end{aligned}
$$

$$
\begin{aligned}
S(t_i) &= 1 - F(t_i) \\
&= 1 - (1 - e^{-(t_i/\theta_i)^{\alpha}}) \\
&= e^{-(t_i/\theta_i)^{\alpha}}
\end{aligned}
$$

$$
h(t_i) = \frac{f(t_i)}{S(t_i)}
$$

$$
\begin{aligned}
f(t_i) &= \left(\frac{\alpha}{\theta_i^\alpha}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^\alpha\right]
\end{aligned}
$$

$$
\begin{aligned}
S(t_i) &= 1 - F(t_i) \\
&= 1 - (1 - e^{-(t_i/\theta_i)^\alpha}) \\
&= e^{-(t_i/\theta_i)^\alpha}
\end{aligned}
$$

$$
\begin{aligned}
h(t_i) &= \frac{f(t_i)}{S(t_i)} \\
&= \frac{\left(\frac{\alpha}{\theta_i^\alpha}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^\alpha\right]}{e^{-(t_i/\theta_i)^\alpha}}
\end{aligned}
$$

$$
\begin{aligned}
f(t_i) &= \left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right] \\
S(t_i) &= 1 - F(t_i) \\
&= 1 - (1 - e^{-(t_i/\theta_i)^{\alpha}}) \\
&= e^{-(t_i/\theta_i)^{\alpha}}
\end{aligned}
$$

$$
\begin{aligned}
h(t_i) &= \frac{f(t_i)}{S(t_i)} \\
&= \frac{\left(\frac{\alpha}{\theta_i^{\alpha}}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^{\alpha}\right]}{e^{-(t_i/\theta_i)^{\alpha}}} \\
&= \left(\frac{\alpha}{\theta_i}\right)\left(\frac{t_i}{\theta_i}\right)^{\alpha-1}
\end{aligned}
$$

$$
\begin{aligned}
f(t_i) &= \left(\frac{\alpha}{\theta_i^\alpha}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^\alpha\right]
\end{aligned}
$$

$$
\begin{aligned}
S(t_i) &= 1 - F(t_i) \\
&= 1 - (1 - e^{-(t_i/\theta_i)^\alpha}) \\
&= e^{-(t_i/\theta_i)^\alpha}
\end{aligned}
$$

$$
\begin{aligned}
h(t_i) &= \frac{f(t_i)}{S(t_i)} \\
&= \frac{\left(\frac{\alpha}{\theta_i^\alpha}\right) t_i^{\alpha-1} \exp\left[-\left(\frac{t_i}{\theta_i}\right)^\alpha\right]}{e^{-(t_i/\theta_i)^\alpha}} \\
&= \left(\frac{\alpha}{\theta_i}\right)\left(\frac{t_i}{\theta_i}\right)^{\alpha-1} \\
&= \left(\frac{\alpha}{\theta_i^\alpha}\right) t_i^{\alpha-1}
\end{aligned}
$$

Hazard monotonicity assumption

$h(t_i)$ is modeled with both λ_i and α and is a function of t_i. Thus, the Weibull model assumes a **monotonic hazard**.

Hazard monotonicity assumption

$h(t_i)$ is modeled with both λ_i and α and is a function of t_i. Thus, the Weibull model assumes a **monotonic hazard**.

- If $\alpha = 1$, $h(t_i)$ is flat and the model is the exponential model.

Hazard monotonicity assumption

$h(t_i)$ is modeled with both λ_i and α and is a function of t_i. Thus, the Weibull model assumes a **monotonic hazard**.

- If $\alpha = 1$, $h(t_i)$ is flat and the model is the exponential model.
- If $\alpha > 1$, $h(t_i)$ is monotonically increasing.

Hazard monotonicity assumption

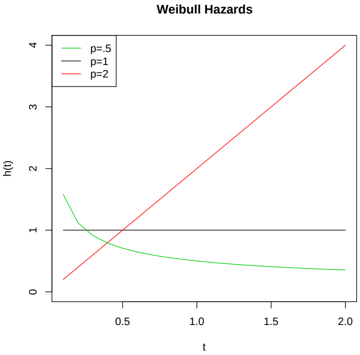
$h(t_i)$ is modeled with both λ_i and α and is a function of t_i. Thus, the Weibull model assumes a **monotonic hazard**.

- If $\alpha = 1$, $h(t_i)$ is flat and the model is the exponential model.
- If $\alpha > 1$, $h(t_i)$ is monotonically increasing.
- If $\alpha < 1$, $h(t_i)$ is monotonically decreasing.

Hazard monotonicity assumption

$h(t_i)$ is modeled with both λ_i and α and is a function of t_i. Thus, the Weibull model assumes a **monotonic hazard**.

- If $\alpha = 1$, $h(t_i)$ is flat and the model is the exponential model.
- If $\alpha > 1$, $h(t_i)$ is monotonically increasing.
- If $\alpha < 1$, $h(t_i)$ is monotonically decreasing.



The shape parameter α for the Weibull distribution is the reciprocal of the scale parameter given by `survreg()`.

The shape parameter α for the Weibull distribution is the reciprocal of the scale parameter given by `survreg()`.

The scale parameter given by `survreg()` is NOT the same as the scale parameter in the Weibull distribution, which should be $\theta_i = e^{\mathbf{x}_i \beta}$.

Hazard Ratios

Hazard Ratios

One quantity of interest is the hazard ratio:

Hazard Ratios

One quantity of interest is the hazard ratio:

$$HR = \frac{h(t|x=1)}{h(t|x=0)}$$

Hazard Ratios

One quantity of interest is the hazard ratio:

$$HR = \frac{h(t|x=1)}{h(t|x=0)}$$

With the Weibull model we make a **proportional hazards** assumption: hazard ratio does not depend t.

Other Parametric Models

- Gompertz model: monotonic hazard

Other Parametric Models

- Gompertz model: monotonic hazard
- Log-logistic or log-normal model: nonmonotonic hazard

Other Parametric Models

- Gompertz model: monotonic hazard
- Log-logistic or log-normal model: nonmonotonic hazard
- Generalized gamma model: nests the exponential, Weibull, log-normal, and gamma models with an extra parameter (see appendix slides)

Other Parametric Models

- Gompertz model: monotonic hazard
- Log-logistic or log-normal model: nonmonotonic hazard
- Generalized gamma model: nests the exponential, Weibull, log-normal, and gamma models with an extra parameter (see appendix slides)

But what if we don't want to make an assumption about the shape of the hazard?

The Cox Proportional Hazards Model

The Cox Proportional Hazards Model

Often described as a semi-parametric model.

The Cox Proportional Hazards Model

Often described as a semi-parametric model.
Pros:

The Cox Proportional Hazards Model

Often described as a semi-parametric model.
Pros:

- Makes no restrictive assumption about the shape of the hazard.

The Cox Proportional Hazards Model

Often described as a semi-parametric model.
Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

The Cox Proportional Hazards Model

Often described as a semi-parametric model.
Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

The Cox Proportional Hazards Model

Often described as a semi-parametric model.

Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

- Only quantities of interest are hazard ratios.

The Cox Proportional Hazards Model

Often described as a semi-parametric model.

Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

- Only quantities of interest are hazard ratios.
- Can be subject to overfitting

The Cox Proportional Hazards Model

Often described as a semi-parametric model.
Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

- Only quantities of interest are hazard ratios.
- Can be subject to overfitting
- Shape of hazard is unknown (although there are semi-parametric ways to derive the hazard and survivor functions)

1. Reconceptualize each t_i as a discrete event time rather than a duration or survival time (non-censored observations only).

1. Reconceptualize each t_i as a discrete event time rather than a duration or survival time (non-censored observations only).
   - $t_i = 5$: An event occurred at month 5, rather than observation i surviving for 5 months.

1. Reconceptualize each t_i as a discrete event time rather than a duration or survival time (non-censored observations only).
   - $t_i = 5$: An event occurred at month 5, rather than observation i surviving for 5 months.
2. Assume there are no tied event times in the data.

1. Reconceptualize each t_i as a discrete event time rather than a duration or survival time (non-censored observations only).
    - $t_i = 5$: An event occurred at month 5, rather than observation i surviving for 5 months.
2. Assume there are no tied event times in the data.

1. Reconceptualize each t_i as a discrete event time rather than a duration or survival time (non-censored observations only).
   - $t_i = 5$: An event occurred at month 5, rather than observation i surviving for 5 months.
2. Assume there are no tied event times in the data.
   - No two events can occur at the same instant. It only seems that way because our unit of measurement is not precise enough.

1. Reconceptualize each t_i as a discrete event time rather than a duration or survival time (non-censored observations only).
   - $t_i = 5$: An event occurred at month 5, rather than observation i surviving for 5 months.
2. Assume there are no tied event times in the data.
   - No two events can occur at the same instant. It only seems that way because our unit of measurement is not precise enough.
   - There are ways to adjust the likelihood to take into account observed ties.

1. Reconceptualize each t_i as a discrete event time rather than a duration or survival time (non-censored observations only).
   - $t_i = 5$: An event occurred at month 5, rather than observation i surviving for 5 months.
2. Assume there are no tied event times in the data.
   - No two events can occur at the same instant. It only seems that way because our unit of measurement is not precise enough.
   - There are ways to adjust the likelihood to take into account observed ties.
3. Assume no events can happen between event times.

We know that exactly one event occurred at each t_i for all non-censored i.

We know that exactly one event occurred at each t_i for all non-censored i.

Define a risk set R_i as the set of all possible observations at risk of an event at time t_i.

We know that exactly one event occurred at each t_i for all non-censored i.

Define a risk set R_i as the set of all possible observations at risk of an event at time t_i.

What observations belong in R_i?

We know that exactly one event occurred at each t_i for all non-censored i.

Define a risk set R_i as the set of all possible observations at risk of an event at time t_i.

What observations belong in R_i?

All observations (censored and non-censored) j such that $t_j \geq t_i$

We know that exactly one event occurred at each t_i for all non-censored i.

Define a risk set R_i as the set of all possible observations at risk of an event at time t_i.

What observations belong in R_i?

All observations (censored and non-censored) j such that $t_j \geq t_i$

For example, if $t_i = 5$ months, then all observations that do not experience the event or are not censored before 5 months are at risk.

We can then create a partial likelihood function:

We can then create a partial likelihood function:

$$\mathcal{L} = \prod_{i=1}^{n} [P(\text{event occurred in } i | \text{event occurred in } R_i)]^{c_i}$$

We can then create a partial likelihood function:

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \left[P(\text{event occurred in } i | \text{event occurred in } R_i)\right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{P(\text{event occurred in } i)}{P(\text{event occurred in } R_i)}\right]^{c_i}
\end{aligned}$$

We can then create a partial likelihood function:

$$\begin{aligned}\mathcal{L} &= \prod_{i=1}^{n} \left[P(\text{event occurred in } i | \text{event occurred in } R_i)\right]^{c_i} \\ &= \prod_{i=1}^{n} \left[\frac{P(\text{event occurred in } i)}{P(\text{event occurred in } R_i)}\right]^{c_i} \\ &= \prod_{i=1}^{n} \left[\frac{h(t_i)}{\sum_{j \in R_i} h(t_j)}\right]^{c_i}\end{aligned}$$

We can then create a partial likelihood function:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [P(\text{event occurred in } i | \text{event occurred in } R_i)]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{P(\text{event occurred in } i)}{P(\text{event occurred in } R_i)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h(t_i)}{\sum_{j \in R_i} h(t_j)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h_0(t) h_i(t_i)}{\sum_{j \in R_i} h_0(t) h_j(t_j)} \right]^{c_i}
\end{aligned}
$$

We can then create a partial likelihood function:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} \left[P(\text{event occurred in } i | \text{event occurred in } R_i) \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{P(\text{event occurred in } i)}{P(\text{event occurred in } R_i)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h(t_i)}{\sum_{j \in R_i} h(t_j)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h_0(t) h_i(t_i)}{\sum_{j \in R_i} h_0(t) h_j(t_j)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h_i(t_i)}{\sum_{j \in R_i} h_j(t_j)} \right]^{c_i}
\end{aligned}
$$

We can then create a partial likelihood function:

$$
\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n} [P(\text{event occurred in } i | \text{event occurred in } R_i)]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{P(\text{event occurred in } i)}{P(\text{event occurred in } R_i)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h(t_i)}{\sum_{j \in R_i} h(t_j)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h_0(t) h_i(t_i)}{\sum_{j \in R_i} h_0(t) h_j(t_j)} \right]^{c_i} \\
&= \prod_{i=1}^{n} \left[\frac{h_i(t_i)}{\sum_{j \in R_i} h_j(t_j)} \right]^{c_i}
\end{aligned}
$$

$h_0(t)$ is the baseline hazard, which is the same for all observations, so it cancels out.

Like in parametric models, $h(t)$ is modeled with covariates:

Like in parametric models, $h(t)$ is modeled with covariates:

$$h_i(t_i) = e^{\mathbf{x}_i \beta}$$

Like in parametric models, $h(t)$ is modeled with covariates:

$$h_i(t_i) = e^{\mathbf{x}_i \beta}$$

Note that a positive β now suggests that an increase in x increases the hazard and decreases survival time.

Like in parametric models, $h(t)$ is modeled with covariates:

$$h_i(t_i) = e^{\mathbf{x}_i \beta}$$

Note that a positive β now suggests that an increase in x increases the hazard and decreases survival time.

$$\mathcal{L} = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i \beta}}{\sum_{j \in R_i} e^{\mathbf{x}_j \beta}} \right]^{c_i}$$

Like in parametric models, $h(t)$ is modeled with covariates:

$$h_i(t_i) = e^{\mathbf{x}_i \beta}$$

Note that a positive β now suggests that an increase in x increases the hazard and decreases survival time.

$$\mathcal{L} = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i \beta}}{\sum_{j \in R_i} e^{\mathbf{x}_j \beta}} \right]^{c_i}$$

There is no β_0 term estimated.

Like in parametric models, $h(t)$ is modeled with covariates:

$$h_i(t_i) = e^{\mathbf{x}_i \beta}$$

Note that a positive β now suggests that an increase in x increases the hazard and decreases survival time.

$$\mathcal{L} = \prod_{i=1}^{n} \left[\frac{e^{\mathbf{x}_i \beta}}{\sum_{j \in R_i} e^{\mathbf{x}_j \beta}} \right]^{c_i}$$

There is no β_0 term estimated. This implies that the shape of the baseline hazard is left unmodeled.

Pros:

Pros:

- Makes no restrictive assumption about the shape of the hazard.

Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

- Only quantities of interest are hazard ratios.

Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

- Only quantities of interest are hazard ratios.
- Can be subject to overfitting

Pros:

- Makes no restrictive assumption about the shape of the hazard.
- A better choice if you want the effects of the covariates and the nature of the time dependence is unimportant.

Cons:

- Only quantities of interest are hazard ratios.
- Can be subject to overfitting
- Shape of hazard is unknown (although there are semi-parametric ways to derive the hazard and survivor functions)

How do I run a Cox proportional hazards model in R?

How do I run a Cox proportional hazards model in R?

Use the `coxph()` function in the `survival` package (also in the `Design` and `Zelig` packages).

Alternatives

- Survival models are cool . . .

Alternatives

- Survival models are cool . . . but hard.

Alternatives

- Survival models are cool . . . but hard.
- There are other things you can model:

Alternatives

- Survival models are cool . . . but hard.
- There are other things you can model:
  - Perhaps some observations are more likely to fail than others: frailty models

Alternatives

- Survival models are cool ... but hard.
- There are other things you can model:
  - Perhaps some observations are more likely to fail than others: frailty models
  - Perhaps some observations you don't expect to fail at all: split population models

Alternatives

- Survival models are cool ... but hard.
- There are other things you can model:
  - Perhaps some observations are more likely to fail than others: frailty models
  - Perhaps some observations you don't expect to fail at all: split population models
  - Perhaps there can be more than one type of event: competing risks model

Alternatives

- Survival models are cool . . . but hard.
- There are other things you can model:
  - Perhaps some observations are more likely to fail than others: frailty models
  - Perhaps some observations you don't expect to fail at all: split population models
  - Perhaps there can be more than one type of event: competing risks model

If you encounter survival data think carefully about the process and then choose a corresponding model.

References:

Box-Steffensmeier, Janet M. and Bradford S. Jones. 2004. Event History Modeling. Cambridge University Press.

Therneau, Terry M., and Patricia M. Grambsch. 2013 Modeling survival data: extending the Cox model. Springer Science & Business Media.

Andersen, Per Kragh, et al. 2012 Statistical models based on counting processes. Springer Science & Business Media.

How Do Survival Models Relate to Duration Dependence in a Logit Model?

How Do Survival Models Relate to Duration Dependence in a Logit Model?

- Based on Beck, Katz, and Tucker (1998)

How Do Survival Models Relate to Duration Dependence in a Logit Model?

- Based on Beck, Katz, and Tucker (1998)
- Suppose we have Time-Series Cross-Sectional Data with a binary dependent variable.

How Do Survival Models Relate to Duration Dependence in a Logit Model?

- Based on Beck, Katz, and Tucker (1998)
- Suppose we have Time-Series Cross-Sectional Data with a binary dependent variable.
  - For example, if we had data on country dyads over 50 years, with the dependent variable being whether there was a war between the two countries in each year.

How Do Survival Models Relate to Duration Dependence in a Logit Model?

- Based on Beck, Katz, and Tucker (1998)
- Suppose we have Time-Series Cross-Sectional Data with a binary dependent variable.
  - For example, if we had data on country dyads over 50 years, with the dependent variable being whether there was a war between the two countries in each year.
- Not all observations are independent. We may see some duration dependence.

How Do Survival Models Relate to Duration Dependence in a Logit Model?

- Based on Beck, Katz, and Tucker (1998)
- Suppose we have Time-Series Cross-Sectional Data with a binary dependent variable.
  - For example, if we had data on country dyads over 50 years, with the dependent variable being whether there was a war between the two countries in each year.
- Not all observations are independent. We may see some duration dependence.
  - Perhaps countries that have been at peace for 100 years may be less likely to go to war than countries that have been at peace for only 2 years.

How Do Survival Models Relate to Duration Dependence in a Logit Model?

- Based on Beck, Katz, and Tucker (1998)
- Suppose we have Time-Series Cross-Sectional Data with a binary dependent variable.
  - For example, if we had data on country dyads over 50 years, with the dependent variable being whether there was a war between the two countries in each year.
- Not all observations are independent. We may see some duration dependence.
  - Perhaps countries that have been at peace for 100 years may be less likely to go to war than countries that have been at peace for only 2 years.

How can we account for this duration dependence in a logit model?

Think of the observations as grouped duration data:

Think of the observations as grouped duration data:

Year	t_k	Dyad	Y_i	T_i
1992	1	US-Iraq	0	
1993	2	US-Iraq	0	
1994	3	US-Iraq	0	
1995	4	US-Iraq	0	
1996	5	US-Iraq	0	
1997	6	US-Iraq	0	
1998	7	US-Iraq	0	12
1999	8	US-Iraq	0	
2000	9	US-Iraq	0	
2001	10	US-Iraq	0	
2002	11	US-Iraq	0	
2003	12	US-Iraq	1	

Then we end up with:

$$P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) \quad = \quad h(t_k | \mathbf{x}_{i,t_k})$$

Then we end up with:

$$\begin{aligned} P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) &= h(t_k | \mathbf{x}_{i,t_k}) \\ &= 1 - P(\text{surviving beyond } t_k | \text{survival up to } t_{k-1}) \end{aligned}$$

Then we end up with:

$$\begin{aligned} P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) &= h(t_k | \mathbf{x}_{i,t_k}) \\ &= 1 - P(\text{surviving beyond } t_k | \text{survival up to } t_{k-1}) \end{aligned}$$

It can be shown in general that

$$S(t) = e^{-\int_0^t h(u) du}$$

Then we end up with:

$$\begin{aligned} P(y_{i,t_k} = 1|\mathbf{x}_{i,t_k}) &= h(t_k|\mathbf{x}_{i,t_k}) \\ &= 1 - P(\text{surviving beyond } t_k|\text{survival up to } t_{k-1}) \end{aligned}$$

It can be shown in general that

$$S(t) = e^{-\int_0^t h(u)du}$$

So then we get

$$P(y_{i,t_k} = 1|\mathbf{x}_{i,t_k}) = 1 - e^{-\int_{t_{k-1}}^{t_k} h(u)du}$$

where we take the integral from t_{k-1} to t_k in order to get the conditional survival.

$$P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) \quad = \quad 1 - \exp\left(-\int_{t_{k-1}}^{t_k} h(u) du\right)$$

$$
\begin{aligned}
P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) &= 1 - \exp\left(- \int_{t_{k-1}}^{t_k} h(u) du \right) \\
&= 1 - \exp\left(- \int_{t_{k-1}}^{t_k} e^{\mathbf{x}_{i,t_k} \beta} h_0(u) du \right)
\end{aligned}
$$

$$\begin{aligned} P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) &= 1 - \exp\left(-\int_{t_{k-1}}^{t_k} h(u) du\right) \\ &= 1 - \exp\left(-\int_{t_{k-1}}^{t_k} e^{\mathbf{x}_{i,t_k}\beta} h_0(u) du\right) \\ &= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta} \int_{t_{k-1}}^{t_k} h_0(u) du\right) \end{aligned}$$

$$
\begin{aligned}
P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) &= 1 - \exp\left(- \int_{t_{k-1}}^{t_k} h(u) du \right) \\
&= 1 - \exp\left(- \int_{t_{k-1}}^{t_k} e^{\mathbf{x}_{i,t_k}\beta} h_0(u) du \right) \\
&= 1 - \exp\left(- e^{\mathbf{x}_{i,t_k}\beta} \int_{t_{k-1}}^{t_k} h_0(u) du \right) \\
&= 1 - \exp\left(- e^{\mathbf{x}_{i,t_k}\beta} \alpha_{t_k} \right)
\end{aligned}
$$

$$
\begin{aligned}
P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) &= 1 - \exp\left(- \int_{t_{k-1}}^{t_k} h(u) du \right) \\
&= 1 - \exp\left(- \int_{t_{k-1}}^{t_k} e^{\mathbf{x}_{i,t_k}\beta} h_0(u) du \right) \\
&= 1 - \exp\left(- e^{\mathbf{x}_{i,t_k}\beta} \int_{t_{k-1}}^{t_k} h_0(u) du \right) \\
&= 1 - \exp\left(- e^{\mathbf{x}_{i,t_k}\beta} \alpha_{t_k} \right) \\
&= 1 - \exp\left(- e^{\mathbf{x}_{i,t_k}\beta + \kappa_{t_k}} \right)
\end{aligned}
$$

$$
\begin{aligned}
P(y_{i,t_k} = 1 | \mathbf{x}_{i,t_k}) &= 1 - \exp\left(-\int_{t_{k-1}}^{t_k} h(u)du\right) \\
&= 1 - \exp\left(-\int_{t_{k-1}}^{t_k} e^{\mathbf{x}_{i,t_k}\beta} h_0(u)du\right) \\
&= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta} \int_{t_{k-1}}^{t_k} h_0(u)du\right) \\
&= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta} \alpha_{t_k}\right) \\
&= 1 - \exp\left(-e^{\mathbf{x}_{i,t_k}\beta + \kappa_{t_k}}\right)
\end{aligned}
$$

This is equivalent to a model with a complementary log-log (cloglog) link and time dummies κ_{t_k}.

- BKT suggest using a logit link instead of a cloglog link because logit is more widely used (and nobody knows what a cloglog link is except you guys!).

- BKT suggest using a logit link instead of a cloglog link because logit is more widely used (and nobody knows what a cloglog link is except you guys!).
- As long as probability of an event does not exceed 50 percent, logit and cloglog links are very similar.

- BKT suggest using a logit link instead of a cloglog link because logit is more widely used (and nobody knows what a cloglog link is except you guys!).
- As long as probability of an event does not exceed 50 percent, logit and cloglog links are very similar.
- The use of time dummies means that we are imposing no structure on the nature of duration dependence (structure of the hazard).

- BKT suggest using a logit link instead of a cloglog link because logit is more widely used (and nobody knows what a cloglog link is except you guys!).
- As long as probability of an event does not exceed 50 percent, logit and cloglog links are very similar.
- The use of time dummies means that we are imposing no structure on the nature of duration dependence (structure of the hazard).
- If we don't use time dummies, we are assuming no duration dependence (flat hazard)

- BKT suggest using a logit link instead of a cloglog link because logit is more widely used (and nobody knows what a cloglog link is except you guys!).
- As long as probability of an event does not exceed 50 percent, logit and cloglog links are very similar.
- The use of time dummies means that we are imposing no structure on the nature of duration dependence (structure of the hazard).
- If we don't use time dummies, we are assuming no duration dependence (flat hazard)
- Using a variable such as "number of years at peace" instead of time dummies imposes a monotonic hazard.

- BKT suggest using a logit link instead of a cloglog link because logit is more widely used (and nobody knows what a cloglog link is except you guys!).
- As long as probability of an event does not exceed 50 percent, logit and cloglog links are very similar.
- The use of time dummies means that we are imposing no structure on the nature of duration dependence (structure of the hazard).
- If we don't use time dummies, we are assuming no duration dependence (flat hazard)
- Using a variable such as "number of years at peace" instead of time dummies imposes a monotonic hazard.
- The use of time dummies may use up a lot of degrees of freedom, so BKT suggest using restricted cubic splines.

Possible complications:

Possible complications:

- Multiple events

Possible complications:

- Multiple events
  - Assumes that multiple events are independent (independence of observations assumption in a survival model).

Possible complications:

- Multiple events
  - Assumes that multiple events are independent (independence of observations assumption in a survival model).
- Left censoring

Possible complications:

- Multiple events
  - Assumes that multiple events are independent (independence of observations assumption in a survival model).
- Left censoring
  - Countries may have been at peace long before we start observing data, and we don't know when that "peace duration" began.

Possible complications:

- Multiple events
  - Assumes that multiple events are independent (independence of observations assumption in a survival model).
- Left censoring
  - Countries may have been at peace long before we start observing data, and we don't know when that "peace duration" began.
- Variables that do not vary across units

Possible complications:

- Multiple events
  - Assumes that multiple events are independent (independence of observations assumption in a survival model).
- Left censoring
  - Countries may have been at peace long before we start observing data, and we don't know when that "peace duration" began.
- Variables that do not vary across units
  - May be collinear with time dummies.

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit
  - categorical: multinomial logit, multinomial probit

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit
  - categorical: multinomial logit, multinomial probit
  - duration: exponential, weibull, cox proportional hazards

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit
  - categorical: multinomial logit, multinomial probit
  - duration: exponential, weibull, cox proportional hazards
- Each model followed a similar pattern:
  1. define a model with a stochastic and systematic component

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit
  - categorical: multinomial logit, multinomial probit
  - duration: exponential, weibull, cox proportional hazards
- Each model followed a similar pattern:
  1. define a model with a stochastic and systematic component
  2. derive the log-likelihood

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit
  - categorical: multinomial logit, multinomial probit
  - duration: exponential, weibull, cox proportional hazards
- Each model followed a similar pattern:
  1. define a model with a stochastic and systematic component
  2. derive the log-likelihood
  3. estimate via MLE

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit
  - categorical: multinomial logit, multinomial probit
  - duration: exponential, weibull, cox proportional hazards
- Each model followed a similar pattern:
  1. define a model with a stochastic and systematic component
  2. derive the log-likelihood
  3. estimate via MLE
  4. interpret quantities of interest

First Half of Course Summary

- In the first six weeks we have covered: maximum likelihood estimation, generalized linear models and a general approach to quantities of interest
- We touched on at least briefly on standard models for most types of data you will encounter:
  - continuous: normal linear model
  - binary: logit, probit
  - count: poisson, negative binomial
  - ordered: ordinal probit, ordinal logit
  - categorical: multinomial logit, multinomial probit
  - duration: exponential, weibull, cox proportional hazards
- Each model followed a similar pattern:
  1. define a model with a stochastic and systematic component
  2. derive the log-likelihood
  3. estimate via MLE
  4. interpret quantities of interest
- You can now interpret these models and learn new ones.

Preview

Preview

- Weeks 7-8 Missing Data
  - Mixture Models and the Expectation Maximization algorithm
  - Missing Data and Multiple Imputation

Preview

- Weeks 7-8 Missing Data
  - Mixture Models and the Expectation Maximization algorithm
  - Missing Data and Multiple Imputation
- Weeks 9-10 Causal Inference
  - Model Dependence and Matching
  - Explanation in Causal Inference with Moderation/Mediation

Preview

- Weeks 7-8 Missing Data
  - Mixture Models and the Expectation Maximization algorithm
  - Missing Data and Multiple Imputation
- Weeks 9-10 Causal Inference
  - Model Dependence and Matching
  - Explanation in Causal Inference with Moderation/Mediation
- Weeks 11-12 Hierarchical Models
  - Regularization and Hierarchical Models
  - More Hierarchical Models and Wrap-up

Preview

- Weeks 7-8 Missing Data
  - Mixture Models and the Expectation Maximization algorithm
  - Missing Data and Multiple Imputation
- Weeks 9-10 Causal Inference
  - Model Dependence and Matching
  - Explanation in Causal Inference with Moderation/Mediation
- Weeks 11-12 Hierarchical Models
  - Regularization and Hierarchical Models
  - More Hierarchical Models and Wrap-up

These topics are a long-term bet on things that will be important in your career. Also a short case-study in reading into a new statistical literature.

Identification

Identification

- Reading: Unifying Political Methodology, Chapter 8

Identification

- Reading: Unifying Political Methodology, Chapter 8
- Definition of "identification"

Identification

- Reading: Unifying Political Methodology, Chapter 8
- Definition of "identification"
  - Qualitative: we can learn the parameter with infinite draws

Identification

- Reading: Unifying Political Methodology, Chapter 8
- Definition of "identification"
  - Qualitative: we can learn the parameter with infinite draws
  - Mathematical: Each parameter value produces unique likelihood value

Identification

- Reading: Unifying Political Methodology, Chapter 8
- Definition of "identification"
  - Qualitative: we can learn the parameter with infinite draws
  - Mathematical: Each parameter value produces unique likelihood value
  - Graphical: A likelihood with a plateau at the maximum

Identification

- Reading: Unifying Political Methodology, Chapter 8
- Definition of "identification"
  - Qualitative: we can learn the parameter with infinite draws
  - Mathematical: Each parameter value produces unique likelihood value
  - Graphical: A likelihood with a plateau at the maximum
- Partially identified models: the likelihood is informative but not about a single point

Identification

- Reading: Unifying Political Methodology, Chapter 8
- Definition of "identification"
  - Qualitative: we can learn the parameter with infinite draws
  - Mathematical: Each parameter value produces unique likelihood value
  - Graphical: A likelihood with a plateau at the maximum
- Partially identified models: the likelihood is informative but not about a single point
- Non-identified models: include those that make little sense, even if hard to tell.

Example 1: Flat Likelihoods

Example 1: Flat Likelihoods

A (dumb) model:

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

$$\lambda_i = 1 + 0\beta$$

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

$$\lambda_i = 1 + 0\beta$$

What do we know about β? (from the likelihood perspective)

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

$$\lambda_i = 1 + 0\beta$$

What do we know about β? (from the likelihood perspective)

$$L(\lambda|y) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

$$\lambda_i = 1 + 0\beta$$

What do we know about β? (from the likelihood perspective)

$$L(\lambda|y) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

and the log-likelihood, with $(1 + 0\beta)$ substituted for λ_i:

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

$$\lambda_i = 1 + 0\beta$$

What do we know about β? (from the likelihood perspective)

$$L(\lambda|y) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

and the log-likelihood, with $(1 + 0\beta)$ substituted for λ_i:

$$\ln L(\beta|y) = \sum_{i=1}^{n} \{-(0\beta + 1) - y_i \ln(0\beta + 1)\}$$

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

$$\lambda_i = 1 + 0\beta$$

What do we know about β? (from the likelihood perspective)

$$L(\lambda|y) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

and the log-likelihood, with $(1 + 0\beta)$ substituted for λ_i:

$$\begin{aligned}\ln L(\beta|y) &= \sum_{i=1}^{n} \{-(0\beta + 1) - y_i \ln(0\beta + 1)\} \\ &= \sum_{i=1}^{n} -1\end{aligned}$$

Example 1: Flat Likelihoods

A (dumb) model:

$$Y_i \sim f_p(y_i|\lambda_i)$$

$$\lambda_i = 1 + 0\beta$$

What do we know about β? (from the likelihood perspective)

$$L(\lambda|y) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!}$$

and the log-likelihood, with $(1 + 0\beta)$ substituted for λ_i:

$$\begin{aligned}
\ln L(\beta|y) &= \sum_{i=1}^{n} \{-(0\beta + 1) - y_i \ln(0\beta + 1)\} \\
&= \sum_{i=1}^{n} -1 \\
&= -n
\end{aligned}$$

Example 1: Flat Likelihoods

Example 1: Flat Likelihoods

1. An identified likelihood has a unique maximum.

Example 1: Flat Likelihoods

1. An identified likelihood has a unique maximum.
2. A likelihood function with a flat region or plateau at the maximum is not identified.

Example 1: Flat Likelihoods

1. An identified likelihood has a unique maximum.
2. A likelihood function with a flat region or plateau at the maximum is not identified.
3. A likelihood with a plateau can be informative, but a unique MLE doesn't exist

Example 2: Non-unique Reparameterization

Example 2: Non-unique Reparameterization

A model

Example 2: Non-unique Reparameterization

A model

$$Y_i \sim f_N(y_i|\mu_i, \sigma^2)$$

Example 2: Non-unique Reparameterization

A model

$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \end{aligned}$$

Example 2: Non-unique Reparameterization

A model

$$Y_i \sim f_N(y_i|\mu_i, \sigma^2)$$
$$\mu_i = x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \qquad \text{where } x_{2i} = x_{3i}$$

Example 2: Non-unique Reparameterization

A model

$$
\begin{aligned}
Y_i &\sim f_N(y_i|\mu_i, \sigma^2) \\
\mu_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \qquad \text{where } x_{2i} = x_{3i} \\
&= x_{1i}\beta_1 + x_{2i}(\beta_2 + \beta_3)
\end{aligned}
$$

Example 2: Non-unique Reparameterization

A model

$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \qquad \text{where } x_{2i} = x_{3i} \\ &= x_{1i}\beta_1 + x_{2i}(\beta_2 + \beta_3) \end{aligned}$$

What is the (unique) MLE of β_2 and β_3? Different parameter values lead to the same values of μ and thus the same likelihood values:

Example 2: Non-unique Reparameterization

A model

$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \qquad \text{where } x_{2i} = x_{3i} \\ &= x_{1i}\beta_1 + x_{2i}(\beta_2 + \beta_3) \end{aligned}$$

What is the (unique) MLE of β_2 and β_3? Different parameter values lead to the same values of μ and thus the same likelihood values:

$$\mu_i = x_{1i}\beta_1 + x_{2i}(5 + 3)$$

Example 2: Non-unique Reparameterization

A model

$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \qquad \text{where } x_{2i} = x_{3i} \\ &= x_{1i}\beta_1 + x_{2i}(\beta_2 + \beta_3) \end{aligned}$$

What is the (unique) MLE of β_2 and β_3? Different parameter values lead to the same values of μ and thus the same likelihood values:

$$\mu_i = x_{1i}\beta_1 + x_{2i}(5 + 3)$$
$$\mu_i = x_{1i}\beta_1 + x_{2i}(3 + 5)$$

Example 2: Non-unique Reparameterization

A model

$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \qquad \text{where } x_{2i} = x_{3i} \\ &= x_{1i}\beta_1 + x_{2i}(\beta_2 + \beta_3) \end{aligned}$$

What is the (unique) MLE of β_2 and β_3? Different parameter values lead to the same values of μ and thus the same likelihood values:

$$\mu_i = x_{1i}\beta_1 + x_{2i}(5+3)$$
$$\mu_i = x_{1i}\beta_1 + x_{2i}(3+5)$$
$$\mu_i = x_{1i}\beta_1 + x_{2i}(7+1)$$

Example 2: Non-unique Reparameterization

A model

$$\begin{aligned} Y_i &\sim f_N(y_i|\mu_i, \sigma^2) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3, \qquad \text{where } x_{2i} = x_{3i} \\ &= x_{1i}\beta_1 + x_{2i}(\beta_2 + \beta_3) \end{aligned}$$

What is the (unique) MLE of β_2 and β_3? Different parameter values lead to the same values of μ and thus the same likelihood values:

$$\begin{aligned} \mu_i &= x_{1i}\beta_1 + x_{2i}(5+3) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}(3+5) \\ \mu_i &= x_{1i}\beta_1 + x_{2i}(7+1) \end{aligned}$$

So $\{\beta_2 = 2, \beta_3 = 5\}$ gives the same likelihood as $\{\beta_2 = 5, \beta_3 = 2\}$.

Introduction to Multiple Equation Models

Introduction to Multiple Equation Models

1. Let Y_i be an $N \times 1$ vector for each i ($i = 1, \ldots, n$)

Introduction to Multiple Equation Models

1. Let Y_i be an $N \times 1$ vector for each i ($i = 1, \ldots, n$)
2. Elements of Y_i are jointly distributed

Introduction to Multiple Equation Models

1. Let Y_i be an $N \times 1$ vector for each i $(i = 1, \ldots, n)$
2. Elements of Y_i are jointly distributed

$$\underset{N\times 1}{Y_i} \sim f(\underset{N\times 1}{\theta_i}, \underset{N\times N}{\alpha})$$

Introduction to Multiple Equation Models

1. Let Y_i be an $N \times 1$ vector for each i $(i = 1, \ldots, n)$
2. Elements of Y_i are jointly distributed

$$\underset{N\times 1}{Y_i} \sim f(\underset{N\times 1}{\theta_i}, \underset{N\times N}{\alpha})$$

3. Systematic components:

Introduction to Multiple Equation Models

1. Let Y_i be an $N \times 1$ vector for each i ($i = 1, \ldots, n$)
2. Elements of Y_i are jointly distributed

$$\underset{N\times 1}{Y_i} \sim f(\underset{N\times 1}{\theta_i}, \underset{N\times N}{\alpha})$$

3. Systematic components:

$$\theta_{1i} = g_1(x_{1i}, \beta_1)$$

Introduction to Multiple Equation Models

1. Let Y_i be an $N \times 1$ vector for each i ($i = 1, \ldots, n$)
2. Elements of Y_i are jointly distributed

$$\underset{N\times 1}{Y_i} \sim f(\underset{N\times 1}{\theta_i}, \underset{N\times N}{\alpha})$$

3. Systematic components:

$$\theta_{1i} = g_1(x_{1i}, \beta_1)$$
$$\theta_{2i} = g_2(x_{2i}, \beta_2)$$

Introduction to Multiple Equation Models

1. Let Y_i be an $N \times 1$ vector for each i ($i = 1, \ldots, n$)
2. Elements of Y_i are jointly distributed

$$\underset{N\times 1}{Y_i} \sim f(\underset{N\times 1}{\theta_i}, \underset{N\times N}{\alpha})$$

3. Systematic components:

$$\begin{aligned}
\theta_{1i} &= g_1(x_{1i}, \beta_1) \\
\theta_{2i} &= g_2(x_{2i}, \beta_2) \\
&\vdots \\
\theta_{Ni} &= g_N(x_{Ni}, \beta_N)
\end{aligned}$$

When are Multiple Equation Models different from N separate equation-by-equation models?

When the elements of Y_i are (conditional on X),

When are Multiple Equation Models different from N separate equation-by-equation models?

When the elements of Y_i are (conditional on X),

1. Stochastically dependent, or

When are Multiple Equation Models different from N separate equation-by-equation models?

When the elements of Y_i are (conditional on X),

1. Stochastically dependent, or
2. Parametrically dependent (shared parameters)

When are Multiple Equation Models different from N separate equation-by-equation models?

When the elements of Y_i are (conditional on X),

1. Stochastically dependent, or
2. Parametrically dependent (shared parameters)

Example and proof:

When are Multiple Equation Models different from N separate equation-by-equation models?

When the elements of Y_i are (conditional on X),

1. Stochastically dependent, or
2. Parametrically dependent (shared parameters)

Example and proof:

Suppose no ancillary parameters, and $N = 2$. The joint density:

When are Multiple Equation Models different from N separate equation-by-equation models?

When the elements of Y_i are (conditional on X),

1. Stochastically dependent, or
2. Parametrically dependent (shared parameters)

Example and proof:

Suppose no ancillary parameters, and $N = 2$. The joint density:

$$f(y|\theta) = \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i})$$

When are Multiple Equation Models different from N separate equation-by-equation models?

When the elements of Y_i are (conditional on X),

1. Stochastically dependent, or
2. Parametrically dependent (shared parameters)

Example and proof:

Suppose no ancillary parameters, and $N = 2$. The joint density:

$$f(y|\theta) = \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i})$$

(BTW, you now know how to form the likelihood for multiple equation models!)

Assuming stochastic independence lets us factor f:

Assuming stochastic independence lets us factor f:

$$P(y|\theta) = \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i})$$

Assuming stochastic independence lets us factor f:

$$\begin{aligned} P(y|\theta) &= \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i}) \\ &= \prod_{i=1}^{n} f(y_{1i}|\theta_{1i}) f(y_{2i}|\theta_{2i}) \end{aligned}$$

Assuming stochastic independence lets us factor f:

$$\begin{aligned} P(y|\theta) &= \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i}) \\ &= \prod_{i=1}^{n} f(y_{1i}|\theta_{1i}) f(y_{2i}|\theta_{2i}) \end{aligned}$$

with log-likelihood

Assuming stochastic independence lets us factor f:

$$\begin{aligned} P(y|\theta) &= \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i}) \\ &= \prod_{i=1}^{n} f(y_{1i}|\theta_{1i}) f(y_{2i}|\theta_{2i}) \end{aligned}$$

with log-likelihood

$$\ln L(\theta_1, \theta_2|y) = \sum_{i=1}^{n} \ln f(y_{1i}|\theta_{1i}) + \sum_{i=1}^{n} \ln f(y_{2i}|\theta_{2i})$$

Assuming stochastic independence lets us factor f:

$$\begin{aligned} P(y|\theta) &= \prod_{i=1}^{n} f(y_{1i}, y_{2i}|\theta_{1i}, \theta_{2i}) \\ &= \prod_{i=1}^{n} f(y_{1i}|\theta_{1i}) f(y_{2i}|\theta_{2i}) \end{aligned}$$

with log-likelihood

$$\ln L(\theta_1, \theta_2|y) = \sum_{i=1}^{n} \ln f(y_{1i}|\theta_{1i}) + \sum_{i=1}^{n} \ln f(y_{2i}|\theta_{2i})$$

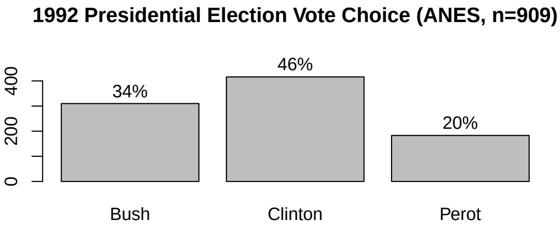
Also assume parametric independence, and you can estimate the equations separately.

Example: 1992 U.S. Presidential Election

Example: 1992 U.S. Presidential Election

Alvarez and Nagler (1995):

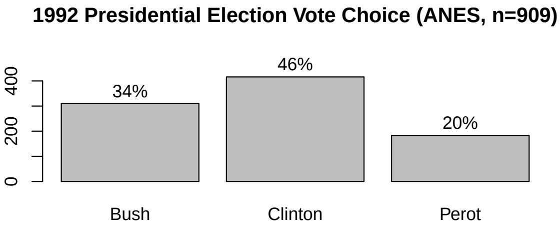
- Y_i: Vote choice in the 1992 U.S. presidential election (1 = Clinton, 2 = Bush, 3 = Perot)



Example: 1992 U.S. Presidential Election

Alvarez and Nagler (1995):

- Y_i: Vote choice in the 1992 U.S. presidential election ($1 =$ Clinton, $2 =$ Bush, $3 =$ Perot)



- Two types of predictors:
  - Voter-specific (V_i): age, gender, education, party, opinions, etc.
  - Candidate-varying (X_{ij}): ideological distance between voter i and candidate j

Multinomial Logit Model

- Generalize the logit model to more than two choices

Multinomial Logit Model

- Generalize the logit model to more than two choices
- The multinomial logit model (MNL):

$$\pi_{ij} \; = \; \Pr(Y_i = j \mid V_i) \; = \; \frac{\exp(V_i^\top \delta_j)}{\sum_{k=1}^J \exp(V_i^\top \delta_k)},$$

where V_i = individual-specific characteristics of unit i (and an intercept)

Multinomial Logit Model

- Generalize the logit model to more than two choices
- The multinomial logit model (MNL):

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid V_i) \;=\; \frac{\exp(V_i^\top \delta_j)}{\sum_{k=1}^{J} \exp(V_i^\top \delta_k)},$$

  where V_i = individual-specific characteristics of unit i (and an intercept)
- Note that $\sum_{j=1}^{J} \pi_{ij} = 1$

Multinomial Logit Model

- Generalize the logit model to more than two choices
- The multinomial logit model (MNL):

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid V_i) \;=\; \frac{\exp(V_i^\top \delta_j)}{\sum_{k=1}^{J} \exp(V_i^\top \delta_k)},$$

  where V_i = individual-specific characteristics of unit i (and an intercept)
- Note that $\sum_{j=1}^{J} \pi_{ij} = 1$
- Need to set the base category for identifiability: $\delta_1 = 0$

Multinomial Logit Model

- Generalize the logit model to more than two choices
- The multinomial logit model (MNL):

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid V_i) \;=\; \frac{\exp(V_i^\top \delta_j)}{\sum_{k=1}^{J} \exp(V_i^\top \delta_k)},$$

  where V_i = individual-specific characteristics of unit i (and an intercept)
- Note that $\sum_{j=1}^{J} \pi_{ij} = 1$
- Need to set the base category for identifiability: $\delta_1 = 0$
- δ_j represents how characteristics of voter i is associated with probability of voting for candidate j

Conditional Logit Model

- We can also incorporate alternative-varying predictors X_{ij}

Conditional Logit Model

- We can also incorporate alternative-varying predictors X_{ij}
- The conditional logit (CL) model:

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid X_{ij}) \;=\; \frac{\exp(X_{ij}^\top \beta)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta)}$$

Conditional Logit Model

- We can also incorporate alternative-varying predictors X_{ij}
- The conditional logit (CL) model:

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid X_{ij}) \;=\; \frac{\exp(X_{ij}^\top \beta)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta)}$$

- β represents how characteristics of candidate j for voter i are associated with voting probabilities

Conditional Logit Model

- We can also incorporate alternative-varying predictors X_{ij}
- The conditional logit (CL) model:

$$\pi_{ij} \; = \; \Pr(Y_i = j \mid X_{ij}) \; = \; \frac{\exp(X_{ij}^\top \beta)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta)}$$

- β represents how characteristics of candidate j for voter i are associated with voting probabilities

- X_{ij} does not have to vary across voters (e.g. whether candidate j is incumbent)

Conditional Logit Model

- We can also incorporate alternative-varying predictors X_{ij}
- The conditional logit (CL) model:

$$\pi_{ij} \;=\; \Pr(Y_i = j \mid X_{ij}) \;=\; \frac{\exp(X_{ij}^\top \beta)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta)}$$

- β represents how characteristics of candidate j for voter i are associated with voting probabilities
- X_{ij} does not have to vary across voters (e.g. whether candidate j is incumbent)
- In that case we suppress the subscript to X_j

MNL as a Special Case of CL

MNL as a Special Case of CL

- Mathematically, MNL can be subsumed under CL using a set of artificial alternative-varying regressors for each V_i:

$$X_{i1} = \begin{pmatrix} V_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad X_{i2} = \begin{pmatrix} 0 \\ V_i \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots, \quad X_{iJ} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ V_i \end{pmatrix}$$

MNL as a Special Case of CL

- Mathematically, MNL can be subsumed under CL using a set of artificial alternative-varying regressors for each V_i:

$$X_{i1} = \begin{pmatrix} V_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad X_{i2} = \begin{pmatrix} 0 \\ V_i \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots, \quad X_{iJ} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ V_i \end{pmatrix}$$

- Set the element of β for X_{ij} to δ_j and you get the MNL model

MNL as a Special Case of CL

- Mathematically, MNL can be subsumed under CL using a set of artificial alternative-varying regressors for each V_i:

$$X_{i1} = \begin{pmatrix} V_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad X_{i2} = \begin{pmatrix} 0 \\ V_i \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots, \quad X_{iJ} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ V_i \end{pmatrix}$$

- Set the element of β for X_{ij} to δ_j and you get the MNL model
- δ_1 must be set to zero for identifiability

MNL as a Special Case of CL

- Mathematically, MNL can be subsumed under CL using a set of artificial alternative-varying regressors for each V_i:

$$X_{i1} = \begin{pmatrix} V_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad X_{i2} = \begin{pmatrix} 0 \\ V_i \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots, \quad X_{iJ} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ V_i \end{pmatrix}$$

- Set the element of β for X_{ij} to δ_j and you get the MNL model
- δ_1 must be set to zero for identifiability

- Thus we can write both models (and their mixture) simply as CL:

$$\pi_{ij} = \Pr(Y_i = j \mid X) = \frac{\exp(X_{ij}^\top \beta)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta)}$$

MNL as a Special Case of CL

- Mathematically, MNL can be subsumed under CL using a set of artificial alternative-varying regressors for each V_i:

$$X_{i1} = \begin{pmatrix} V_i \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad X_{i2} = \begin{pmatrix} 0 \\ V_i \\ \vdots \\ 0 \end{pmatrix}, \quad \cdots, \quad X_{iJ} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ V_i \end{pmatrix}$$

- Set the element of β for X_{ij} to δ_j and you get the MNL model
- δ_1 must be set to zero for identifiability
- Thus we can write both models (and their mixture) simply as CL:

$$\pi_{ij} = \Pr(Y_i = j \mid X) = \frac{\exp(X_{ij}^\top \beta)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta)}$$

- We use the names CL and MNL interchangeably from here on

Predictor Types and Data Formats

Predictor Types and Data Formats

Discrete choice data usually come in one of the two formats:

Predictor Types and Data Formats

Discrete choice data usually come in one of the two formats:

1. Wide format: N rows, $\#V + J \cdot \#X$ predictors

Predictor Types and Data Formats

Discrete choice data usually come in one of the two formats:

1. Wide format: N rows, $\#V + J \cdot \#X$ predictors

choice	women	educ	idist.Clinton	idist.Bush	idist.Perot	
Bush	1	3	4.0804	0.1024	0.2601	
Bush	1	4	4.0804	0.1024	0.2601	
Clinton	1	2	1.0404	1.7424	0.2401	
Bush	0	6	0.0004	5.3824	2.2201	
Clinton	1	3	0.9604	11.0220	6.2001	...

Predictor Types and Data Formats

Discrete choice data usually come in one of the two formats:

1. Wide format: N rows, $\#V + J \cdot \#X$ predictors

```
 choice women educ idist.Clinton idist.Bush idist.Perot
   Bush     1    3        4.0804     0.1024      0.2601
   Bush     1    4        4.0804     0.1024      0.2601
Clinton     1    2        1.0404     1.7424      0.2401
   Bush     0    6        0.0004     5.3824      2.2201
Clinton     1    3        0.9604    11.0220      6.2001 ...
```

2. Long format: NJ rows, $\#V + \#X$ predictors

Predictor Types and Data Formats

Discrete choice data usually come in one of the two formats:

1. Wide format: N rows, $\#V + J \cdot \#X$ predictors

choice	women	educ	idist.Clinton	idist.Bush	idist.Perot	
Bush	1	3	4.0804	0.1024	0.2601	
Bush	1	4	4.0804	0.1024	0.2601	
Clinton	1	2	1.0404	1.7424	0.2401	
Bush	0	6	0.0004	5.3824	2.2201	
Clinton	1	3	0.9604	11.0220	6.2001	...

2. Long format: NJ rows, $\#V + \#X$ predictors

chid	alt	choice	women	educ	idist	
1	Bush	TRUE	1	3	0.1024	
1	Clinton	FALSE	1	3	4.0804	
1	Perot	FALSE	1	3	0.2601	
2	Bush	TRUE	1	4	0.1024	
2	Clinton	FALSE	1	4	4.0804	
2	Perot	FALSE	1	4	0.2601	...

Predictor Types and Data Formats

Discrete choice data usually come in one of the two formats:

1. Wide format: N rows, $\#V + J \cdot \#X$ predictors

choice	women	educ	idist.Clinton	idist.Bush	idist.Perot	
Bush	1	3	4.0804	0.1024	0.2601	
Bush	1	4	4.0804	0.1024	0.2601	
Clinton	1	2	1.0404	1.7424	0.2401	
Bush	0	6	0.0004	5.3824	2.2201	
Clinton	1	3	0.9604	11.0220	6.2001	...

2. Long format: NJ rows, $\#V + \#X$ predictors

chid	alt	choice	women	educ	idist	
1	Bush	TRUE	1	3	0.1024	
1	Clinton	FALSE	1	3	4.0804	
1	Perot	FALSE	1	3	0.2601	
2	Bush	TRUE	1	4	0.1024	
2	Clinton	FALSE	1	4	4.0804	
2	Perot	FALSE	1	4	0.2601	...

- Use `reshape` to change between wide and long

Predictor Types and Data Formats

Discrete choice data usually come in one of the two formats:

1. Wide format: N rows, $\#V + J \cdot \#X$ predictors

choice	women	educ	idist.Clinton	idist.Bush	idist.Perot	
Bush	1	3	4.0804	0.1024	0.2601	
Bush	1	4	4.0804	0.1024	0.2601	
Clinton	1	2	1.0404	1.7424	0.2401	
Bush	0	6	0.0004	5.3824	2.2201	
Clinton	1	3	0.9604	11.0220	6.2001	...

2. Long format: NJ rows, $\#V + \#X$ predictors

chid	alt	choice	women	educ	idist	
1	Bush	TRUE	1	3	0.1024	
1	Clinton	FALSE	1	3	4.0804	
1	Perot	FALSE	1	3	0.2601	
2	Bush	TRUE	1	4	0.1024	
2	Clinton	FALSE	1	4	4.0804	
2	Perot	FALSE	1	4	0.2601	...

- Use `reshape` to change between wide and long
- Some estimation functions (e.g. `mlogit`) can take both formats

Latent Variable Interpretation

- Recall the random utility model:

$$Y_{ij}^* \;=\; X_{ij}^\top \beta + \epsilon_{ij},$$

where $\begin{cases} Y_{ij}^* & = & \text{latent utility from choosing } j \text{ for } i \\ \epsilon_{ij} & = & \text{stochastic component of the utility} \end{cases}$

Latent Variable Interpretation

- Recall the random utility model:

$$Y_{ij}^* = X_{ij}^\top \beta + \epsilon_{ij},$$

where $\begin{cases} Y_{ij}^* & = & \text{latent utility from choosing } j \text{ for } i \\ \epsilon_{ij} & = & \text{stochastic component of the utility} \end{cases}$

- Assume that voter chooses the most preferred candidate, i.e.,

$$Y_i = j \quad \text{if} \quad Y_{ij}^* \geq Y_{ij'}^* \quad \text{for any} \quad j' \in \{1, ..., J\}$$

Latent Variable Interpretation

- Recall the random utility model:

$$Y_{ij}^* \ = \ X_{ij}^\top \beta + \epsilon_{ij},$$

where $\begin{cases} Y_{ij}^* & = & \text{latent utility from choosing } j \text{ for } i \\ \epsilon_{ij} & = & \text{stochastic component of the utility} \end{cases}$

- Assume that voter chooses the most preferred candidate, i.e.,

$$Y_i = j \quad \text{if} \quad Y_{ij}^* \geq Y_{ij'}^* \quad \text{for any} \quad j' \in \{1, ..., J\}$$

- Assuming $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim}$ type I extreme value distribution, this setup implies MNL (McFadden 1974)

Latent Variable Interpretation

- Recall the random utility model:

$$Y_{ij}^* \;=\; X_{ij}^\top \beta + \epsilon_{ij},$$

where $\begin{cases} Y_{ij}^* & = & \text{latent utility from choosing } j \text{ for } i \\ \epsilon_{ij} & = & \text{stochastic component of the utility} \end{cases}$

- Assume that voter chooses the most preferred candidate, i.e.,

$$Y_i = j \quad \text{if} \quad Y_{ij}^* \geq Y_{ij'}^* \quad \text{for any} \quad j' \in \{1, ..., J\}$$

- Assuming $\epsilon_{ij} \overset{\text{i.i.d.}}{\sim}$ type I extreme value distribution, this setup implies MNL (McFadden 1974)

- Proof for $J = 2$:

$$\begin{aligned} \Pr(Y_i = 1 \mid X) &= \Pr(Y_{i1}^* \geq Y_{i2}^* \mid X) \\ &= \Pr\left(\epsilon_{i2} - \epsilon_{i1} \leq (X_{i1} - X_{i2})^\top \beta\right) \\ &= \frac{\exp\left((X_{i1} - X_{i2})^\top \beta\right)}{1 + \exp\left((X_{i1} - X_{i2})^\top \beta\right)} = \frac{\exp(X_{i1}^\top \beta)}{\exp(X_{i1}^\top \beta) + \exp(X_{i2}^\top \beta)} \end{aligned}$$

Estimation and Inference

- Estimation via MLE

Estimation and Inference

- Estimation via MLE
- Likelihood for a random sample of size n:

$$L(\beta \mid Y, X) \;=\; \prod_{i=1}^{n} \prod_{j=1}^{J} \pi_{ij}^{1\{Y_i = j\}}$$

Estimation and Inference

- Estimation via MLE
- Likelihood for a random sample of size n:

$$L(\beta \mid Y, X) \; = \; \prod_{i=1}^{n} \prod_{j=1}^{J} \pi_{ij}^{1\{Y_i = j\}}$$

- It can be shown that the log-likelihood is globally concave
  $\Rightarrow$ guaranteed convergence to the true (not local) MLE

Interpreting MNL/CL Coefficients

In MNL/CL, β itself is not necessarily informative about the effect of X

Interpreting MNL/CL Coefficients

In MNL/CL, β itself is not necessarily informative about the effect of X

1. The coefficients are all with respect to the baseline category
   $\longrightarrow$ Testing $\beta_j = 0$ does not generally make sense
   (unless comparison to the baseline is the goal)

Interpreting MNL/CL Coefficients

In MNL/CL, β itself is not necessarily informative about the effect of X

1. The coefficients are all with respect to the baseline category
   $\longrightarrow$ Testing $\beta_j = 0$ does not generally make sense
   (unless comparison to the baseline is the goal)
2. Changing X_{ij} has impact on $\Pr(Y_i = k \mid X)$, $k \neq j$:

Interpreting MNL/CL Coefficients

In MNL/CL, β itself is not necessarily informative about the effect of X

1. The coefficients are all with respect to the baseline category
   $\longrightarrow$ Testing $\beta_j = 0$ does not generally make sense
   (unless comparison to the baseline is the goal)
2. Changing X_{ij} has impact on $\Pr(Y_i = k \mid X)$, $k \neq j$:
   - For individual-specific characteristics (V_i), even sign of δ_j may not agree with the direction of the change in response probability for j

Interpreting MNL/CL Coefficients

In MNL/CL, β itself is not necessarily informative about the effect of X

1. The coefficients are all with respect to the baseline category
   $\longrightarrow$ Testing $\beta_j = 0$ does not generally make sense
   (unless comparison to the baseline is the goal)

2. Changing X_{ij} has impact on $\Pr(Y_i = k \mid X)$, $k \neq j$:
   - For individual-specific characteristics (V_i), even sign of δ_j may not agree with the direction of the change in response probability for j
   - For alternative-varying characteristics (X_{ij}), sign of β does indicate the direction of the effect, but magnitude is hard to interpret

Interpreting MNL/CL Coefficients

In MNL/CL, β itself is not necessarily informative about the effect of X

1. The coefficients are all with respect to the baseline category
   $\longrightarrow$ Testing $\beta_j = 0$ does not generally make sense
   (unless comparison to the baseline is the goal)
2. Changing X_{ij} has impact on $\Pr(Y_i = k \mid X)$, $k \neq j$:
   - For individual-specific characteristics (V_i), even sign of δ_j may not agree with the direction of the change in response probability for j
   - For alternative-varying characteristics (X_{ij}), sign of β does indicate the direction of the effect, but magnitude is hard to interpret

Compute a quantity that has a clear substantive interpretation!

Calculating Quantities of Interest

1. Choice probability:

$$\pi_j(x) \;=\; \Pr(Y_i = j \mid X = x)$$

e.g. How likely is a female college-educated conservative Republican voter to vote for Perot?

Calculating Quantities of Interest

1. Choice probability:

$$\pi_j(x) \ = \ \Pr(Y_i = j \mid X = x)$$

   e.g. How likely is a female college-educated conservative Republican voter to vote for Perot?

2. Predicted vote share:

$$p_j(x_1) \ \equiv \ \mathbb{E}\left[1\left\{\pi_j(X_{i1} = x_1, X_{i2}) \geq \pi_k(X_{i1} = x_1, X_{i2}) \text{ for all } k\right\}\right]$$

   where X_{i1} is the predictor(s) of interest and X_{i2} is all other predictors
   e.g. What would Perot's vote share be if all voters supported abortion?

Calculating Quantities of Interest

1. Choice probability:

$$\pi_j(x) \ = \ \Pr(Y_i = j \mid X = x)$$

e.g. How likely is a female college-educated conservative Republican voter to vote for Perot?

2. Predicted vote share:

$$p_j(x_1) \ \equiv \ \mathbb{E}\left[1\left\{\pi_j(X_{i1} = x_1, X_{i2}) \geq \pi_k(X_{i1} = x_1, X_{i2}) \text{ for all } k\right\}\right]$$

where X_{i1} is the predictor(s) of interest and X_{i2} is all other predictors
e.g. What would Perot's vote share be if all voters supported abortion?

3. Average partial (treatment) effects:

$$\tau_{jk} \ = \ \mathbb{E}\left[\pi_j(T_{ik} = 1, T_{i*}, W_i) - \pi_j(T_{ik} = 0, T_{i*}, W_i)\right]$$

where T_{ik} is treatment on candidate k, T_{i*} is treatment on others, W_i is pre-treatment covariates

- "Direct effect" if $j = k$; "indirect effect" if $j \neq k$
- If T is individual-specific, $\tau_j = \mathbb{E}[\pi_j(T_i = 1, W_i) - \pi_j(T_i = 0, W_i)]$

Calculating Quantities of Interest

1. Choice probability:

$$\pi_j(x) \;=\; \Pr(Y_i = j \mid X = x)$$

   e.g. How likely is a female college-educated conservative Republican voter to vote for Perot?

2. Predicted vote share:

$$p_j(x_1) \;\equiv\; \mathbb{E}\left[1\left\{\pi_j(X_{i1} = x_1, X_{i2}) \geq \pi_k(X_{i1} = x_1, X_{i2}) \text{ for all } k\right\}\right]$$

   where X_{i1} is the predictor(s) of interest and X_{i2} is all other predictors
   e.g. What would Perot's vote share be if all voters supported abortion?

3. Average partial (treatment) effects:

$$\tau_{jk} \;=\; \mathbb{E}\left[\pi_j(T_{ik} = 1, T_{i*}, W_i) - \pi_j(T_{ik} = 0, T_{i*}, W_i)\right]$$

   where T_{ik} is treatment on candidate k, T_{i*} is treatment on others, W_i is pre-treatment covariates

   - "Direct effect" if $j = k$; "indirect effect" if $j \neq k$
   - If T is individual-specific, $\tau_j = \mathbb{E}[\pi_j(T_i = 1, W_i) - \pi_j(T_i = 0, W_i)]$

- Estimate by plugging in sample analogues (e.g. $\pi_j \rightarrow \hat{\pi}_j$, $\mathbb{E} \rightarrow \frac{1}{n}\sum$)

Example: 1992 U.S. Presidential Election

- Model specification (Alvarez and Nagler 1995):

$$\pi_{ij} = \frac{\exp(X_{ij}^{\top}\beta + V_i^{\top}\delta_j)}{\sum_{k=1}^{J}\exp(X_{ik}^{\top}\beta + V_i^{\top}\delta_k)}$$

where

$$\begin{aligned} X_{ij} &= \{\text{ideological distance}\} \\ V_i &= \{1, \text{issue opinions, party, gender, education, age, ...}\} \end{aligned}$$

Example: 1992 U.S. Presidential Election

- Model specification (Alvarez and Nagler 1995):

$$\pi_{ij} = \frac{\exp(X_{ij}^\top \beta + V_i^\top \delta_j)}{\sum_{k=1}^{J} \exp(X_{ik}^\top \beta + V_i^\top \delta_k)}$$

where

$$\begin{aligned} X_{ij} &= \{\text{ideological distance}\} \\ V_i &= \{1, \text{issue opinions, party, gender, education, age, ...}\} \end{aligned}$$

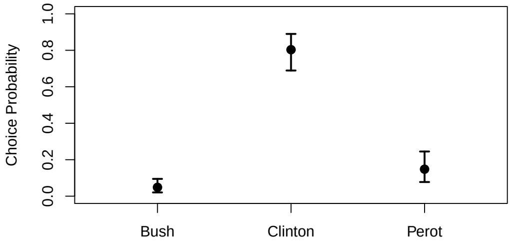
- Estimated coefficients:

$$\hat{\beta} = -0.11 \ (0.02)$$

$$\hat{\delta} = \begin{bmatrix} \hat{\delta}_{\text{Bush}} & \hat{\delta}_{\text{Clinton}} \end{bmatrix} = \begin{array}{c} \begin{bmatrix} 0.67\ (0.94) & -0.41\ (0.45) \\ -0.52\ (0.11) & -0.02\ (0.12) \\ 0.54\ (0.23) & 0.30\ (0.22) \\ \vdots & \vdots \end{bmatrix} \end{array} \begin{array}{l} \text{(intercept)} \\ \text{(support abortion)} \\ \text{(female)} \\ \vdots \end{array}$$

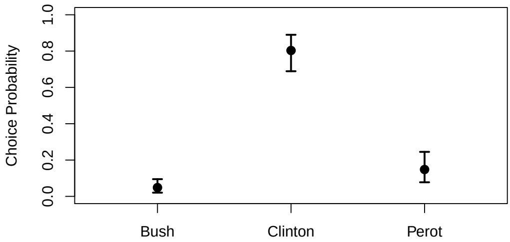
Example: 1992 U.S. Presidential Election

- Estimated choice probabilities for a “typical” voter from South:

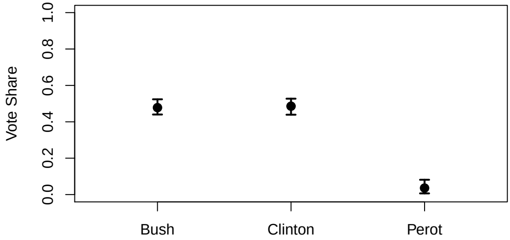


Example: 1992 U.S. Presidential Election

- Estimated choice probabilities for a "typical" voter from South:



- Predicted vote shares if everyone opposed abortion:



Assumptions in Multinomial Logit Models

- Recall that MNL assumes ϵ_{ij} is i.i.d.

Assumptions in Multinomial Logit Models

- Recall that MNL assumes ϵ_{ij} is i.i.d.
- In particular, $\epsilon_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $j \neq k$

Assumptions in Multinomial Logit Models

- Recall that MNL assumes ϵ_{ij} is i.i.d.
- In particular, $\epsilon_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $j \neq k$
- This implies that unobserved factors affecting Y^*_{ij} are unrelated to those affecting Y^*_{ik}

Assumptions in Multinomial Logit Models

- Recall that MNL assumes ϵ_{ij} is i.i.d.
- In particular, $\epsilon_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $j \neq k$
- This implies that unobserved factors affecting Y^*_{ij} are unrelated to those affecting Y^*_{ik}
- When is this assumption plausible?

Assumptions in Multinomial Logit Models

- Recall that MNL assumes ϵ_{ij} is i.i.d.
- In particular, $\epsilon_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $j \neq k$
- This implies that unobserved factors affecting Y_{ij}^* are unrelated to those affecting Y_{ik}^*
- When is this assumption plausible?
- Example: Multiparty election with parties R, L1 and L2.

Assumptions in Multinomial Logit Models

- Recall that MNL assumes ϵ_{ij} is i.i.d.
- In particular, $\epsilon_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $j \neq k$
- This implies that unobserved factors affecting Y^*_{ij} are unrelated to those affecting Y^*_{ik}
- When is this assumption plausible?
- Example: Multiparty election with parties R, L1 and L2.
- Do voters' unobserved ideological preferences affect $\Pr(Y_i = \text{L1})$ independently of their effect on $\Pr(Y_i = \text{L2})$?

Assumptions in Multinomial Logit Models

- Recall that MNL assumes ϵ_{ij} is i.i.d.
- In particular, $\epsilon_{ij} \perp\!\!\!\perp \epsilon_{ik}$ for $j \neq k$
- This implies that unobserved factors affecting Y^*_{ij} are unrelated to those affecting Y^*_{ik}
- When is this assumption plausible?
- Example: Multiparty election with parties R, L1 and L2.
- Do voters' unobserved ideological preferences affect $\Pr(Y_i = \text{L1})$ independently of their effect on $\Pr(Y_i = \text{L2})$? Probably not.

Independence of Irrelevant Alternatives

- MNL makes the Independence of irrelevant alternatives (IIA) assumption:

$$\frac{\Pr(\text{Choose } j \mid j \text{ or } k)}{\Pr(\text{Choose } k \mid j \text{ or } k)} = \frac{\Pr(\text{Choose } j \mid j \text{ or } k \text{ or } l)}{\Pr(\text{Choose } k \mid j \text{ or } k \text{ or } l)} \quad \text{for any } l \in \{1, \ldots J\}$$

Independence of Irrelevant Alternatives

- MNL makes the Independence of irrelevant alternatives (IIA) assumption:

$$\frac{\Pr(\text{Choose } j \mid j \text{ or } k)}{\Pr(\text{Choose } k \mid j \text{ or } k)} = \frac{\Pr(\text{Choose } j \mid j \text{ or } k \text{ or } l)}{\Pr(\text{Choose } k \mid j \text{ or } k \text{ or } l)} \quad \text{for any } l \in \{1, \ldots J\}$$

- A classical example of IIA violation: the red bus-blue bus problem

Independence of Irrelevant Alternatives

- MNL makes the Independence of irrelevant alternatives (IIA) assumption:

$$\frac{\Pr(\text{Choose } j \mid j \text{ or } k)}{\Pr(\text{Choose } k \mid j \text{ or } k)} = \frac{\Pr(\text{Choose } j \mid j \text{ or } k \text{ or } l)}{\Pr(\text{Choose } k \mid j \text{ or } k \text{ or } l)} \quad \text{for any } l \in \{1, \ldots J\}$$

- A classical example of IIA violation: the red bus-blue bus problem

- Relative risk of j over k does not depend on other alternatives:

$$\frac{\Pr(Y_i = j \mid X_i)}{\Pr(Y_i = k \mid X_i)} = \exp\{(X_{ij} - X_{ik})^\top \beta\}$$

Independence of Irrelevant Alternatives

- MNL makes the Independence of irrelevant alternatives (IIA) assumption:

$$\frac{\Pr(\text{Choose } j \mid j \text{ or } k)}{\Pr(\text{Choose } k \mid j \text{ or } k)} = \frac{\Pr(\text{Choose } j \mid j \text{ or } k \text{ or } l)}{\Pr(\text{Choose } k \mid j \text{ or } k \text{ or } l)} \quad \text{for any } l \in \{1, ...J\}$$

- A classical example of IIA violation: the red bus-blue bus problem

- Relative risk of j over k does not depend on other alternatives:

$$\frac{\Pr(Y_i = j \mid X_i)}{\Pr(Y_i = k \mid X_i)} = \exp\{(X_{ij} - X_{ik})^\top \beta\}$$

- That is, the multinomial choice reduces to a series of independent pairwise comparisons

Multinomial Probit Model

- How can we relax the IIA assumption?

Multinomial Probit Model

- How can we relax the IIA assumption?
- Instead of assuming ϵ_{ij} to be i.i.d. across alternatives j, we allow ϵ_{ij} to be correlated across j within each voter i

Multinomial Probit Model

- How can we relax the IIA assumption?
- Instead of assuming ϵ_{ij} to be i.i.d. across alternatives j, we allow ϵ_{ij} to be correlated across j within each voter i
- Multinomial probit model (MNP):

$$Y_i^* = X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \ \cdots \ Y_{iJ}^*]^\top \\ X_i = [X_{i1} \ \cdots \ X_{iJ}]^\top \end{cases}$$

Multinomial Probit Model

- How can we relax the IIA assumption?
- Instead of assuming ϵ_{ij} to be i.i.d. across alternatives j, we allow ϵ_{ij} to be correlated across j within each voter i
- Multinomial probit model (MNP):

$$Y_i^* \; = \; X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \; \cdots \; Y_{iJ}^*]^\top \\ X_i = [X_{i1} \; \cdots \; X_{iJ}]^\top \end{cases}$$

- Restrictions on the model for identifiability:

Multinomial Probit Model

- How can we relax the IIA assumption?
- Instead of assuming ϵ_{ij} to be i.i.d. across alternatives j, we allow ϵ_{ij} to be correlated across j within each voter i
- Multinomial probit model (MNP):

$$Y_i^* \;=\; X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \;\cdots\; Y_{iJ}^*]^\top \\ X_i = [X_{i1} \;\cdots\; X_{iJ}]^\top \end{cases}$$

- Restrictions on the model for identifiability:
  - The (absolute) level of Y_i^* shouldn't matter

Multinomial Probit Model

- How can we relax the IIA assumption?
- Instead of assuming ϵ_{ij} to be i.i.d. across alternatives j, we allow ϵ_{ij} to be correlated across j within each voter i
- Multinomial probit model (MNP):

$$Y_i^* \;=\; X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \;\cdots\; Y_{iJ}^*]^\top \\ X_i = [X_{i1} \;\cdots\; X_{iJ}]^\top \end{cases}$$

- Restrictions on the model for identifiability:
  - The (absolute) level of Y_i^* shouldn't matter
    $\longrightarrow$ Subtract the 1st equation from all the other equations and work with a system of $J-1$ equations with $\tilde{\epsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tilde{\Sigma}_{J-1})$

Multinomial Probit Model

- How can we relax the IIA assumption?
- Instead of assuming ϵ_{ij} to be i.i.d. across alternatives j, we allow ϵ_{ij} to be correlated across j within each voter i
- Multinomial probit model (MNP):

$$Y_i^* \;=\; X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \;\; \cdots \;\; Y_{iJ}^*]^\top \\ X_i = [X_{i1} \;\; \cdots \;\; X_{iJ}]^\top \end{cases}$$

- Restrictions on the model for identifiability:
  - The (absolute) level of Y_i^* shouldn't matter
    $\longrightarrow$ Subtract the 1st equation from all the other equations and work with a system of $J-1$ equations with $\tilde{\epsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tilde{\Sigma}_{J-1})$
  - The scale of Y_i^* also shouldn't matter

Multinomial Probit Model

- How can we relax the IIA assumption?
- Instead of assuming ϵ_{ij} to be i.i.d. across alternatives j, we allow ϵ_{ij} to be correlated across j within each voter i
- Multinomial probit model (MNP):

$$Y_i^* \ = \ X_i^\top \beta + \epsilon_i \quad \text{where} \quad \begin{cases} \epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Sigma_J) \\ Y_i^* = [Y_{i1}^* \ \cdots \ Y_{iJ}^*]^\top \\ X_i = [X_{i1} \ \cdots \ X_{iJ}]^\top \end{cases}$$

- Restrictions on the model for identifiability:
  - The (absolute) level of Y_i^* shouldn't matter
    $\longrightarrow$ Subtract the 1st equation from all the other equations and work with a system of $J-1$ equations with $\tilde{\epsilon}_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tilde{\Sigma}_{J-1})$
  - The scale of Y_i^* also shouldn't matter
    $\longrightarrow \tilde{\Sigma}_{(1,1)} = 1$

Multinomial Probit Walkthrough

$$U_i^* \sim N(u_i^*|\mu_i, \Sigma)$$

Multinomial Probit Walkthrough

$$U_i^* \sim N(u_i^*|\mu_i, \Sigma)$$
$$\mu_{ij} = x_{ij}\beta_j$$

Multinomial Probit Walkthrough

$$U_i^* \sim N(u_i^*|\mu_i, \Sigma)$$
$$\mu_{ij} = x_{ij}\beta_j$$

with observation mechanism:

Multinomial Probit Walkthrough

$$U_i^* \sim N(u_i^* | \mu_i, \Sigma)$$
$$\mu_{ij} = x_{ij}\beta_j$$

with observation mechanism:

$$Y_{ij} = \begin{cases} 1 & \text{if } U_{ij}^* > U_{ij'}^* \; \forall \; j \neq j' \\ 0 & \text{otherwise} \end{cases}$$

The stochastic component:

$$\text{for} \quad i = 1, \ldots, n$$

The stochastic component:

$$\Pr(Y_{ij} = 1) = \pi_{ij}, \quad \text{s.t.} \quad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for} \quad i = 1, \dots, n$$

The stochastic component:

$$\Pr(Y_{ij} = 1) = \pi_{ij}, \quad \text{s.t.} \quad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for} \quad i = 1, \ldots, n$$

Systematic component. Let $Y^*_{ij} = U^*_{ij} - U^*_{ij'}$, so the observation mechanism is

The stochastic component:

$$\Pr(Y_{ij} = 1) = \pi_{ij}, \quad \text{s.t.} \quad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for} \quad i = 1, \dots, n$$

Systematic component. Let $Y_{ij}^* = U_{ij}^* - U_{ij'}^*$, so the observation mechanism is

$$Y_{ij} = \begin{cases} 1 & \text{if } Y_{ij}^* > 0 \\ 0 & \text{otherwise} \end{cases}$$

The stochastic component:

$$\Pr(Y_{ij} = 1) = \pi_{ij}, \quad \text{s.t.} \quad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for} \quad i = 1, \ldots, n$$

Systematic component. Let $Y^*_{ij} = U^*_{ij} - U^*_{ij'}$, so the observation mechanism is

$$Y_{ij} = \begin{cases} 1 & \text{if } Y^*_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\pi_{ij} = \Pr(y_{ij} = 1)$$

The stochastic component:

$$\Pr(Y_{ij} = 1) = \pi_{ij}, \quad \text{s.t.} \quad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for} \quad i = 1, \ldots, n$$

Systematic component. Let $Y^*_{ij} = U^*_{ij} - U^*_{ij'}$, so the observation mechanism is

$$Y_{ij} = \begin{cases} 1 & \text{if } Y^*_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \pi_{ij} &= \Pr(y_{ij} = 1) \\ &= \Pr(Y^*_{i1} \leq 0, \ldots, {\color{red} Y^*_{ij} > 0}, \ldots, Y^*_{iJ} \leq 0) \end{aligned}$$

The stochastic component:

$$\Pr(Y_{ij} = 1) = \pi_{ij}, \quad \text{s.t.} \quad \sum_{j=1}^{J} \pi_{ij} = 1 \quad \text{for} \quad i = 1, \dots, n$$

Systematic component. Let $Y^*_{ij} = U^*_{ij} - U^*_{ij'}$, so the observation mechanism is

$$Y_{ij} = \begin{cases} 1 & \text{if } Y^*_{ij} > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} \pi_{ij} &= \Pr(y_{ij} = 1) \\ &= \Pr(Y^*_{i1} \leq 0, \dots, Y^*_{ij} > 0, \dots, Y^*_{iJ} \leq 0) \\ &= \int_{-\infty}^{0} \cdots \int_{0}^{\infty} \cdots \int_{-\infty}^{0} N(y|\mu_i, \Sigma) dy_{i1} \cdots dy_{ij} \cdots dy_{iJ} \end{aligned}$$

Computational and Estimation issues

Computational and Estimation issues

- No analytical solution is known to the integral

Computational and Estimation issues

- No analytical solution is known to the integral
- Moreover, # of parameters in Σ_J increases as J gets large, but data contain little information about Σ_J:

J	3	4	5	6	7
# of elements in Σ_J	6	10	15	21	28
# of parameters identified	2	5	9	14	20

Computational and Estimation issues

- No analytical solution is known to the integral
- Moreover, # of parameters in Σ_J increases as J gets large, but data contain little information about Σ_J:

J	3	4	5	6	7
# of elements in Σ_J	6	10	15	21	28
# of parameters identified	2	5	9	14	20

- Consequently, MNP is only feasible when J is small

Estimating the Dispersion Parameter

Estimating the Dispersion Parameter

In GLMs, ϕ is typically estimated sequentially after $\hat{\beta}$ is obtained by MLE.

Estimating the Dispersion Parameter

In GLMs, ϕ is typically estimated sequentially after $\hat{\beta}$ is obtained by MLE.
Two common methods for the common case of $a(\phi) = \phi/\omega_i$:

1. Calculate the (unscaled) deviance:

$$D(Y;\hat{\mu}) \;\equiv\; \phi D^*(Y;\hat{\mu}) \;=\; 2\sum_{i=1}^{n}\omega_i\left\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))\right\}$$

   which approximately follows $\phi \cdot \chi^2_{n-k}$ (because $D^*(Y;\hat{\mu}) \overset{approx.}{\sim} \chi^2_{n-k}$)

Estimating the Dispersion Parameter

In GLMs, ϕ is typically estimated sequentially after $\hat{\beta}$ is obtained by MLE.
Two common methods for the common case of $a(\phi) = \phi/\omega_i$:

1. Calculate the (unscaled) deviance:

$$D(Y;\hat{\mu}) \;\equiv\; \phi D^*(Y;\hat{\mu}) \;=\; 2\sum_{i=1}^{n}\omega_i\left\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))\right\}$$

which approximately follows $\phi \cdot \chi^2_{n-k}$ (because $D^*(Y;\hat{\mu}) \overset{approx.}{\sim} \chi^2_{n-k}$)

Then estimate ϕ by $\hat{\phi}_D \;=\; \frac{D(Y;\hat{\mu})}{n-k}$ (because $\mathbb{E}\left[\chi^2_{n-k}\right] = n - k$)

Estimating the Dispersion Parameter

In GLMs, ϕ is typically estimated sequentially after $\hat{\beta}$ is obtained by MLE. Two common methods for the common case of $a(\phi) = \phi/\omega_i$:

1. Calculate the (unscaled) deviance:

$$D(Y;\hat{\mu}) \;\equiv\; \phi D^*(Y;\hat{\mu}) \;=\; 2\sum_{i=1}^{n}\omega_i\left\{Y_i(\tilde{\theta}_i-\hat{\theta}_i)-(b(\tilde{\theta}_i)-b(\hat{\theta}_i))\right\}$$

   which approximately follows $\phi\cdot\chi^2_{n-k}$ (because $D^*(Y;\hat{\mu}) \overset{approx.}{\sim} \chi^2_{n-k}$)

   Then estimate ϕ by $\hat{\phi}_D \;=\; \frac{D(Y;\hat{\mu})}{n-k}$ (because $\mathbb{E}\left[\chi^2_{n-k}\right] = n-k$)

2. Calculate the generalized Pearson χ^2 statistic:

$$\mathcal{X}^2 \;\equiv\; \sum_{i=1}^{n}\frac{\omega_i(Y_i-\hat{\mu}_i)^2}{b''(\hat{\theta}_i)} \;=\; \phi\sum_{i=1}^{n}\frac{(Y_i-\hat{\mu}_i)^2}{\widehat{V}_i}$$

   which also approximately follows $\phi\cdot\chi^2_{n-k}$

Estimating the Dispersion Parameter

In GLMs, ϕ is typically estimated sequentially after $\hat{\beta}$ is obtained by MLE. Two common methods for the common case of $a(\phi) = \phi/\omega_i$:

1. Calculate the (unscaled) deviance:

$$D(Y; \hat{\mu}) \equiv \phi D^*(Y; \hat{\mu}) = 2\sum_{i=1}^{n} \omega_i \left\{ Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i)) \right\}$$

   which approximately follows $\phi \cdot \chi^2_{n-k}$ (because $D^*(Y; \hat{\mu}) \overset{approx.}{\sim} \chi^2_{n-k}$)

   Then estimate ϕ by $\hat{\phi}_D = \frac{D(Y;\hat{\mu})}{n-k}$ (because $\mathbb{E}\left[\chi^2_{n-k}\right] = n - k$)

2. Calculate the generalized Pearson χ^2 statistic:

$$\mathcal{X}^2 \equiv \sum_{i=1}^{n} \frac{\omega_i(Y_i - \hat{\mu}_i)^2}{b''(\hat{\theta}_i)} = \phi \sum_{i=1}^{n} \frac{(Y_i - \hat{\mu}_i)^2}{\widehat{V}_i}$$

   which also approximately follows $\phi \cdot \chi^2_{n-k}$

   Then estimate ϕ by $\hat{\phi}_P = \frac{\mathcal{X}^2}{n-k}$ (by the same logic)

Estimating the Dispersion Parameter

In GLMs, ϕ is typically estimated sequentially after $\hat{\beta}$ is obtained by MLE. Two common methods for the common case of $a(\phi) = \phi/\omega_i$:

1. Calculate the (unscaled) deviance:

$$D(Y;\hat{\mu}) \;\equiv\; \phi D^*(Y;\hat{\mu}) \;=\; 2\sum_{i=1}^{n}\omega_i\left\{Y_i(\tilde{\theta}_i - \hat{\theta}_i) - (b(\tilde{\theta}_i) - b(\hat{\theta}_i))\right\}$$

   which approximately follows $\phi \cdot \chi^2_{n-k}$ (because $D^*(Y;\hat{\mu}) \overset{approx.}{\sim} \chi^2_{n-k}$)

   Then estimate ϕ by $\hat{\phi}_D \;=\; \frac{D(Y;\hat{\mu})}{n-k}$ (because $\mathbb{E}\left[\chi^2_{n-k}\right] = n-k$)

2. Calculate the generalized Pearson χ^2 statistic:

$$\mathcal{X}^2 \;\equiv\; \sum_{i=1}^{n}\frac{\omega_i(Y_i - \hat{\mu}_i)^2}{b''(\hat{\theta}_i)} \;=\; \phi\sum_{i=1}^{n}\frac{(Y_i - \hat{\mu}_i)^2}{\widehat{V}_i}$$

   which also approximately follows $\phi \cdot \chi^2_{n-k}$

   Then estimate ϕ by $\hat{\phi}_P = \frac{\mathcal{X}^2}{n-k}$ (by the same logic)

When $Y_i \sim \mathcal{N}$, $D = \mathcal{X}^2 \sim \phi\chi^2_{n-k}$ exactly, and $\hat{\phi}_D$ and $\hat{\phi}_P$ are identical and MLE

Relaxing the Distributional Assumption

- Note that in the overdispersed Poisson model, Y_i *cannot* be Poisson distributed
- How can this seemingly arbitrary modification justified?

Relaxing the Distributional Assumption

- Note that in the overdispersed Poisson model, Y_i *cannot* be Poisson distributed
- How can this seemingly arbitrary modification justified?
- In GLMs, we can replace the distributional assumption with the variance function assumption:
  1. Systematic component: $X_i^\top \beta = X_i^\top \beta$
  2. Link function: $X_i^\top \beta = g(\mu_i)$ where $\mu_i \equiv \mathbb{E}(Y_i \mid X)$
  3. Variance function: $\mathbb{V}(Y_i \mid X) = \phi \psi(\mu_i)$

Relaxing the Distributional Assumption

- Note that in the overdispersed Poisson model, Y_i *cannot* be Poisson distributed
- How can this seemingly arbitrary modification justified?
- In GLMs, we can replace the distributional assumption with the variance function assumption:
  1. Systematic component: $X_i^\top \beta = X_i^\top \beta$
  2. Link function: $X_i^\top \beta = g(\mu_i)$ where $\mu_i \equiv \mathbb{E}(Y_i \mid X)$
  3. Variance function: $\mathbb{V}(Y_i \mid X) = \phi\psi(\mu_i)$

Relaxing the Distributional Assumption

- Note that in the overdispersed Poisson model, Y_i *cannot* be Poisson distributed
- How can this seemingly arbitrary modification justified?
- In GLMs, we can replace the distributional assumption with the variance function assumption:
  1. Systematic component: $X_i^\top \beta = X_i^\top \beta$
  2. Link function: $X_i^\top \beta = g(\mu_i)$ where $\mu_i \equiv \mathbb{E}(Y_i \mid X)$
  3. Variance function: $\mathbb{V}(Y_i \mid X) = \phi\psi(\mu_i)$
- That is, we specify mean and variance, but remain agnostic about the rest of $f(Y)$ (i.e. likelihood)

Relaxing the Distributional Assumption

- Note that in the overdispersed Poisson model, Y_i *cannot* be Poisson distributed
- How can this seemingly arbitrary modification justified?
- In GLMs, we can replace the distributional assumption with the variance function assumption:
  1. Systematic component: $X_i^\top \beta = X_i^\top \beta$
  2. Link function: $X_i^\top \beta = g(\mu_i)$ where $\mu_i \equiv \mathbb{E}(Y_i \mid X)$
  3. Variance function: $\mathbb{V}(Y_i \mid X) = \phi\psi(\mu_i)$
- That is, we specify mean and variance, but remain agnostic about the rest of $f(Y)$ (i.e. likelihood)
- With this reduced set of assumptions, what can we learn?

Relaxing the Distributional Assumption

- Note that in the overdispersed Poisson model, Y_i *cannot* be Poisson distributed
- How can this seemingly arbitrary modification justified?
- In GLMs, we can replace the distributional assumption with the variance function assumption:
  1. Systematic component: $X_i^\top \beta = X_i^\top \beta$
  2. Link function: $X_i^\top \beta = g(\mu_i)$ where $\mu_i \equiv \mathbb{E}(Y_i \mid X)$
  3. Variance function: $\mathbb{V}(Y_i \mid X) = \phi\psi(\mu_i)$
- That is, we specify mean and variance, but remain agnostic about the rest of $f(Y)$ (i.e. likelihood)
- With this reduced set of assumptions, what can we learn?
- Bottom line: We lose **nothing**, thanks to the properties of the exponential family

Negative Binomial Regression Models

- The overdispersed Poisson model is also called the negative binomial 1 (NB1) model

Negative Binomial Regression Models

- The overdispersed Poisson model is also called the negative binomial 1 (NB1) model
- An alternative parameterization for allowing overdispersion:

$$\mathbb{E}(Y_i \mid X_i) = \mu_i \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = \mu_i + \mu_i^2/\gamma > \mathbb{E}(Y_i \mid X_i)$$

- This is called the negative binomial 2 (NB2) model

Negative Binomial Regression Models

- The overdispersed Poisson model is also called the negative binomial 1 (NB1) model
- An alternative parameterization for allowing overdispersion:

$$\mathbb{E}(Y_i \mid X_i) = \mu_i \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = \mu_i + \mu_i^2/\gamma > \mathbb{E}(Y_i \mid X_i)$$

- This is called the negative binomial 2 (NB2) model
- The NB2 model *corresponds to* the following PMF:

$$p(Y_i \mid \mu_i, \gamma) \;=\; \frac{\Gamma(Y_i + \gamma)}{Y_i!\Gamma(\gamma)} \left(\frac{\mu_i}{\mu_i + \gamma}\right)^{Y_i} \left(\frac{\gamma}{\mu_i + \gamma}\right)^{\gamma}$$

where $\mu_i = \exp(X_i^\top \beta)$

Negative Binomial Regression Models

- The overdispersed Poisson model is also called the negative binomial 1 (NB1) model
- An alternative parameterization for allowing overdispersion:

$$\mathbb{E}(Y_i \mid X_i) = \mu_i \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = \mu_i + \mu_i^2/\gamma > \mathbb{E}(Y_i \mid X_i)$$

- This is called the negative binomial 2 (NB2) model
- The NB2 model *corresponds to* the following PMF:

$$p(Y_i \mid \mu_i, \gamma) = \frac{\Gamma(Y_i + \gamma)}{Y_i!\Gamma(\gamma)} \left(\frac{\mu_i}{\mu_i + \gamma}\right)^{Y_i} \left(\frac{\gamma}{\mu_i + \gamma}\right)^{\gamma}$$

  where $\mu_i = \exp(X_i^\top \beta)$

- (But keep in mind, you don't need to exactly *assume* this PMF!)

Negative Binomial Regression Models

- The overdispersed Poisson model is also called the negative binomial 1 (NB1) model
- An alternative parameterization for allowing overdispersion:

$$\mathbb{E}(Y_i \mid X_i) = \mu_i \quad \text{and} \quad \mathbb{V}(Y_i \mid X_i) = \mu_i + \mu_i^2/\gamma > \mathbb{E}(Y_i \mid X_i)$$

- This is called the negative binomial 2 (NB2) model
- The NB2 model *corresponds to* the following PMF:

$$p(Y_i \mid \mu_i, \gamma) \;=\; \frac{\Gamma(Y_i + \gamma)}{Y_i!\Gamma(\gamma)} \left(\frac{\mu_i}{\mu_i + \gamma}\right)^{Y_i} \left(\frac{\gamma}{\mu_i + \gamma}\right)^{\gamma}$$

  where $\mu_i = \exp(X_i^\top \beta)$

- (But keep in mind, you don't need to exactly *assume* this PMF!)
- Estimation via (Q)MLE

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i} \pi_i^{y_i} (1-\pi_i)^{N_i - y_i}$$

where

$$\pi_i = [1 + e^{-x_i \beta}]^{-1}$$

which implies

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

and a likelihood of

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i - y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

and a likelihood of

$$L(\pi|y) \propto \prod_{i=1}^{n} \text{Binomial}(y_i|\pi_i)$$

Grouped Uncorrelated Binary Variables

Same model as binary logit, but we only observe sums of iid groups of Bernoulli trials. E.g., the number of times you voted out of the last 5 elections.

$$Y_i \sim \text{Binomial}(y_i|\pi_i) = \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i}$$

where

$$\pi_i = [1 + e^{-x_i\beta}]^{-1}$$

which implies

$$E(Y_i) \equiv \mu_i = N_i\pi_i = N_i[1 + e^{-x_i\beta}]^{-1}$$

and a likelihood of

$$\begin{aligned} L(\pi|y) &\propto \prod_{i=1}^{n} \text{Binomial}(y_i|\pi_i) \\ &= \prod_{i=1}^{n} \binom{N_i}{y_i}\pi_i^{y_i}(1-\pi_i)^{N_i-y_i} \end{aligned}$$

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

and after substituting in the systematic component:

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

and after substituting in the systematic component:

$$\ln L(\beta|y) \doteq \sum_{i=1}^{n} \left\{ -y_i \ln[1 + e^{-x_i\beta}] + (N_i - y_i) \ln \left(1 - [1 + e^{-x_i\beta}]^{-1} \right) \right\}$$

Grouped Uncorrelated Binary Variables

The Log-likelihood is then:

$$\ln L(\pi|y) = \sum_{i=1}^{n} \left\{ \ln \binom{N_i}{y_i} + y_i \ln \pi_i + (N_i - y_i) \ln(1 - \pi_i) \right\}$$

and after substituting in the systematic component:

$$\begin{aligned} \ln L(\beta|y) &\doteq \sum_{i=1}^{n} \left\{ -y_i \ln[1 + e^{-x_i\beta}] + (N_i - y_i) \ln\left(1 - [1 + e^{-x_i\beta}]^{-1}\right) \right\} \\ &= \sum_{i=1}^{n} \left\{ (N_i - y_i) \ln(1 + e^{x_i\beta}) - y_i \ln(1 + e^{-x_i\beta}) \right\} \end{aligned}$$

Grouped Uncorrelated Binary Variables

Notes:

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
    (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:

$$\tilde{\pi}_c = [1 + e^{-x_c \tilde{\beta}}]^{-1}$$

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c \tilde{\beta}}]^{-1}$$
   (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.

Grouped Uncorrelated Binary Variables

Notes:

1. Similar log-likelihood to binary logit
2. All inference is about the same π as in binary logit
3. How to simulate and compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ from the multivariate normal with mean vector $\tilde{\beta}$ and the variance matrix that come from optim.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c \tilde{\beta}}]^{-1}$$
   (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.
   (f) If simulations of y are needed, go one more step and draw $\tilde{y}$ from Binomial$(y_i|\pi_i)$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1 - \pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$f_{ebb}(y_i|\pi_i, \gamma) = \Pr(Y_i = y_i|\pi_i, \gamma, N)$$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$\begin{aligned} f_{ebb}(y_i|\pi_i,\gamma) &= \Pr(Y_i = y_i|\pi_i, \gamma, N) \\ &= \frac{N!}{y_i!(N-y_i)!} \prod_{j=0}^{y_i-1}(\pi_i + \gamma j) \prod_{j=0}^{N-y_i-1}(1-\pi_i+\gamma j) / \prod_{j=0}^{N-1}(1+\gamma j) \end{aligned}$$

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$\begin{aligned} f_{ebb}(y_i|\pi_i,\gamma) &= \Pr(Y_i = y_i|\pi_i,\gamma,N) \\ &= \frac{N!}{y_i!(N-y_i)!}\prod_{j=0}^{y_i-1}(\pi_i+\gamma j)\prod_{j=0}^{N-y_i-1}(1-\pi_i+\gamma j)/\prod_{j=0}^{N-1}(1+\gamma j) \end{aligned}$$

and

Grouped Correlated Binary Variables

In the binomial-logit model, $V(Y) = \pi_i(1-\pi_i)/N_i$, with no σ^2-like parameter to take up slack. The beta-binomial (or extended BB) adds this extra parameter. The model:

$$Y_i \sim f_{ebb}(y_i|\pi_i, \gamma)$$

where, recall

$$\begin{aligned} f_{ebb}(y_i|\pi_i,\gamma) &= \Pr(Y_i = y_i|\pi_i,\gamma,N) \\ &= \frac{N!}{y_i!(N-y_i)!} \prod_{j=0}^{y_i-1}(\pi_i+\gamma j) \prod_{j=0}^{N-y_i-1}(1-\pi_i+\gamma j) / \prod_{j=0}^{N-1}(1+\gamma j) \end{aligned}$$

and

$$\pi_i = \frac{1}{1+e^{-x_i\beta}}$$

The probability model of all the data:

The probability model of all the data:

$$
\begin{aligned}
\Pr(Y = y|\beta, \gamma; N) = & \prod_{i=1}^{n} \left(\frac{N!}{y_i!(N - y_i)!} \right) \\
& \times \prod_{j=0}^{y_i - 1} \left\{ [1 + \exp(-x_i\beta)]^{-1} + \gamma j \right\} \\
& \times \prod_{j=0}^{N - y_i - 1} \left\{ [1 + \exp(x_i\beta)]^{-1} + \gamma j \right\} / \prod_{j=0}^{N-1} (1 + \gamma j)
\end{aligned}
$$

The probability model of all the data:

$$
\begin{aligned}
\Pr(Y = y|\beta, \gamma; N) = & \prod_{i=1}^{n} \left(\frac{N!}{y_i!(N - y_i)!} \right) \\
& \times \prod_{j=0}^{y_i - 1} \left\{ [1 + \exp(-x_i\beta)]^{-1} + \gamma j \right\} \\
& \times \prod_{j=0}^{N - y_i - 1} \left\{ [1 + \exp(x_i\beta)]^{-1} + \gamma j \right\} / \prod_{j=0}^{N-1} (1 + \gamma j) \\
\ln L(\beta, \gamma|y) = & \sum_{i=1}^{n} \left\{ \ln \left(\frac{N!}{y_i!(N - y_i)!} \right) \right. \\
& + \sum_{j=0}^{y_i - 1} \ln \left\{ [1 + \exp(-x_i\beta)]^{-1} + \gamma j \right\} \\
& \left. + \sum_{j=0}^{N - y_i - 1} \ln \left\{ [1 + \exp(x_i\beta)]^{-1} + \gamma j \right\} - \sum_{j=0}^{N-1} \ln(1 + \gamma j) \right\}
\end{aligned}
$$

The probability model of all the data:

$$
\begin{aligned}
\Pr(Y = y|\beta,\gamma;N) &= \prod_{i=1}^{n}\left(\frac{N!}{y_i!(N-y_i)!}\right)\\
&\times \prod_{j=0}^{y_i-1}\left\{[1+\exp(-x_i\beta)]^{-1}+\gamma j\right\}\\
&\times \prod_{j=0}^{N-y_i-1}\left\{[1+\exp(x_i\beta)]^{-1}+\gamma j\right\} / \prod_{j=0}^{N-1}(1+\gamma j)\\
\ln L(\beta,\gamma|y) &= \sum_{i=1}^{n}\left\{\ln\left(\frac{N!}{y_i!(N-y_i)!}\right)\right.\\
&+ \sum_{j=0}^{y_i-1}\ln\left\{[1+\exp(-x_i\beta)]^{-1}+\gamma j\right\}\\
&+ \left.\sum_{j=0}^{N-y_i-1}\ln\left\{[1+\exp(x_i\beta)]^{-1}+\gamma j\right\} - \sum_{j=0}^{N-1}\ln(1+\gamma j)\right\}\\
&\doteq \sum_{i=1}^{n}\left\{\sum_{j=0}^{y_i-1}\ln\left\{[1+\exp(-x_i\beta)]^{-1}+\gamma j\right\}\right.\\
&+ \left.\sum_{j=0}^{N-y_i-1}\ln\left\{[1+\exp(x_i\beta)]^{-1}+\gamma j\right\} - \sum_{j=0}^{N-1}\ln(1+\gamma j)\right\}
\end{aligned}
$$

Notes:

Notes:

1. The math <u>looks</u> complicated.

Notes:

1. The math looks complicated.
2. The use of this model is simple.

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.

Notes:

1. The math <u>looks</u> complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\tilde{\gamma}$ from the multivariate normal with mean vector $\vec{(}\hat{\beta}, \hat{\gamma})$ and the variance matrix that come from `optim`.

Notes:

1. The math <u>looks</u> complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\tilde{\gamma}$ from the multivariate normal with mean vector $\vec{(}\hat{\beta}, \hat{\gamma})$ and the variance matrix that come from `optim`.
   (c) Set X to your choice of values, X_c

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\tilde{\gamma}$ from the multivariate normal with mean vector $\vec{(}\hat{\beta}, \hat{\gamma})$ and the variance matrix that come from `optim`.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\tilde{\gamma}$ from the multivariate normal with mean vector $\vec{(}\hat{\beta}, \hat{\gamma})$ and the variance matrix that come from `optim`.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c \tilde{\beta}}]^{-1}$$

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
   (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
   (b) Draw many values of $\tilde{\beta}$ and $\tilde{\gamma}$ from the multivariate normal with mean vector $\vec{(}\hat{\beta}, \hat{\gamma})$ and the variance matrix that come from `optim`.
   (c) Set X to your choice of values, X_c
   (d) Calculate simulations of the probability that any of the component binary variables is a one:
   $$\tilde{\pi}_c = [1 + e^{-x_c \tilde{\beta}}]^{-1}$$
   (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.

Notes:

1. The math looks complicated.
2. The use of this model is simple.
3. γ soaks up binomial misspecification
4. Assuming binomial when EBB is the right model causes se's to be wrong.
5. How to simulate to compute quantities of interest?
    (a) Run `optim`, and get $\hat{\beta}$, $\hat{\gamma}$ and the variance matrix.
    (b) Draw many values of $\tilde{\beta}$ and $\tilde{\gamma}$ from the multivariate normal with mean vector $\vec{(}\hat{\beta}, \hat{\gamma})$ and the variance matrix that come from `optim`.
    (c) Set X to your choice of values, X_c
    (d) Calculate simulations of the probability that any of the component binary variables is a one:
    $$\tilde{\pi}_c = [1 + e^{-x_c \tilde{\beta}}]^{-1}$$
    (e) If π is of interest, summarize with mean, SD, CI's, or histogram as needed.
    (f) If simulations of y are needed, go one more step and draw $\tilde{y}$ from $f_{ebb}(y_i|\pi_i)$

Generalized Gamma Distribution

The Weibull and Exponential distributions are special cases of the Generalized Gamma distribution, $Y \sim GGamma(\nu, \lambda, p)$:

$$f_Y(y) = \frac{p\lambda^{p\nu}}{\Gamma(\nu)} y^{p\nu-1} exp(-\lambda y)^p$$

Generalized Gamma Distribution

The Weibull and Exponential distributions are special cases of the Generalized Gamma distribution, $Y \sim GGamma(\nu, \lambda, p)$:

$$f_Y(y) = \frac{p\lambda^{p\nu}}{\Gamma(\nu)} y^{p\nu - 1} exp(-\lambda y)^p$$

When $p = 1$, $Y \sim Gamma(\nu, \lambda)$:

$f_Y(y) = \frac{\lambda^\nu}{\Gamma(\nu)} y^{\nu - 1} exp(-\lambda y)$

Generalized Gamma Distribution

The Weibull and Exponential distributions are special cases of the Generalized Gamma distribution, $Y \sim GGamma(\nu, \lambda, p)$:

$$f_Y(y) = \frac{p\lambda^{p\nu}}{\Gamma(\nu)} y^{p\nu - 1} exp(-\lambda y)^p$$

When $p = 1$, $Y \sim Gamma(\nu, \lambda)$:

$f_Y(y) = \frac{\lambda^\nu}{\Gamma(\nu)} y^{\nu-1} exp(-\lambda y)$

When $\nu = 1$, $Y \sim Weibull(\frac{1}{\lambda}, p)$:

$f_Y(y) = p\lambda^p y^{p-1} exp(-\lambda y)^p$

Generalized Gamma Distribution

The Weibull and Exponential distributions are special cases of the Generalized Gamma distribution, $Y \sim GGamma(\nu, \lambda, p)$:

$$f_Y(y) = \frac{p\lambda^{p\nu}}{\Gamma(\nu)} y^{p\nu - 1} exp(-\lambda y)^p$$

When $p = 1$, $Y \sim Gamma(\nu, \lambda)$:

$f_Y(y) = \frac{\lambda^\nu}{\Gamma(\nu)} y^{\nu-1} exp(-\lambda y)$

When $\nu = 1$, $Y \sim Weibull(\frac{1}{\lambda}, p)$:

$f_Y(y) = p\lambda^p y^{p-1} exp(-\lambda y)^p$

When $p = 1$ and $\nu = 1$, $Y \sim Expo(\lambda)$:

$f_Y(y) = \lambda exp(-\lambda y)$

When would a Gamma regression be appropriate?

For positive random variables with a skewed distribution, the variance often increases with the mean.

- Poisson random variable: $Var(Y) = E[Y] = \mu$.

When would a Gamma regression be appropriate?

For positive random variables with a skewed distribution, the variance often increases with the mean.

- Poisson random variable: $Var(Y) = E[Y] = \mu$.

Another case occurs where the standard-deviation increases linearly with the mean:

$$\sqrt{Var(Y)} \propto E(Y)$$

When would a Gamma regression be appropriate?

For positive random variables with a skewed distribution, the variance often increases with the mean.

- Poisson random variable: $Var(Y) = E[Y] = \mu$.

Another case occurs where the standard-deviation increases linearly with the mean:

$$\sqrt{Var(Y)} \propto E(Y)$$

In this case, the coefficient of variation (ratio of standard deviation to expectation) is constant:

$$c.v. = \frac{\sqrt{Var(Y)}}{E(Y)}$$

When would a Gamma regression be appropriate?

For positive random variables with a skewed distribution, the variance often increases with the mean.

- Poisson random variable: $Var(Y) = E[Y] = \mu$.

Another case occurs where the standard-deviation increases linearly with the mean:

$$\sqrt{Var(Y)} \propto E(Y)$$

In this case, the coefficient of variation (ratio of standard deviation to expectation) is constant:

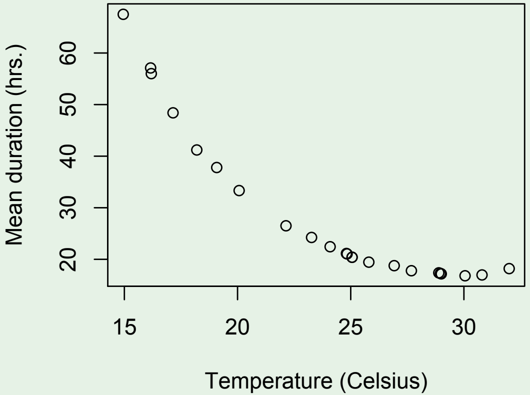
$$c.v. = \frac{\sqrt{Var(Y)}}{E(Y)}$$

The **Gamma distribution** has this property.

When would a Gamma regression be appropriate?

Example

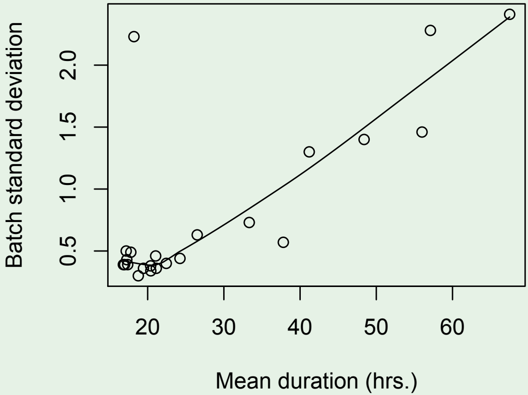
Mean duration of developmental period in *Drosphila melanogaster* (McCullagh & Nelder, 1989)



When would a Gamma regression be appropriate?

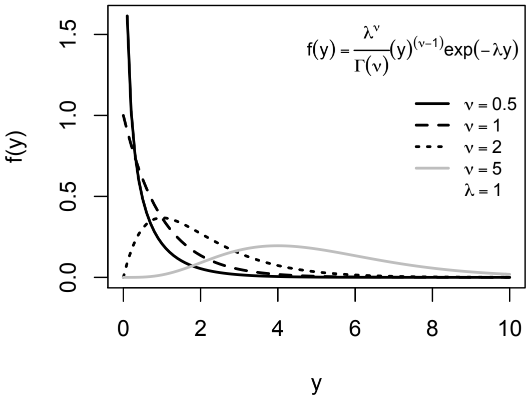
Example

Mean duration of developmental period in *Drosphila melanogaster* (McCullagh & Nelder, 1989)



Gamma shapes

ν is the shape parameter, λ is the scale parameter



Special cases:

$$\nu = 1 \implies \textit{Exponential}$$

$$\nu \to \infty \implies \textit{Normal}$$

Gamma as an EDF

$$
\begin{aligned}
f_Y(y) &= \frac{\lambda^{\nu}}{\Gamma(\nu)} y^{\nu-1} exp(-\lambda y) \\
&= exp\left(\frac{-\frac{\lambda}{\nu} y + ln(\frac{\lambda}{\nu})}{\nu^{-1}} + \nu ln(\nu y) - ln(y) - ln(\Gamma(\nu)) \right)
\end{aligned}
$$

Gamma as an EDF

$$
\begin{aligned}
f_Y(y) &= \frac{\lambda^\nu}{\Gamma(\nu)} y^{\nu-1} exp(-\lambda y) \\
&= exp\left(\frac{-\frac{\lambda}{\nu} y + ln(\frac{\lambda}{\nu})}{\nu^{-1}} + \nu ln(\nu y) - ln(y) - ln(\Gamma(\nu)) \right)
\end{aligned}
$$

Where

$$
\begin{aligned}
\theta &= -\frac{\lambda}{\nu} \\
\phi &= \nu^{-1} = \sigma^2 \\
b(\theta) &= -ln\left(\frac{\lambda}{\nu}\right) \\
E[Y] &= b'(\theta) = \frac{\nu}{\lambda} = \mu \\
Var(Y) &= \phi b''(\theta) = \frac{1}{\nu}\frac{\nu^2}{\lambda^2} = \sigma^2 \mu^2
\end{aligned}
$$

Link functions

Canonical link

$$\eta = \theta = -\frac{1}{\mu}$$

The reciprocal transformation does not map the range of μ onto the whole real line.

The requirement that $\mu > 0$ places restrictions on β's.

The canonical link is rarely used.

Link functions

Inverse polynomial: linear

$$\eta = \mu^{-1} = \beta_0 + \beta_1/x$$

Inverse polynomial: quadratic

$$\eta = \mu^{-1} = \beta_0 + \beta_1 x + \beta_2/x$$

Inverse polynomials have appealing property that η is everywhere positive and bounded.

Application: sometimes used in plant density experiments, where yield per plant (y_i) varies inversely with plant density (x_i)

Link functions

Log link

$$\eta = ln(\mu) = \beta_0 + \beta_1 x$$

$$\eta = ln(\mu) = \beta_0 + \beta_1 x + \beta_2 / x$$

Application: useful for describing functions that have turning points, but are noticeably asymmetric around that point.

Link functions

Identity link

$$\eta = \mu = \beta_0 + \beta_1 x$$

Application: used for modeling variance components.

Maximum Likelihood Estimation

$$\mathcal{L} = \prod_{i=1}^{n} \frac{\lambda^{\nu}}{\Gamma(\nu)} y_i^{\nu-1} exp(-\lambda y_i)$$

$$\begin{aligned}
\ln\mathcal{L} &= \sum_{i=1}^{n} \ln\left[\frac{\lambda^{\nu}}{\Gamma(\nu)} y_i^{\nu-1} exp(-\lambda y_i)\right] \\
&= \sum_{i=1}^{n} \nu \ln\lambda - \ln\Gamma(\nu) + (\nu-1)\ln y_i - \lambda y_i
\end{aligned}$$

Gamma regression with weights

Suppose your data consist of n observations, each from a separate group $i \in \{1, \ldots, n\}$. Each group has n_i individuals.

Gamma regression with weights

Suppose your data consist of n observations, each from a separate group $i \in \{1, \ldots, n\}$. Each group has n_i individuals.

Example

Y_i is the duration of embryonic period in n_i batches of fruit flies

$$Y_i = \sum_{j=1}^{n_i} Y_{ij} \quad Y_{ij} = \text{duration for } j \text{ embryo in } i\text{-th batch}$$

$$Y_i^s = Y_i / n_i = \text{ average duration in } i\text{-th batch}$$

Gamma regression with weights

Suppose your data consist of n observations, each from a separate group $i \in \{1, \ldots, n\}$. Each group has n_i individuals.

Example

Y_i is the duration of embryonic period in n_i batches of fruit flies

$$Y_i = \sum_{j=1}^{n_i} Y_{ij} \quad Y_{ij} = \text{duration for } j \text{ embryo in } i\text{-th batch}$$

$$Y_i^s = Y_i/n_i = \text{ average duration in } i\text{-th batch}$$

If $Y_{ij} \sim Gamma(\lambda_i, \nu)$, independent, with $\lambda_i = \nu/\mu_i$:

$$E[Y_i^s] = \frac{1}{n_i}\sum_{j=1}^{n_i} E[Y_{ij}] = \frac{1}{n_i}\sum_{j=1}^{n_i} \mu_i = \mu_i$$

$$Var(Y_i^s) = \frac{1}{n_i} Var(Y_i) = \frac{\sigma^2 \mu_i^2}{n_i} \qquad \text{weights} = n_i$$

Application: *Drosophila melanogaster*

4 models estimated:

1. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i$

Application: *Drosophila melanogaster*

4 models estimated:

1. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i$
2. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 / \text{Temp}_i$

Application: *Drosophila melanogaster*

4 models estimated:

1. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i$
2. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 / \text{Temp}_i$
3. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 / \text{Temp}_i$
   (weighted by batch size)

Application: *Drosophila melanogaster*

4 models estimated:

1. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i$
2. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 / \text{Temp}_i$
3. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 / \text{Temp}_i$
   (weighted by batch size)
4. $log(\text{Duration}_i) = \beta_0 + \beta_1 \text{Temp}_i + \beta_2 / (\text{Temp}_i - \delta)$
   (weighted by batch size)

Application: *Drosophila melanogaster*