

Week 5: Simple Linear Regression

Brandon Stewart¹

Princeton

September 28-October 2, 2020

¹These slides are heavily influenced by Matt Blackwell, Adam Glynn, Erin Hartman and Jens Hainmueller. Illustrations by Shay O'Brien.

Where We've Been and Where We're Going...

- Last Week
 - ▶ hypothesis testing
 - ▶ what is regression
- This Week
 - ▶ mechanics and properties of simple linear regression
 - ▶ inference and measures of model fit
 - ▶ confidence intervals for regression
 - ▶ goodness of fit
- Next Week
 - ▶ mechanics with two regressors
 - ▶ omitted variables, multicollinearity
- Long Run
 - ▶ probability → inference → regression → causal inference

Macrostructure—This Semester

The next few weeks,

- Linear Regression with Two Regressors
- Break Week and Multiple Linear Regression
- Rethinking Regression
- Regression in the Social Sciences
- Causality with Measured Confounding
- Unmeasured Confounding and Instrumental Variables
- Repeated Observations and Panel Data
- Review and Final Discussion

- 1 Mechanics of OLS
- 2 Classical Perspective (Part 1, Unbiasedness)
 - Sampling Distributions
 - Classical Assumptions 1–4
- 3 Classical Perspective: Variance
 - Sampling Variance
 - Gauss-Markov
 - Large Samples
 - Small Samples
 - Agnostic Perspective
- 4 Inference
 - Hypothesis Tests
 - Confidence Intervals
 - Goodness of fit
 - Interpretation
- 5 Non-linearities
 - Log Transformations
 - Fun With Logs
 - LOESS

Narrow Goal: Understand `lm()` Output

Call:

```
lm(formula = sr ~ pop15, data = LifeCycleSavings)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.637	-2.374	0.349	2.022	11.155

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	17.49660	2.27972	7.675	6.85e-10	***
pop15	-0.22302	0.06291	-3.545	0.000887	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.03 on 48 degrees of freedom

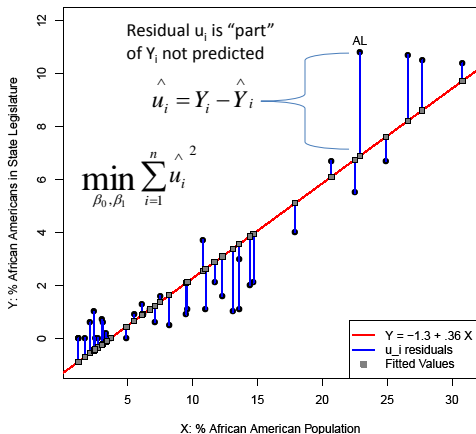
Multiple R-squared: 0.2075, Adjusted R-squared: 0.191

F-statistic: 12.57 on 1 and 48 DF, p-value: 0.0008866

Reminder

How do we fit the regression line $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ to the data?

Answer: We will **minimize the squared sum of residuals**



The Population Quantity

- Broadly speaking we are interested in the **conditional expectation function** (CEF) in part because it minimizes the **mean squared error**.
- The CEF has a potentially **arbitrary** shape but there is always a **best linear predictor** (BLP) or **linear projection** which is the line given by:

$$g(X) = \beta_0 + \beta_1 X$$

$$\beta_0 = E[Y] - \frac{\text{Cov}[X, Y]}{V[X]} E[X]$$

$$\beta_1 = \frac{\text{Cov}[X, Y]}{V[X]}$$

- This **may** not be a good approximation depending on how non-linear the true CEF is. However, it provides us with a reasonable **target** that always exists.
- Define deviations from the BLP as

$$u = Y - g(X)$$

then, the following properties hold:

$$(1) E[u] = 0, \quad (2) E[Xu] = 0, \quad (3) \text{Cov}[X, u] = 0$$

What is OLS?

- The **best linear predictor** is the line that minimizes

$$(\beta_0, \beta_1) = \arg \min_{b_0, b_1} E[(Y - b_0 - b_1 X)^2]$$

- **Ordinary Least Squares** (OLS) is a method for minimizing the sample analog of this quantity. It solves the optimization problem:

$$(\hat{\beta}_0, \hat{\beta}_1) = \arg \min_{b_0, b_1} \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

- In words, the OLS estimates are the intercept and slope that minimize the **sum of the squared residuals**.
- There are many loss functions, but OLS uses the **squared error loss** which is connected to the **conditional expectation function**. If we chose a different loss, we would target a different feature of the conditional distribution.

Deriving the OLS estimator

- Let's think about n pairs of sample observations:
 $(Y_1, X_1), (Y_2, X_2), \dots, (Y_n, X_n)$
- Let $\{b_0, b_1\}$ be possible values for $\{\beta_0, \beta_1\}$
- Define the **least squares objective function**:

$$S(b_0, b_1) = \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2.$$

- How do we derive the LS estimators for β_0 and β_1 ? We want to minimize this function, which is actually a very well-defined calculus problem.
 - 1 Take partial derivatives of S with respect to b_0 and b_1 .
 - 2 Set each of the partial derivatives to 0
 - 3 Solve for $\{b_0, b_1\}$ and replace them with the solutions
- We are going to step through this process together.

Step 1: Take Partial Derivatives

$$\begin{aligned}S(b_0, b_1) &= \sum_{i=1}^n (Y_i - b_0 - X_i b_1)^2 \\ &= \sum_{i=1}^n (Y_i^2 - 2Y_i b_0 - 2Y_i b_1 X_i + b_0^2 + 2b_0 b_1 X_i + b_1^2 X_i^2)\end{aligned}$$

$$\begin{aligned}\frac{\partial S(b_0, b_1)}{\partial b_0} &= \sum_{i=1}^n (-2Y_i + 2b_0 + 2b_1 X_i) \\ &= -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i)\end{aligned}$$

$$\begin{aligned}\frac{\partial S(b_0, b_1)}{\partial b_1} &= \sum_{i=1}^n (-2Y_i X_i + 2b_0 X_i + 2b_1 X_i^2) \\ &= -2 \sum_{i=1}^n X_i (Y_i - b_0 - b_1 X_i)\end{aligned}$$

Solving for the Intercept

$$\begin{aligned}\frac{\partial S(b_0, b_1)}{\partial b_0} &= -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) \\ 0 &= -2 \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) \\ 0 &= \sum_{i=1}^n (Y_i - b_0 - b_1 X_i) \\ 0 &= \sum_{i=1}^n Y_i - \sum_{i=1}^n b_0 - \sum_{i=1}^n b_1 X_i \\ b_0 n &= \left(\sum_{i=1}^n Y_i \right) - b_1 \left(\sum_{i=1}^n X_i \right) \\ b_0 &= \bar{Y} - b_1 \bar{X}\end{aligned}$$

A Helpful Lemma on Deviations from Means

Lemmas are like helper results that are often invoked repeatedly.

Lemma (Deviations from the Mean Sum to 0)

$$\begin{aligned}\sum_{i=1}^n (X_i - \bar{X}) &= \left(\sum_{i=1}^n X_i \right) - n\bar{X} \\ &= \left(\sum_{i=1}^n X_i \right) - n \sum_{i=1}^n X_i / n \\ &= \left(\sum_{i=1}^n X_i \right) - \sum_{i=1}^n X_i \\ &= 0\end{aligned}$$

Solving for the Slope

$$0 = -2 \sum_{i=1}^n X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^n X_i(Y_i - b_0 - b_1 X_i)$$

$$0 = \sum_{i=1}^n X_i(Y_i - (\bar{Y} - b_1 \bar{X}) - b_1 X_i) \quad (\text{sub in } b_0)$$

$$0 = \sum_{i=1}^n X_i(Y_i - \bar{Y} - b_1(X_i - \bar{X}))$$

$$0 = \sum_{i=1}^n X_i(Y_i - \bar{Y}) - b_1 \sum_{i=1}^n X_i(X_i - \bar{X})$$

$$b_1 \sum_{i=1}^n X_i(X_i - \bar{X}) = \sum_{i=1}^n X_i(Y_i - \bar{Y}) - \bar{X} \sum_{i=1}^n (Y_i - \bar{Y}) \quad (\text{add } 0)$$

Solving for the Slope

$$b_1 \sum_{i=1}^n X_i(X_i - \bar{X}) = \sum_{i=1}^n X_i(Y_i - \bar{Y}) - \bar{X} \sum_{i=1}^n (Y_i - \bar{Y})$$

$$b_1 \sum_{i=1}^n X_i(X_i - \bar{X}) = \sum_{i=1}^n X_i(Y_i - \bar{Y}) - \sum_{i=1}^n \bar{X}(Y_i - \bar{Y})$$

$$b_1 \left(\sum_{i=1}^n X_i(X_i - \bar{X}) - \sum_{i=1}^n \bar{X}(X_i - \bar{X}) \right) = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

add 0

$$b_1 \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) = \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

$$b_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

The OLS estimator

- Now we're done! Here are the **OLS estimators**:

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

Intuition of the OLS estimator

- The intercept equation tells us that the regression line goes through the point (\bar{Y}, \bar{X}) :

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X}$$

- The slope for the regression line can be written as the following:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

- The higher the **covariance** between X and Y , the higher the **slope** will be.
- Negative covariances \rightarrow negative slopes;
positive covariances \rightarrow positive slopes
- If X_i doesn't vary, the denominator is undefined.
- If Y_i doesn't vary, you get a flat line.

Mechanical properties of OLS

- Later we'll see that under certain assumptions, OLS will have nice statistical properties.
- But some properties are mechanical since they can be derived from the first order conditions of OLS.
- ① The sample mean of the residuals will be zero:

$$\frac{1}{n} \sum_{i=1}^n \hat{u}_i = 0$$

- ② The residuals will be uncorrelated with the predictor ($\widehat{\text{Cov}}$ is the sample covariance):

$$\sum_{i=1}^n X_i \hat{u}_i = 0 \implies \widehat{\text{Cov}}(X_i, \hat{u}_i) = 0$$

- ③ The residuals will be uncorrelated with the fitted values:

$$\sum_{i=1}^n \hat{Y}_i \hat{u}_i = 0 \implies \widehat{\text{Cov}}(\hat{Y}_i, \hat{u}_i) = 0$$

OLS slope as a weighted sum of the outcomes

- One useful derivation is to write the OLS estimator for the slope as a weighted sum of the outcomes.

$$\hat{\beta}_1 = \sum_{i=1}^n W_i Y_i$$

- Where here we have the weights, W_i as:

$$W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- This is important for two reasons. First, it'll make derivations later much easier. And second, it shows that is just the sum of a random variable. Therefore it is also a random variable.

Lemma 2: OLS as a Weighted Sum of Outcomes

Lemma (OLS as Weighted Sum of Outcomes)

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} - \frac{\sum_{i=1}^n (X_i - \bar{X})\bar{Y}}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i}{\sum_{i=1}^n (X_i - \bar{X})^2} \\ &= \sum_{i=1}^n W_i Y_i\end{aligned}$$

Where the weights, W_i are:

$$W_i = \frac{(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

We Covered

- A brief review of regression
- Derivation of the OLS estimator
- OLS as a weighted sum of outcomes

Next Time: The Classical Perspective

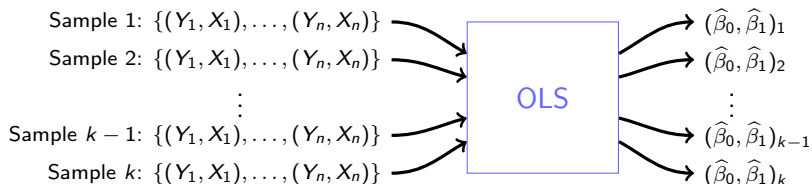
Where We've Been and Where We're Going...

- Last Week
 - ▶ hypothesis testing
 - ▶ what is regression
- This Week
 - ▶ mechanics and properties of simple linear regression
 - ▶ inference and measures of model fit
 - ▶ confidence intervals for regression
 - ▶ goodness of fit
- Next Week
 - ▶ mechanics with two regressors
 - ▶ omitted variables, multicollinearity
- Long Run
 - ▶ probability → inference → regression → causal inference

- 1 Mechanics of OLS
- 2 Classical Perspective (Part 1, Unbiasedness)
 - Sampling Distributions
 - Classical Assumptions 1–4
- 3 Classical Perspective: Variance
 - Sampling Variance
 - Gauss-Markov
 - Large Samples
 - Small Samples
 - Agnostic Perspective
- 4 Inference
 - Hypothesis Tests
 - Confidence Intervals
 - Goodness of fit
 - Interpretation
- 5 Non-linearities
 - Log Transformations
 - Fun With Logs
 - LOESS

Sampling distribution of the OLS estimator

- Remember: OLS is an estimator—it's a machine that we plug samples into and we get out estimates.

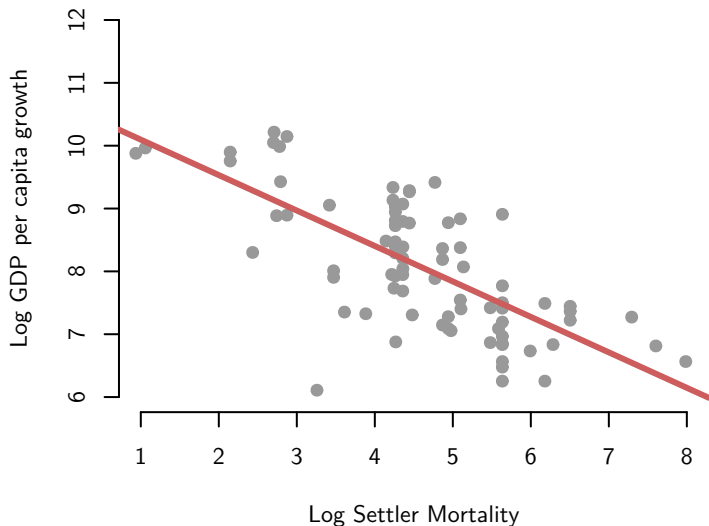


- Just like the sample mean, sample difference in means, or the sample variance
- It has a sampling distribution, with a sampling variance/standard error, etc.
- Let's take a simulation approach to demonstrate:
 - Let's use some data from Acemoglu, Daron, Simon Johnson, and James A. Robinson. "The colonial origins of comparative development: An empirical investigation." 2000
 - See how the line varies from sample to sample

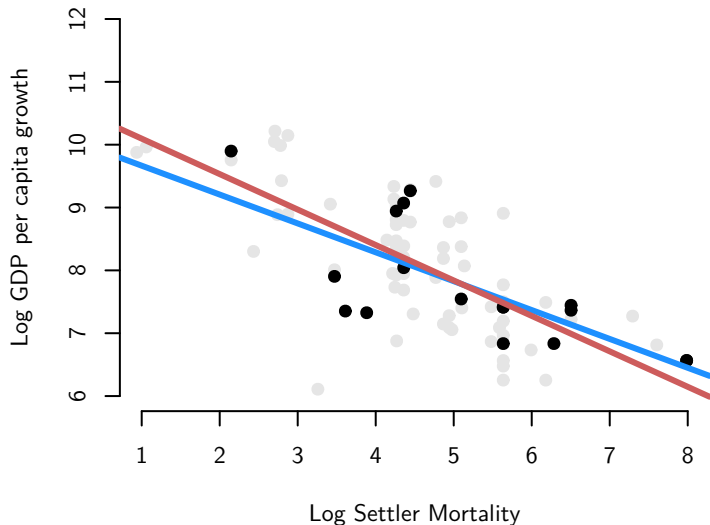
Simulation procedure

- 1 Draw a random sample of size $n = 30$ with replacement using `sample()`
- 2 Use `lm()` to calculate the OLS estimates of the slope and intercept
- 3 Plot the estimated regression line

Population Regression



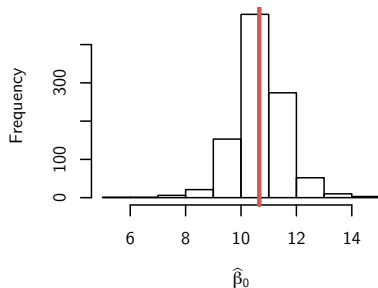
Randomly sample from AJR



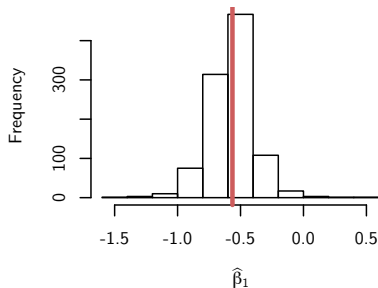
Sampling distribution of OLS

- You can see that the estimated slopes and intercepts vary from sample to sample, but that the “average” of the lines looks about right.

Sampling distribution of intercepts

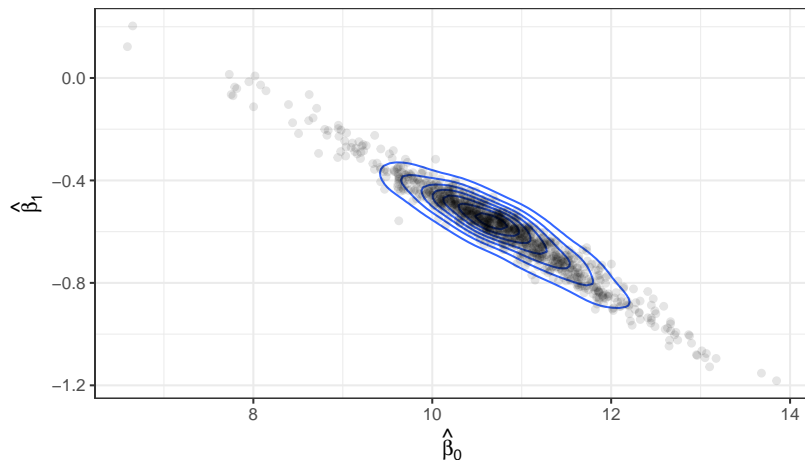


Sampling distribution of slopes



The Sampling Distribution is a Joint Distribution!

While both the intercept and the slope vary, they vary together.



Sample Mean Properties Review

- In the last few weeks we derived the properties of the sampling distribution for the sample mean, \bar{X}_n .
- Under essentially only the **iid assumption** (plus finite mean and variance) we derived the large sample distribution as

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- ▶ This means the estimator is unbiased for the population mean:
 $E[\bar{X}_n] = \mu$.
- ▶ has sampling variance: σ^2/n
- ▶ and standard error: σ/\sqrt{n}
- This in turn gave us confidence intervals and hypothesis tests.
- We will use the same strategy here!

Our goal

- What is the sampling distribution of the OLS slope?

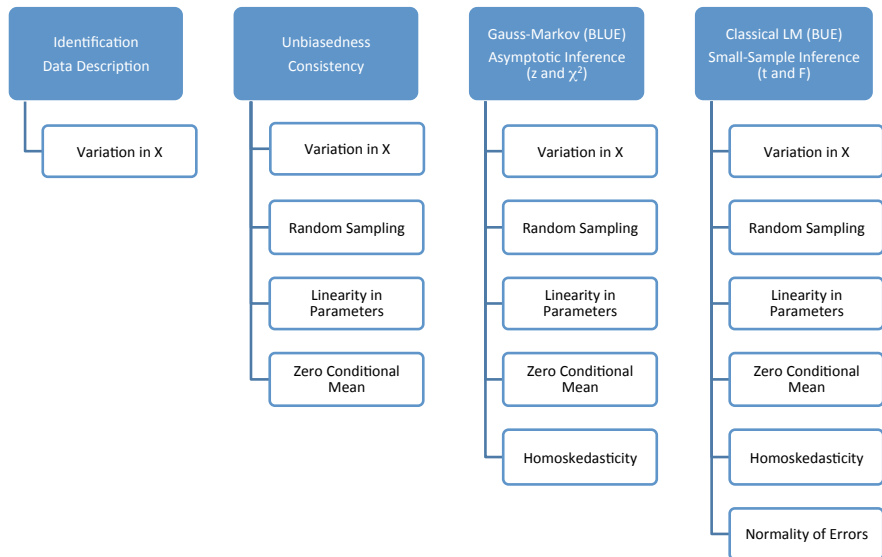
$$\hat{\beta}_1 \sim ?(?, ?)$$

- We need fill in those ?s.
- We'll start with the mean of the sampling distribution. Is the estimator centered at the true value, β_1 ?

Classical Model: OLS Assumptions Preview

- 1 **Linearity in Parameters:** The population model is linear in its parameters and correctly specified
- 2 **Random Sampling:** The observed data represent a random sample from the population described by the model.
- 3 **Variation in X :** There is variation in the explanatory variable.
- 4 **Zero conditional mean:** Expected value of the error term is zero conditional on all values of the explanatory variable
- 5 **Homoskedasticity:** The error term has the same variance conditional on all values of the explanatory variable.
- 6 **Normality:** The error term is independent of the explanatory variables and normally distributed.

Hierarchy of OLS Assumptions



OLS Assumption I

Assumption (I. Linearity in Parameters)

The population regression model is linear in its parameters and correctly specified as:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- Note that it can be nonlinear *in variables*
 - ▶ OK: $Y_i = \beta_0 + \beta_1 X_i + u_i$ or
 $Y_i = \beta_0 + \beta_1 X_i^2 + u_i$ or
 $Y_i = \beta_0 + \beta_1 \log(X_i) + u$
 - ▶ Not OK: $Y_i = \beta_0 + \beta_1^2 X_i + u_i$ or
 $Y_i = \beta_0 + \exp(\beta_1) X_i + u_i$
- β_0, β_1 : Population **parameters** — fixed and unknown
- u_i : Unobserved random variable with $E[u_i] = 0$ — captures all other factors influencing Y_i other than X_i
- We assume this to be the structural model, i.e., the model describing the true process generating Y_i

OLS Assumption II

Assumption (II. Random Sampling)

The observed data:

$$(y_i, x_i) \text{ for } i = 1, \dots, n$$

represent an i.i.d. random sample of size n following the population model.

Data examples consistent with this assumption:

- A cross-sectional survey where the units are sampled randomly

Potential Violations:

- Time series data (regressor values may exhibit persistence)
- Sample selection problems (sample not representative of the population)

OLS Assumption III

Assumption (III. Variation in X ; a.k.a. No Perfect Collinearity)

The observed data:

$$x_i \text{ for } i = 1, \dots, n$$

are not all the same value.

Satisfied as long as there is some variation in the regressor X in the sample.

Why do we need this?

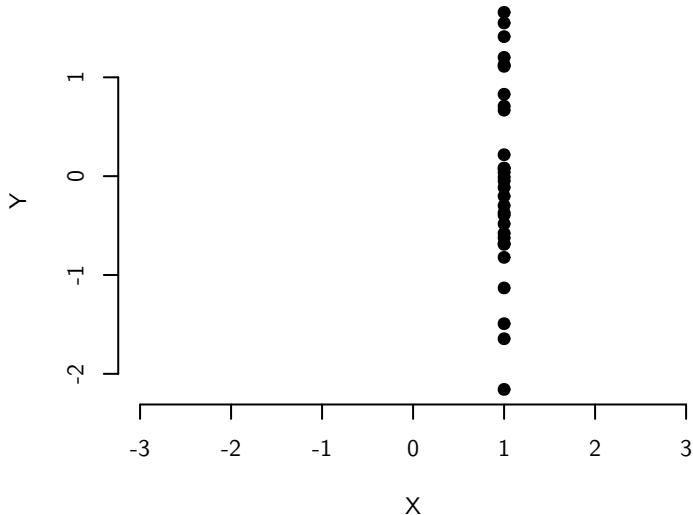
$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

This assumption is needed just to calculate $\hat{\beta}$.

Only assumption needed for using OLS as a pure data summary.

Stuck in a moment

- Why does this matter? How would you draw the line of best fit through this scatterplot, which is a violation of this assumption?



OLS Assumption IV

Assumption (IV. Zero Conditional Mean)

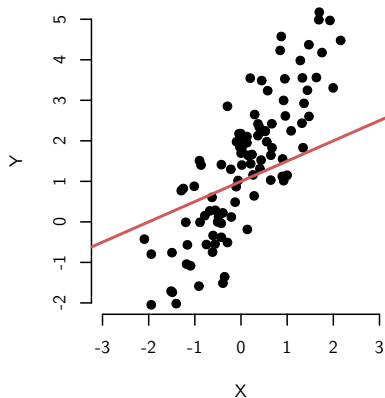
The expected value of the error term is zero conditional on any value of the explanatory variable:

$$E[u_i | X_i = x] = 0$$

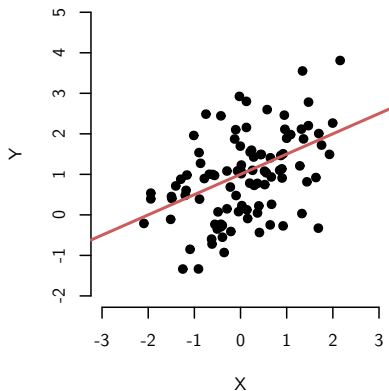
- $E[u_i | X] = 0$ implies a slightly weaker condition $\text{Cov}(X, u) = 0$
- Given random sampling, $E[u | X] = 0$ also implies $E[u_i | x_i] = 0$ for all i

Violating the zero conditional mean assumption

Assumption 4 violated

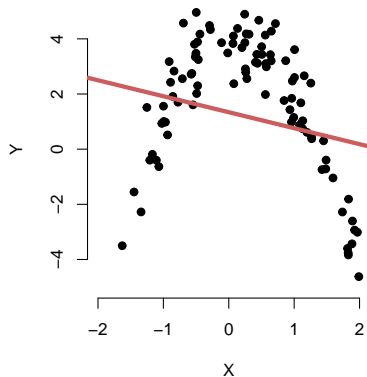


Assumption 4 not violated

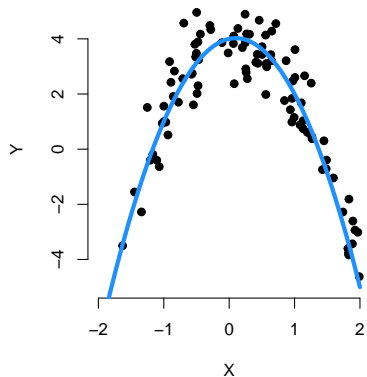


Violating the zero conditional mean assumption

Assumption 4 Violated



Assumption 4 Not Violated



Unbiasedness

With Assumptions 1-4, we can show that the OLS estimator for the slope is unbiased, that is $E[\hat{\beta}_1] = \beta_1$.

Let's prove it!

Lemma 3: Weighted Combinations of X_i

Lemma ($\sum_i W_i X_i = 1$)

$$\begin{aligned}\sum_{i=1}^n W_i X_i &= \sum_{i=1}^n \frac{X_i(X_i - \bar{X})}{\sum_{j=1}^n (X_j - \bar{X})^2} \\ &= \frac{1}{\sum_{j=1}^n (X_j - \bar{X})^2} \sum_{i=1}^n X_i(X_i - \bar{X}) \\ &= \frac{1}{\sum_{j=1}^n (X_j - \bar{X})^2} \left[\sum_{i=1}^n X_i(X_i - \bar{X}) - \sum_{i=1}^n \bar{X}(X_i - \bar{X}) \right] \\ &= \frac{1}{\sum_{j=1}^n (X_j - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X}) \\ &= 1\end{aligned}$$

Lemma 4: Estimation Error

Lemma

$$\begin{aligned}\hat{\beta}_1 &= \sum_{i=1}^n W_i Y_i \\ &= \sum_{i=1}^n W_i (\beta_0 + \beta_1 X_i + u_i) \\ &= \beta_0 \left(\sum_{i=1}^n W_i \right) + \beta_1 \left(\sum_{i=1}^n W_i X_i \right) + \sum_{i=1}^n W_i u_i \\ &= \beta_1 + \sum_{i=1}^n W_i u_i \\ \hat{\beta}_1 - \beta_1 &= \sum_{i=1}^n W_i u_i\end{aligned}$$

Unbiasedness Proof

$$\begin{aligned} E[\hat{\beta}_1 - \beta_1 | X] &= E \left[\sum_{i=1}^n W_i u_i | X \right] \\ &= \sum_{i=1}^n E[W_i u_i | X] \\ &= \sum_{i=1}^n W_i E[u_i | X] \\ &= \sum_{i=1}^n W_i 0 \\ &= 0 \end{aligned}$$

Using iterated expectations we can show that it is also unconditionally biased $E[\hat{\beta}_1] = E[E[\hat{\beta}_1 | X]] = E[\beta_1] = \beta_1$.

Consistency

- Recall the estimation error,

$$\hat{\beta}_1 = \beta_1 + \sum_{i=1}^n W_i u_i$$

- Under iid sampling we have

$$\sum_{i=1}^n W_i u_i = \frac{\sum_{i=1}^n (X_i - \bar{X}) u_i}{\sum_{i=1}^n (X_i - \bar{X})} \xrightarrow{P} \frac{\text{Cov}[X_i, u_i]}{V[X_i]}$$

- Under A4 (zero conditional mean error) we have the slightly weaker property $\text{Cov}[X_i, u_i] = 0$ so as long as $V[X] > 0$, then we have,

$$\hat{\beta}_1 \xrightarrow{P} \beta_1$$

We Covered

- The first four assumptions of the classical model
- We showed that these four were sufficient to establish unbiasedness and consistency.
- We even proved it to ourselves!

Next Time: The Classical Perspective Part 2: Variance.

Where We've Been and Where We're Going...

- Last Week
 - ▶ hypothesis testing
 - ▶ what is regression
- This Week
 - ▶ mechanics and properties of simple linear regression
 - ▶ inference and measures of model fit
 - ▶ confidence intervals for regression
 - ▶ goodness of fit
- Next Week
 - ▶ mechanics with two regressors
 - ▶ omitted variables, multicollinearity
- Long Run
 - ▶ probability → inference → regression → causal inference

- 1 Mechanics of OLS
- 2 Classical Perspective (Part 1, Unbiasedness)
 - Sampling Distributions
 - Classical Assumptions 1–4
- 3 Classical Perspective: Variance
 - Sampling Variance
 - Gauss-Markov
 - Large Samples
 - Small Samples
 - Agnostic Perspective
- 4 Inference
 - Hypothesis Tests
 - Confidence Intervals
 - Goodness of fit
 - Interpretation
- 5 Non-linearities
 - Log Transformations
 - Fun With Logs
 - LOESS

Where are we?

- Now we know that, under Assumptions 1-4, we know that

$$\hat{\beta}_1 \sim ?(\beta_1, ?)$$

- That is we know that the sampling distribution is **centered on the true population slope**, but we don't know the population variance.

Sampling variance of estimated slope

- In order to derive the sampling variance of the OLS estimator,
 - 1 Linearity
 - 2 Random (iid) sample
 - 3 Variation in X_i
 - 4 Zero conditional mean of the errors
 - 5 Homoskedasticity

Variance of OLS Estimators

How can we derive $\text{Var}[\hat{\beta}_0]$ and $\text{Var}[\hat{\beta}_1]$? Let's make the following additional assumption:

Assumption (V. Homoskedasticity)

The conditional variance of the error term is constant and does not vary as a function of the explanatory variable:

$$\text{Var}[u|X] = \sigma_u^2$$

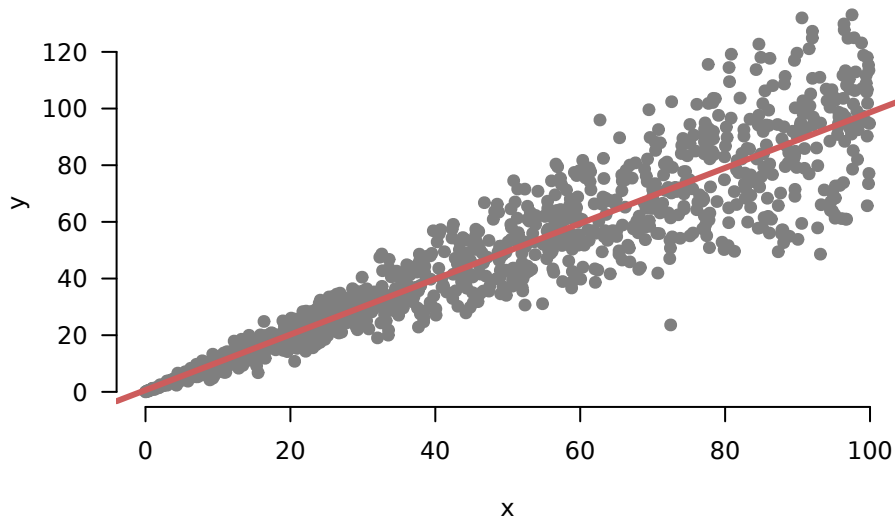
- This implies $\text{Var}[u] = \sigma_u^2$
→ all errors have an identical **error variance** ($\sigma_{u_i}^2 = \sigma_u^2$ for all i)
- Taken together, Assumptions I–V imply:

$$E[Y|X] = \beta_0 + \beta_1 X$$

$$\text{Var}[Y|X] = \sigma_u^2$$

- Violation: $\text{Var}[u|X = x_1] \neq \text{Var}[u|X = x_2]$ called **heteroskedasticity**.
- Assumptions I–V are collectively known as the **Gauss-Markov assumptions**

Heteroskedasticity



Deriving the sampling variance

$$V[\hat{\beta}_1 | \mathbf{X}] = ??$$

$$\begin{aligned} V[\hat{\beta}_1 | \mathbf{X}] &= V \left[\sum_{i=1}^n W_i u_i \mid \mathbf{X} \right] \\ &= \sum_{i=1}^n W_i^2 V[u_i | \mathbf{X}] && \text{(A2: iid)} \\ &= \sum_{i=1}^n W_i^2 \sigma_u^2 && \text{(A5: homoskedastic)} \\ &= \sigma_u^2 \sum_{i=1}^n \left(\frac{(X_i - \bar{X})}{\sum_{i'=1}^n (X_{i'} - \bar{X})^2} \right)^2 \\ &= \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \end{aligned}$$

Variance of OLS Estimators

Theorem (Variance of OLS Estimators)

Given OLS Assumptions I–V (Gauss-Markov Assumptions):

$$V[\hat{\beta}_1 | X] = \frac{\sigma_u^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$V[\hat{\beta}_0 | X] = \sigma_u^2 \left\{ \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right\}$$

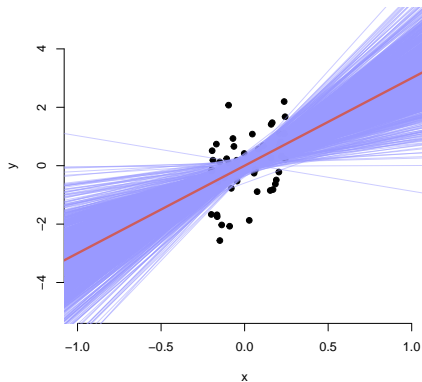
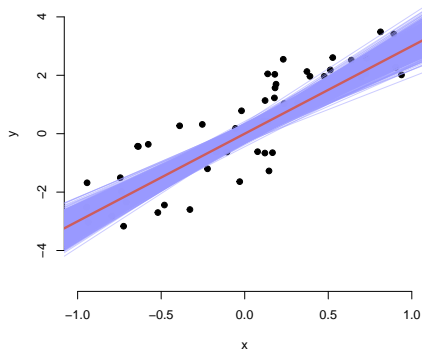
where $V[u | X] = \sigma_u^2$ (the error variance).

Understanding the sampling variance

$$V[\hat{\beta}_1 | X_1, \dots, X_n] = \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

- What drives the sampling variability of the OLS estimator?
 - ▶ The higher the variance of $Y_i | X_i$, the higher the sampling variance
 - ▶ The lower the variance of X_i , the higher the sampling variance
 - ▶ As we increase n , the denominator gets large, while the numerator is fixed and so the sampling variance shrinks to 0.

Variance in X Reduces Standard Errors



Estimating the Variance of OLS Estimators

How can we estimate the unobserved error variance $\text{Var}[u] = \sigma_u^2$?

We can derive an estimator based on the **residuals**:

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

Recall: The **errors** u_i are NOT the same as the residuals \hat{u}_i .

Intuitively, the scatter of the residuals around the fitted regression line should reflect the unseen scatter about the true population regression line.

We can measure scatter with the mean squared deviation:

$$MSD(\hat{u}) \equiv \frac{1}{n} \sum_{i=1}^n (\hat{u}_i - \bar{\hat{u}})^2 = \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2$$

Estimating the Variance of OLS Estimators

- By construction, the regression line is closer since it is drawn to fit the sample we observe
- Specifically, the regression line is drawn so as to minimize the sum of the squares of the distances between it and the observations
- So the spread of the residuals $MSD(\hat{u})$ will slightly *underestimate* the error variance $\text{Var}[u] = \sigma_u^2$ on average
- In fact, we can show that with a single regressor X we have:

$$E[MSD(\hat{u})] = \frac{n-2}{n} \sigma_u^2 \text{ (degrees of freedom adjustment)}$$

- Thus, an **unbiased estimator** for the error variance is:

$$\hat{\sigma}_u^2 = \frac{n}{n-2} MSD(\hat{u}) = \frac{n}{n-2} \frac{1}{n} \sum_{i=1}^n \hat{u}_i^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$$

We plug this estimate into the variance estimators for $\hat{\beta}_0$ and $\hat{\beta}_1$.

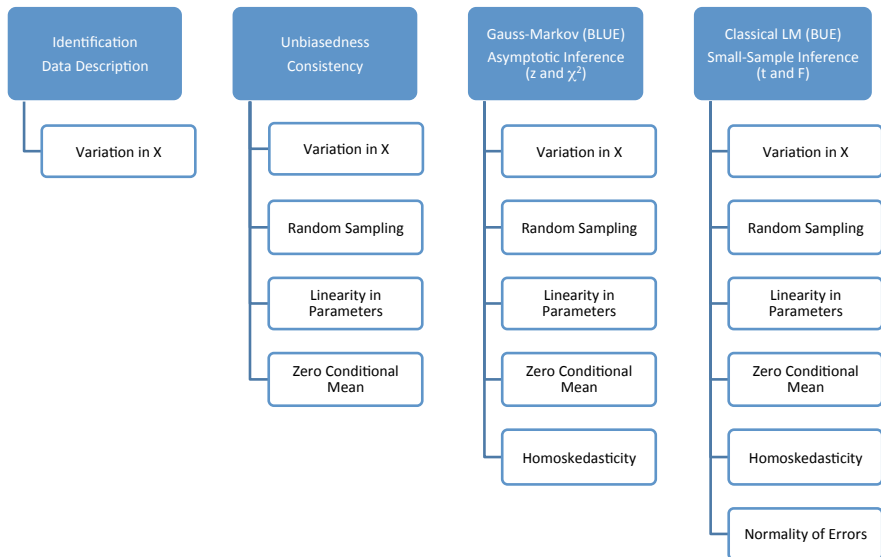
Where are we?

- Under Assumptions 1-5, we know that

$$\hat{\beta}_1 \sim? \left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

- Now we know the mean and sampling variance of the sampling distribution.
- How does this compare to other estimators for the population slope?

Where are we?



OLS is BLUE :(

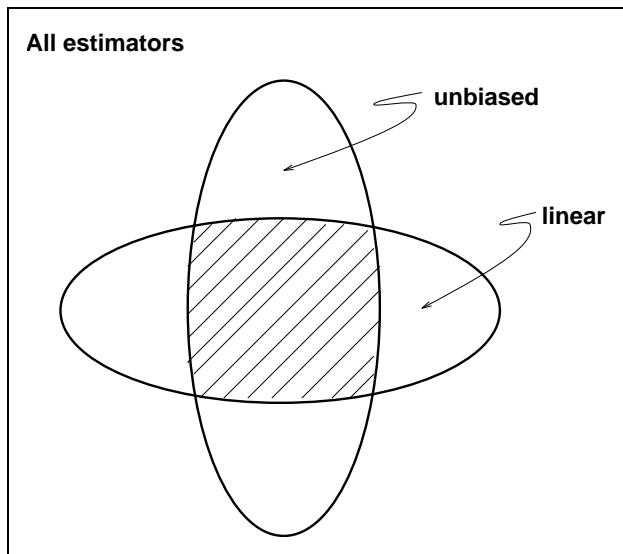
Theorem (Gauss-Markov)

Given OLS Assumptions I–V, the OLS estimator is **BLUE**, i.e. the

- 1 **B**est: Lowest variance in class
- 2 **L**inear: Among Linear estimators
- 3 **U**nbiased: Among Linear Unbiased estimators
- 4 **E**stimator.

- A **linear** estimator is one that can be written as $\hat{\beta} = \mathbf{W}y$
- Assumptions 1-5 are called the “Gauss Markov Assumptions”
- Result fails to hold when the assumptions are violated!

Gauss-Markov Theorem



Where are we?

- Under Assumptions 1-5, we know that

$$\hat{\beta}_1 \sim? \left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2} \right)$$

- And we know that $\frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}$ is the lowest variance of any linear estimator of β_1
- What about the last question mark? What's the form of the distribution?

Large-sample distribution of OLS estimators

- Remember that the OLS estimator is the sum of independent r.v.'s:

$$\hat{\beta}_1 = \sum_{i=1}^n W_i Y_i$$

- Mantra of the Central Limit Theorem:
“the sums and means of random variables tend to be Normally distributed in large samples.”
- True here as well, so we know that in large samples:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1)$$

- Can also replace SE with an estimate:

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \sim N(0, 1)$$

Where are we?

Under Assumptions 1-5 and in large samples, we know that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$



Sampling distribution in small samples

- What if we have a small sample? What can we do then?
- Can't get something for nothing, but we can make progress if we make another assumption:
 - 1 Linearity
 - 2 Random (iid) sample
 - 3 Variation in X_i
 - 4 Zero conditional mean of the errors
 - 5 Homoskedasticity
 - 6 **Errors are conditionally Normal**

OLS Assumptions VI

Assumption (VI. Normality)

The population error term is independent of the explanatory variable, $u \perp\!\!\!\perp X$, and is normally distributed with mean zero and variance σ_u^2 :

$$u \sim N(0, \sigma_u^2), \text{ which implies } Y|X \sim N(\beta_0 + \beta_1 X, \sigma_u^2)$$

Note: This also implies homoskedasticity and zero conditional mean.

- Together Assumptions I–VI are the **classical linear model (CLM) assumptions**.
- The CLM assumptions imply that OLS is **BUE** (i.e. minimum variance among all linear or non-linear unbiased estimators)
- Non-normality of the errors is a serious concern in small samples. We can *partially* check this assumption by looking at the residuals (more in coming weeks)
- Variable transformations can help to come closer to normality
- Reminder: we don't need normality assumption in large samples

Sampling distribution of OLS slope

- If we have Y_i given X_i is distributed $N(\beta_0 + \beta_1 X_i, \sigma_u^2)$, then we have the following at any sample size:

$$\frac{\hat{\beta}_1 - \beta_1}{SE[\hat{\beta}_1]} \sim N(0, 1)$$

- Furthermore, if we replace the true standard error with the estimated standard error, then we get the following:

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \sim t_{n-2}$$

- The standardized coefficient follows a t distribution $n - 2$ degrees of freedom. We take off an extra degree of freedom because we had to estimate one more parameter than just the sample mean.
- All of this depends on Normal errors!

Where are we?

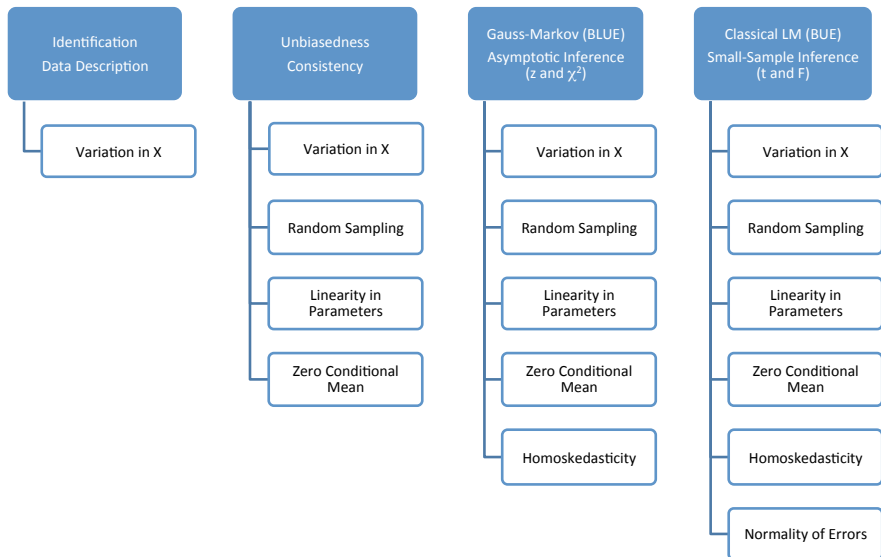
- Under Assumptions 1-5 and in large samples, we know that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

- Under Assumptions 1-6 and in any sample, we know that

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \sim t_{n-2}$$

Hierarchy of OLS Assumptions



Regression as parametric modeling

Let's summarize the parametric view we have taken thus far.

- Gauss-Markov assumptions:
 - ▶ (A1) linearity, (A2) i.i.d. sample, (A3) variation in X , (A4) zero conditional mean error, (A5) homoskedasticity.
 - ▶ basically, **assume the model is right**
- \rightsquigarrow OLS is BLUE, plus (A6) normality of the errors and we get small sample SEs and BUE.
- What is the basic approach here?

- ▶ A1 defines a linear model for the outcome:

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

- ▶ A2 and A4 let us write the CEF as function of X_i alone.

$$E[Y_i|X_i] = \mu_i = \beta_0 + \beta_1 X_i$$

- ▶ A5-6, define a probabilistic model for the conditional distribution:

$$Y_i|X_i \sim \mathcal{N}(\mu_i, \sigma^2)$$

- ▶ A3 covers the edge-case that the β s are indistinguishable.

Agnostic views on regression

- These assumptions assume we know **a lot** about how Y_i is 'generated'.
- Justifications for using OLS (like BLUE/BUE) often invoke these assumptions which are unlikely to hold exactly.
- Alternative: take an **agnostic** view on regression.
 - ▶ use OLS without believing these assumptions.
 - ▶ lean on two things: **A2 i.i.d. sample**, **asymptotics** (large-sample properties)
- Lose the distributional assumptions and focus on approximating the best linear predictor.
- If the true CEF happens to be linear, the best linear predictor is it.

Unbiasedness Result

- One of the results most people know is that OLS is unbiased, but **unbiased for what?**
- It is unbiased for the CEF under the assumption that the model is correctly specified.
- However, this could be a quite poor approximation to the true CEF if there is a great deal of non-linearity.
- We will often use OLS as a means to approximate the CEF, but don't forget that it is just an **approximation!**
- We will return in a few weeks to how you diagnose this approximation.

Pedagogical Note

- For now we are going to move forward with the **classical worldview** and we will return to some alternative approaches later in the semester once we are comfortable with the matrix representation of regression.
- This will lead to techniques like **robust standard errors** which don't rely on the assumptions of homoskedasticity (but have other tradeoffs!)
- For now, just remember that regression is a **linear approximation** to the CEF!

We Covered

- Sampling Variance
- Gauss Markov
- Large Sample and Small Sample Properties

Next Time: Inference

Where We've Been and Where We're Going...

- Last Week
 - ▶ hypothesis testing
 - ▶ what is regression
- This Week
 - ▶ mechanics and properties of simple linear regression
 - ▶ inference and measures of model fit
 - ▶ confidence intervals for regression
 - ▶ goodness of fit
- Next Week
 - ▶ mechanics with two regressors
 - ▶ omitted variables, multicollinearity
- Long Run
 - ▶ probability → inference → regression → causal inference

- 1 Mechanics of OLS
- 2 Classical Perspective (Part 1, Unbiasedness)
 - Sampling Distributions
 - Classical Assumptions 1–4
- 3 Classical Perspective: Variance
 - Sampling Variance
 - Gauss-Markov
 - Large Samples
 - Small Samples
 - Agnostic Perspective
- 4 Inference
 - Hypothesis Tests
 - Confidence Intervals
 - Goodness of fit
 - Interpretation
- 5 Non-linearities
 - Log Transformations
 - Fun With Logs
 - LOESS

Where are we?

- Under Assumptions 1-5 and in large samples, we know that

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

- Under Assumptions 1-6 and in any sample, we know that

$$\frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \sim t_{n-2}$$

Null and alternative hypotheses review

- Null: $H_0 : \beta_1 = 0$
 - ▶ The null is the straw man we want to knock down.
 - ▶ With regression, almost always null of no relationship
- Alternative: $H_a : \beta_1 \neq 0$
 - ▶ Claim we want to test
 - ▶ Could do one-sided test, but you shouldn't
- Notice these are statements about the population parameters, not the OLS estimates.

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

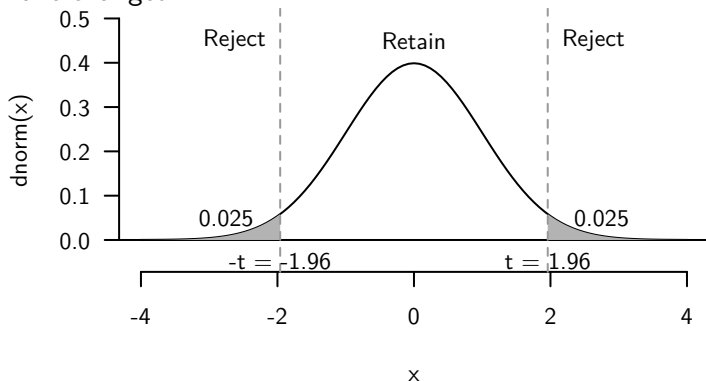
- Under the null hypothesis:
 - ▶ large samples: $T \sim \mathcal{N}(0, 1)$
 - ▶ any size sample with normal errors: $T \sim t_{n-2}$
 - ▶ conservative to use t_{n-2} anyways since t_{n-2} is approximately normal in large samples.
- Thus, under the null, we know the distribution of T and can use that to formulate a rejection region and calculate p-values.
- By default, R shows you the test statistic for $\beta_1 = 0$ and uses the t distribution.

Rejection region

- Choose a level of the test, α , and find rejection regions that correspond to that value under the null distribution:

$$\mathbb{P}(-t_{\alpha/2, n-2} < T < t_{\alpha/2, n-2}) = 1 - \alpha$$

- This is exactly the same as with sample means and sample differences in means, except that the degrees of freedom on the t distribution have changed.



p-value

- The interpretation of the p-value is the same: the probability of seeing a test statistic at least this extreme if the null hypothesis were true
- Mathematically:

$$\mathbb{P} \left(\left| \frac{\hat{\beta}_1 - c}{\widehat{SE}[\hat{\beta}_1]} \right| \geq |T_{obs}| \right)$$

- If the p-value is less than α we would reject the null at the α level.

Confidence intervals

- Very similar to the approach with sample means. By the sampling distribution of the OLS estimator, we know that we can find t -values such that:

$$\mathbb{P}\left(-t_{\alpha/2, n-2} \leq \frac{\hat{\beta}_1 - \beta_1}{\widehat{SE}[\hat{\beta}_1]} \leq t_{\alpha/2, n-2}\right) = 1 - \alpha$$

- If we rearrange this as before, we can get an expression for confidence intervals:

$$\mathbb{P}\left(\hat{\beta}_1 - t_{\alpha/2, n-2}\widehat{SE}[\hat{\beta}_1] \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2, n-2}\widehat{SE}[\hat{\beta}_1]\right) = 1 - \alpha$$

- Thus, we can write the confidence intervals as:

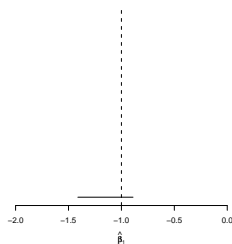
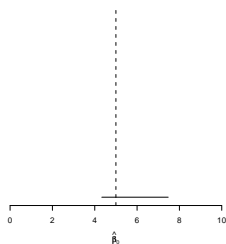
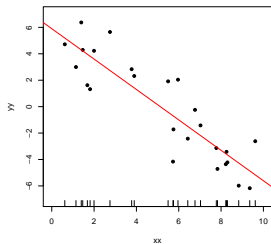
$$\hat{\beta}_1 \pm t_{\alpha/2, n-2}\widehat{SE}[\hat{\beta}_1]$$

- We can derive these for the intercept as well:

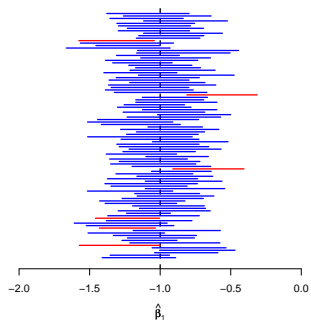
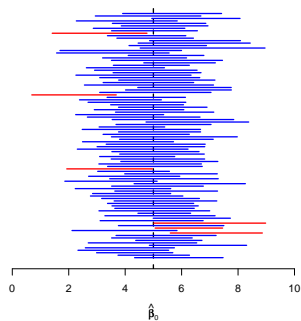
$$\hat{\beta}_0 \pm t_{\alpha/2, n-2}\widehat{SE}[\hat{\beta}_0]$$

CI Simulation Example

Returning to our simulation example we can simulate the sampling distributions of the 95 % confidence interval estimates for $\hat{\beta}_1$ and $\hat{\beta}_0$



CI Simulation Example



Prediction error

- How do we judge how well a line fits the data?
- One way is to find out how much better we do at predicting Y once we include X into the regression model.
- Prediction errors without X : best prediction is the mean, so our squared errors, or the **total sum of squares** (SS_{tot}) would be:

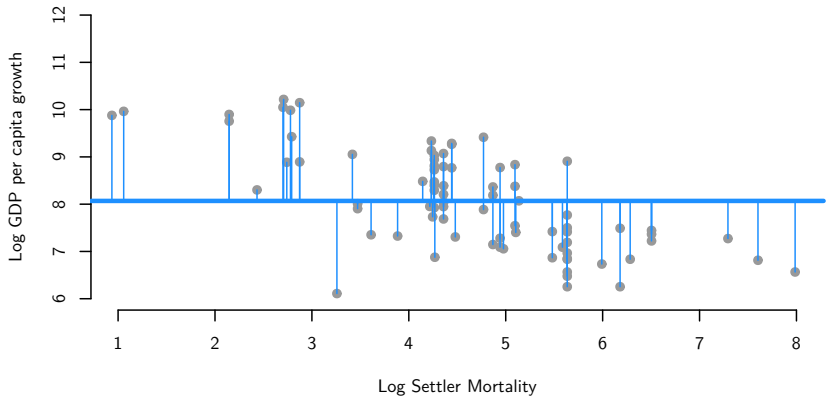
$$SS_{tot} = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res} :

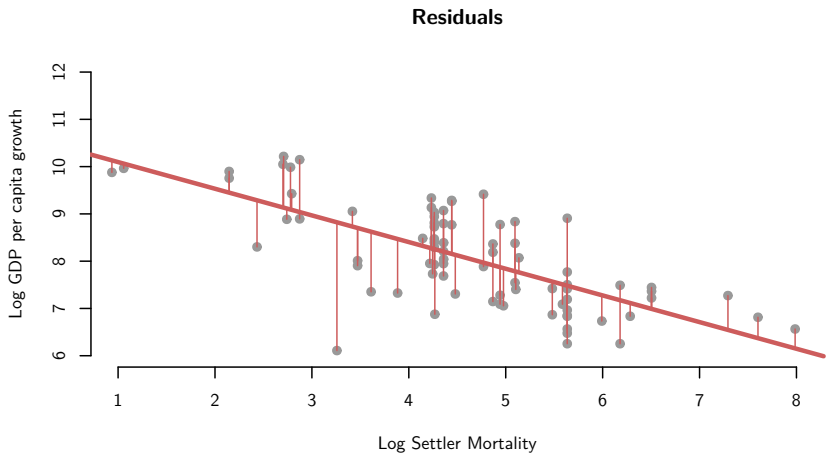
$$SS_{res} = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

Sum of Squares

Total Prediction Errors



Sum of Squares



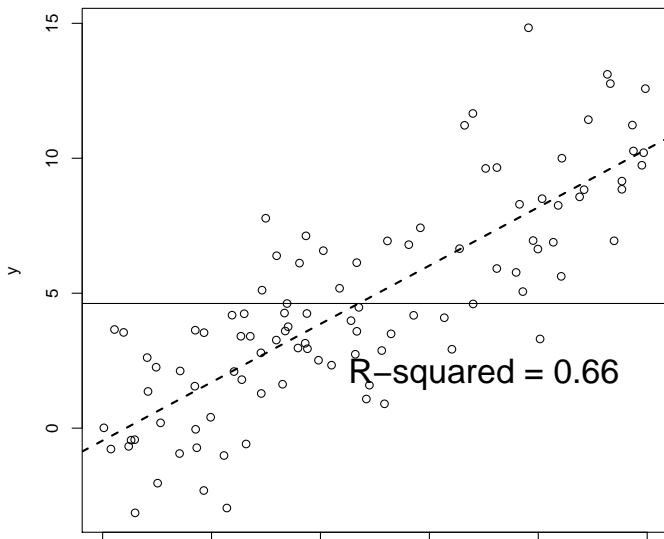
R-square

- By definition, the residuals have to be smaller than the deviations from the mean, so we might ask the following: how much lower is the SS_{res} compared to the SS_{tot} ?
- We quantify this question with the **coefficient of determination** or R^2 . This is the following:

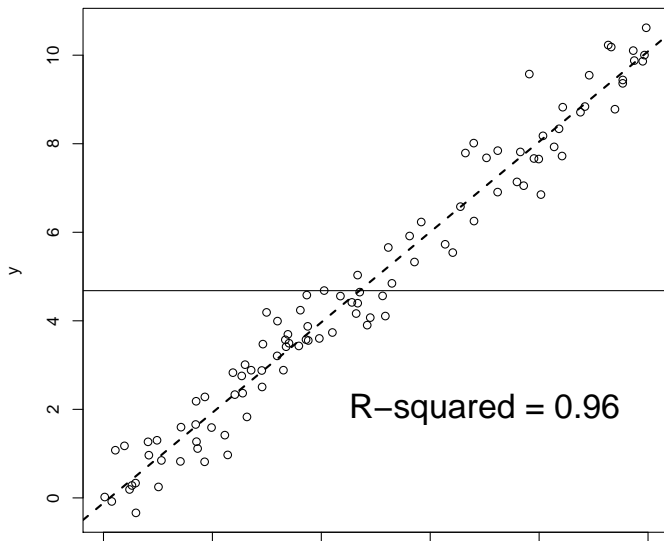
$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on X .
- Alternatively, this is the fraction of the variation in Y is “explained by” X .
- $R^2 = 0$ means no relationship
- $R^2 = 1$ implies perfect linear fit

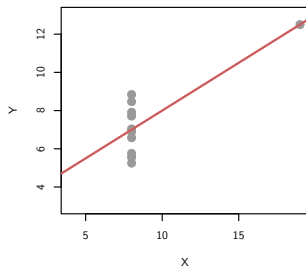
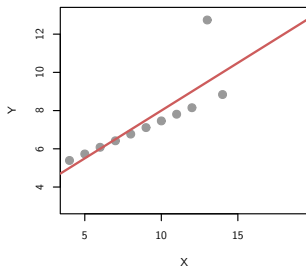
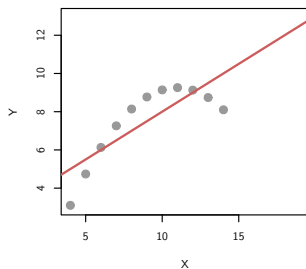
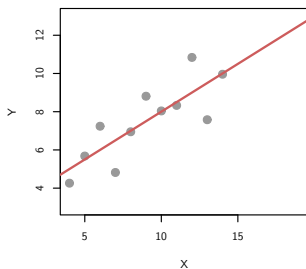
Is R-squared useful?



Is R-squared useful?

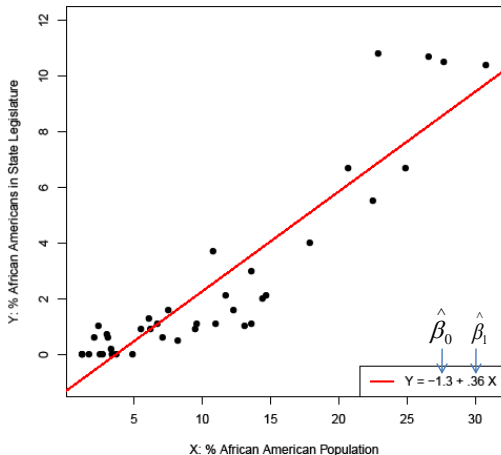


Is R-squared useful?



Interpreting a Regression

Let's have a quick chat about interpretation.



State Legislators and African American Population

Interpretations of increasing quality:

```
> summary(lm(beo ~ bpop, data = D))
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	-1.31489	0.32775	-4.012	0.000264	***
bpop	0.35848	0.02519	14.232	< 2e-16	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.317 on 39 degrees of freedom

Multiple R-squared: 0.8385, Adjusted R-squared: 0.8344

F-statistic: 202.6 on 1 and 39 DF, p-value: < 2.2e-16

“In states where an additional .01 proportion of the population is African American, we observe on average .035 proportion more African American state legislators (between .03 and .04 with 95% confidence).”

(still not perfect, the best will be subject matter specific. is fairly clear it is non-causal, gives uncertainty.)

Ground Rules: Interpretation of the Slope

I almost didn't include the last example in the slides. It is **hard** to give ground rules that cover all cases. Regressions are a part of marshaling evidence in an argument which makes them naturally specific to context.

- ① Give a short, but precise interpretation of the association using interpretable **language** and **units**
- ② If the association has a **causal** interpretation explain why, otherwise do not imply a causal interpretation.
- ③ Provide a meaningful sense of **uncertainty**
- ④ Indicate the **practical** significance of the finding for your argument.

Goal Check: Understand `lm()` Output

Call:

```
lm(formula = sr ~ pop15, data = LifeCycleSavings)
```

Residuals:

Min	1Q	Median	3Q	Max
-8.637	-2.374	0.349	2.022	11.155

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	17.49660	2.27972	7.675	6.85e-10	***
pop15	-0.22302	0.06291	-3.545	0.000887	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.03 on 48 degrees of freedom

Multiple R-squared: 0.2075, Adjusted R-squared: 0.191

F-statistic: 12.57 on 1 and 48 DF, p-value: 0.0008866

We Covered

- Hypothesis tests
- Confidence intervals
- Goodness of fit measures

Next Time: Non-linearities

Where We've Been and Where We're Going...

- Last Week
 - ▶ hypothesis testing
 - ▶ what is regression
- This Week
 - ▶ mechanics and properties of simple linear regression
 - ▶ inference and measures of model fit
 - ▶ confidence intervals for regression
 - ▶ goodness of fit
- Next Week
 - ▶ mechanics with two regressors
 - ▶ omitted variables, multicollinearity
- Long Run
 - ▶ probability → inference → regression → causal inference

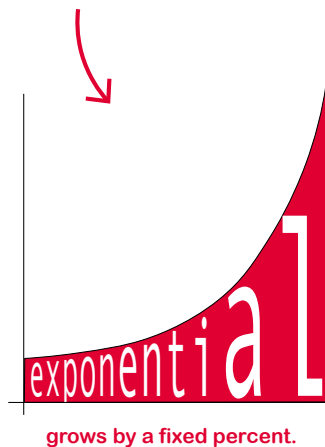
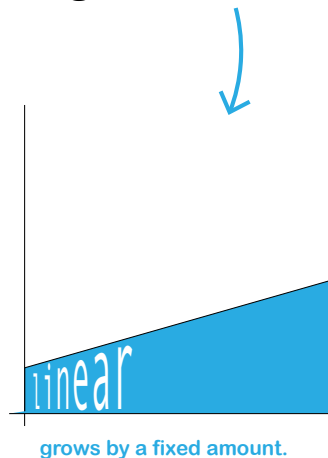
- 1 Mechanics of OLS
- 2 Classical Perspective (Part 1, Unbiasedness)
 - Sampling Distributions
 - Classical Assumptions 1–4
- 3 Classical Perspective: Variance
 - Sampling Variance
 - Gauss-Markov
 - Large Samples
 - Small Samples
 - Agnostic Perspective
- 4 Inference
 - Hypothesis Tests
 - Confidence Intervals
 - Goodness of fit
 - Interpretation
- 5 Non-linearities
 - Log Transformations
 - Fun With Logs
 - LOESS

Non-linear CEFs

- When we say that CEFs are linear with regression, we mean **linear in parameters** but by including transformations of our variables we can make non-linear shapes of pre-specified functional forms.
- Many of these **non-linear transformations** are made by creating multiple variables out of a single X and so will have to wait for future weeks.
- The function $\log(\cdot)$ is one common transformation that has only one parameter.
- This is particularly useful for **positive** and **right-skewed** variables.

Why does everyone keep logging stuff??

Logs **linearize** **exponential** growth.



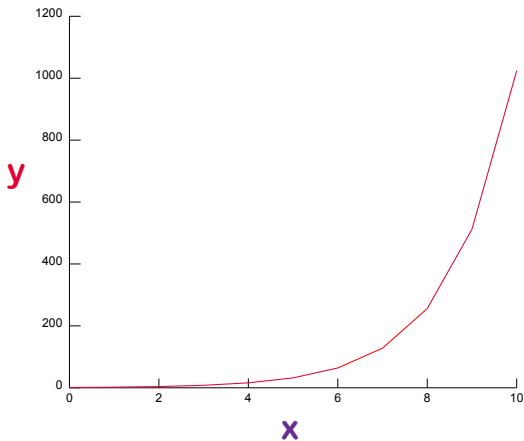
How? Let's look.

First, here's a graph showing **exponential growth**.

We're going to use $y = 2^x$, but any other exponent will work

$$x \quad y = (2^x)$$

0	1
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024



What happens when we take the log of y ?

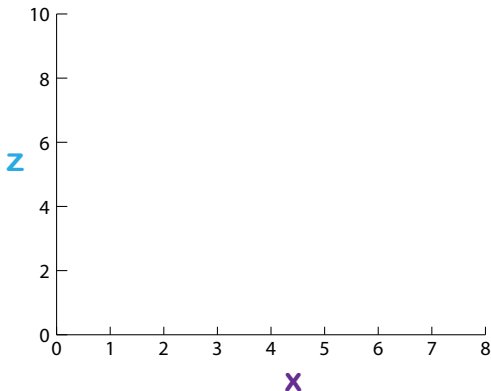
$$\log y = z$$

$$e^z = y$$

We're going to use $y = 2^x$, but any other exponent will work

x	y
0	1
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024

z



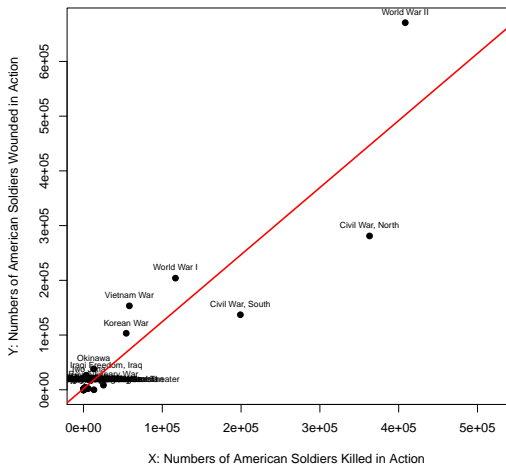
What happens when we take the log of v ?

Interpretation

The log transformation changes the interpretation of β_1 :

- Regress $\log(Y)$ on $X \rightarrow \beta_1$ approximates **percent increase** in our prediction of Y associated with one unit increase in X .
- Regress Y on $\log(X) \rightarrow \beta_1$ approximates increase in Y associated with a **percent increase** in X .
- Note that these approximations work only for small increments.
- In particular, they do not work when X is a discrete random variable.

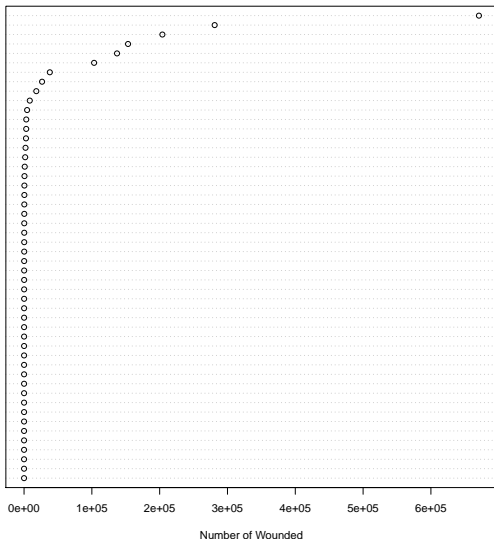
Example from the American War Library



$\hat{\beta}_1 = 1.23 \rightarrow$ One additional soldier killed predicts 1.23 additional soldiers wounded

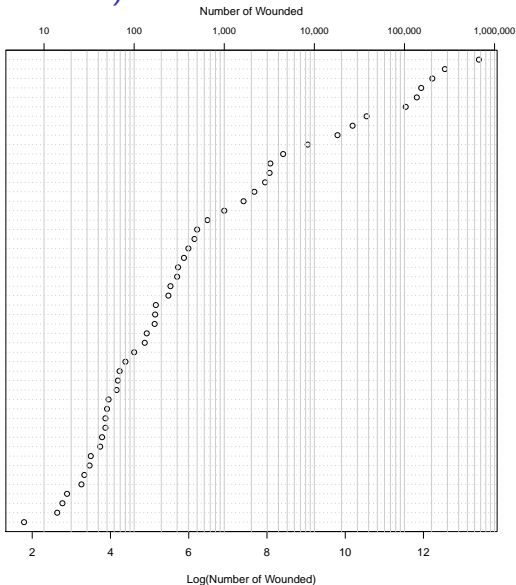
Wounded (Scale in Levels)

World War II
Civil War, North
World War I
Vietnam War
Civil War, South
Korean War
Okinawa
Operation Iraqi Freedom, Iraq
Iwo Jima
Revolutionary War
War of 1812
Aleutian Campaign
D-Day
Philippines War
Indian Wars
Spanish American War
Terrorism, World Trade Center
Yemen, USS Cole
Terrorism Khobar Towers, Saudi Arabia
Persian Gulf
Terrorism Oklahoma City
Persian Gulf, Op Desert Shield/Storm
Russia North Expedition
Moro Campaigns
China Boxer Rebellion
Panama
Dominican Republic
Israel Attack/USS Liberty
Lebanon
Texas War Of Independence
South Korea
Grenada
China Yangtze Service
Mexico
Nicaragua
Barbary Wars
Russia Siberia Expedition
Dominican Republic
China Civil War
Terrorism Riyadh, Saudi Arabia
North Atlantic Naval War
Franco-Amer Naval War
Operation Enduring Freedom, Afghanistan
Mexican War
Operation Enduring Freedom, Afghanistan Theater
Haiti
Texas Border Cortina War
Nicaragua
Italy Trieste
Japan

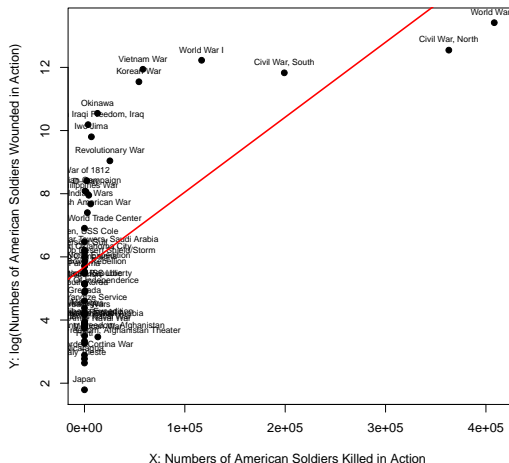


Wounded (Logarithmic Scale)

World War II
 Civil War, North
 World War I
 Vietnam War
 Civil War, South
 Korean War
 Okinawa
 Operation Iraqi Freedom, Iraq
 Iwo Jima
 Revolutionary War
 War of 1812
 Aleutian Campaign
 D-Day
 Philippines War
 Indian Wars
 Spanish American War
 Terrorism, World Trade Center
 Yemen, USS Cole
 Terrorism Khobar Towers, Saudi Arabia
 Persian Gulf
 Terrorism Oklahoma City
 Persian Gulf, Op Desert Shield/Storm
 Russia North Expedition
 Moro Campaigns
 China Boxer Rebellion
 Panama
 Dominican Republic
 Israel Attack/USS Liberty
 Lebanon
 Texas War Of Independence
 South Korea
 Grenada
 China Yangtze Service
 Mexico
 Nicaragua
 Barbary Wars
 Russia Siberia Expedition
 Dominican Republic
 China Civil War
 Terrorism Riyadh, Saudi Arabia
 North Atlantic Naval War
 Franco-Amer Naval War
 Operation Enduring Freedom, Afghanistan
 Mexican War
 Operation Enduring Freedom, Afghanistan Theater
 Haiti
 Texas Border Cortina War
 Nicaragua
 Italy Trieste
 Japan



Regression: Log-Level



$\hat{\beta}_1 = 0.0000237 \rightarrow$ One additional soldier killed predicts 0.0023 percent increase in the number of soldiers wounded

Four Most Commonly Used Models

Model	Equation	β_1 Interpretation
Level-Level	$Y = \beta_0 + \beta_1 X$	$\Delta Y = \beta_1 \Delta X$
Log-Level	$\log(Y) = \beta_0 + \beta_1 X$	$\% \Delta Y = 100 \beta_1 \Delta X$
Level-Log	$Y = \beta_0 + \beta_1 \log(X)$	$\Delta Y = (\beta_1 / 100) \% \Delta X$
Log-Log	$\log(Y) = \beta_0 + \beta_1 \log(X)$	$\% \Delta Y = \beta_1 \% \Delta X$

Why Does This Approximation Work?

A useful thing to know is that for small x ,

$$\log(1 + x) \approx x$$

$$\exp(x) \approx 1 + x$$

This can be derived from a series expansion of the log function. Numerically, when $|x| \leq .1$, the approximation is within 0.001.

Why Does This Approximation Work?

Take two numbers $a > b > 0$. The percentage difference between a and b is

$$p = 100 \left(\frac{a - b}{b} \right)$$

We can rewrite this as

$$\frac{a}{b} = 1 + \frac{p}{100}$$

Taking natural logs

$$\log(a) - \log(b) = \log \left(1 + \frac{p}{100} \right)$$

Applying our approximation and multiplying by 100 we find,

$$p \approx 100 (\log(a) - \log(b))$$

Be Careful: Log-Level with binary X

Assume we have: $\log(Y) = \beta_0 + \beta_1 X$ where X is binary with values 1 or 0. Assume $\beta_1 > .2$. What is the problem with saying that a one unit increase in X is associated with a $\beta_1 \cdot 100$ percent change in Y ?

Log approximation is inaccurate for large changes like going from $X = 0$ to $X = 1$. Instead the percent change in Y when X goes from 0 to 1 needs to be computed using:

$$\begin{aligned} 100(Y_{X=1} - Y_{X=0})/Y_{X=0} &= 100((Y_{X=1}/Y_{X=0}) - 1) \\ &= 100((\exp(\beta_1) - 1)) \\ &= 100(\exp(\beta_1) - 1) \end{aligned}$$

Recall: $\log(Y_{X=1}) - \log(Y_{X=0}) = \log(Y_{X=1}/Y_{X=0}) = \beta_1$.

A one unit change in X (ie. going from 0 to 1) is associated with a $100(\exp(\beta_1) - 1)$ percent increase in Y .

Interpreting a Logged Outcome


- On the last few slides, there was a bit that was a little dodgy.
- When we log the **outcome**, we are no longer approximating $E[Y|X]$ we are approximating $E[\log(Y)|X]$.
- Jensen's inequality gives us information on this relation:
 $f(E[X]) \leq E[f(X)]$ for any convex function $f()$.
- In practice, this means we are no longer characterizing the expectation of Y and it is technically inaccurate to talk about Y 'on average' changing in a certain way.
- What are we characterizing? The **geometric mean**.

Geometric Mean

$$\begin{aligned}\exp(E(\log(Y))) &= \exp\left(\frac{1}{N} \sum_{i=1}^N \log(Y_i)\right) \\ &= \exp\left(\frac{1}{N} \log\left(\prod_{i=1}^N Y_i\right)\right) \\ &= \exp\left(\log\left(\left(\prod_{i=1}^N Y_i\right)^{\frac{1}{N}}\right)\right) \\ &= \left(\prod_{i=1}^N Y_i\right)^{\frac{1}{N}} \\ &= \text{Geometric Mean}(Y)\end{aligned}$$

The **geometric mean** is a robust measure of central tendency.

THE INTERGENERATIONAL ELASTICITY OF WHAT? THE CASE FOR REDEFINING THE WORKHORSE MEASURE OF ECONOMIC MOBILITY

*Pablo A. Mitnik** 
*David B. Grusky**

Abstract

The intergenerational elasticity (IGE) has been assumed to refer to the expectation of children's income when in fact it pertains to the geometric mean of children's income. We show that mobility analyses based on the conventional IGE have been widely misinterpreted, are subject to selection bias, and cannot disentangle the different channels for transmitting economic status across generations. The solution to these problems—estimating the IGE of expected income or earnings—returns the field to what it has long meant to estimate. Under this approach, intergenerational persistence is found to be substantially higher, thus raising the possibility that the field's stock results are misleading.

Keywords

intergenerational economic mobility, elasticity of expected income, selection bias, gender, marriage and economic mobility

Core Idea

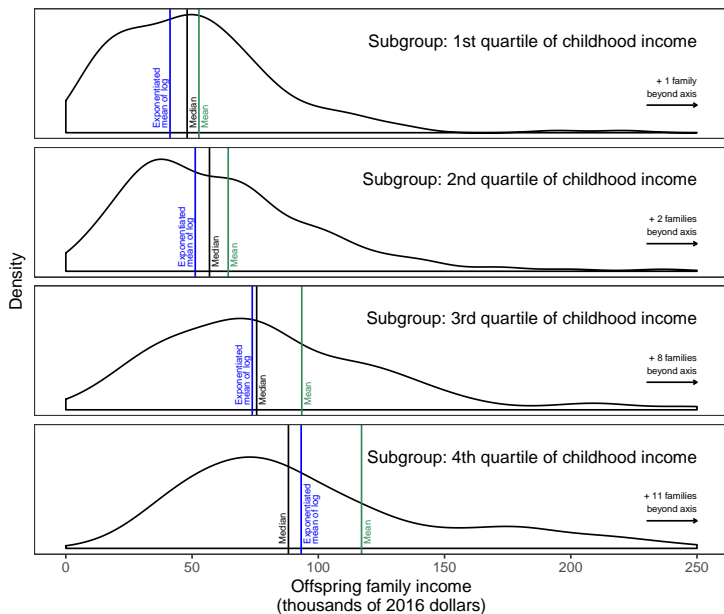
Classic approach :

$$\underbrace{E(\log(Y) | X)}_{\substack{\text{Mean of log} \\ \text{offspring income } Y \\ \text{given parent income } X}} = \beta_0 + \underbrace{\beta_1}_{\substack{\text{Intergenerational} \\ \text{elasticity} \\ \text{(IGE)}}} \underbrace{\log(X)}_{\substack{\text{Log parent} \\ \text{income}}}$$

MG proposal :

$$\underbrace{\log(E(Y | X))}_{\substack{\text{Log of mean} \\ \text{offspring income } Y \\ \text{given parent income } X}} = \alpha_0 + \underbrace{\alpha_1}_{\substack{\text{Intergenerational} \\ \text{elasticity of the} \\ \text{expectation (IGEE)}}} \underbrace{\log(X)}_{\substack{\text{Log parent} \\ \text{income}}}$$

Geometric Mean is Closer to the Median Than the Mean



COMMENT: SUMMARIZING INCOME MOBILITY WITH MULTIPLE SMOOTH QUANTILES INSTEAD OF PARAMETERIZED MEANS

*Ian Lundberg**

*Brandon M. Stewart**

*Department of Sociology and Office of Population Research, Princeton University,
Princeton, NJ, USA

Corresponding Author: Ian Lundberg, ilundberg@princeton.edu

DOI: 10.1177/0081175020931126

Single-number summaries that capture the relationship of socioeconomic outcomes across generations are a cornerstone of economic mobility research. Studies often focus on the intergenerational elasticity (IGE) of income: the coefficient β_1 on parent log income in a model predicting offspring log income (e.g., Aaronson and Mazumder 2008; Björklund and Jäntti 1997; Solon 2004). A large β_1 is often interpreted as evidence that incomes persist to a substantial degree across generations.

Images from this section are from this paper or earlier drafts of it.

Two Implicit Choices

(1) Summary Statistics for the Conditional Distribution

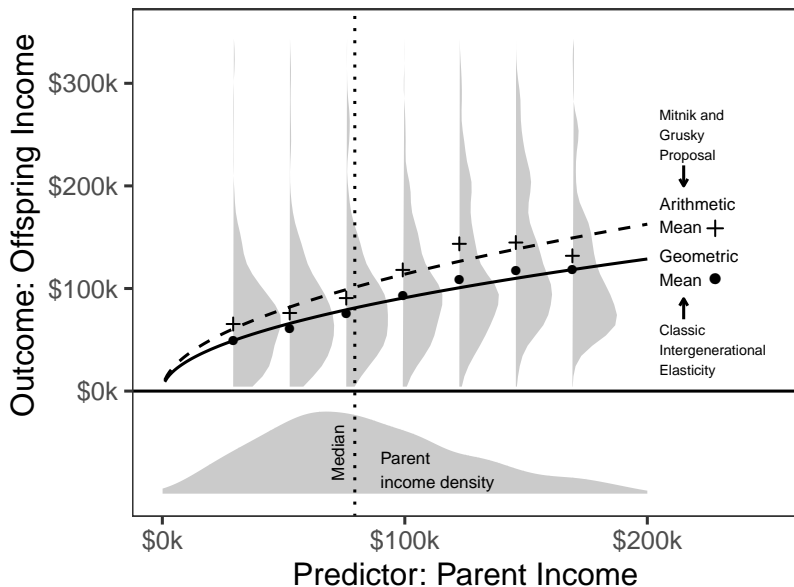
(gets you down to one number per value of x)

and

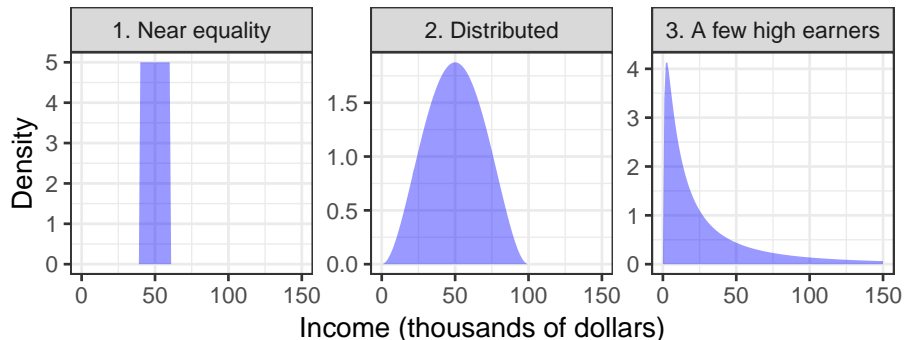
(2) Assume or Learn a Functional Form

(potentially simplifies the set of summary statistics to a single number)

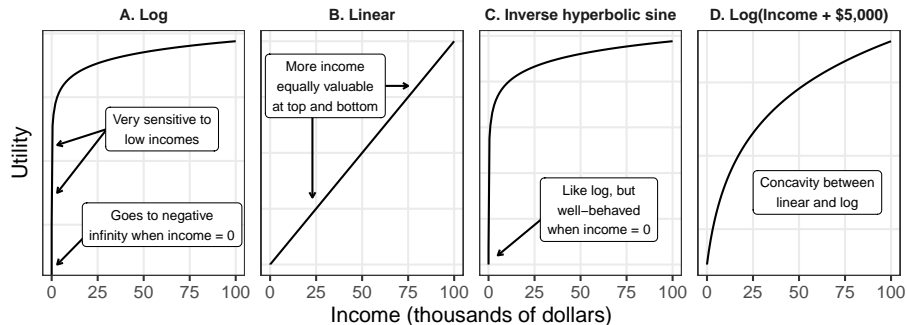
Visualizing the MG Proposal



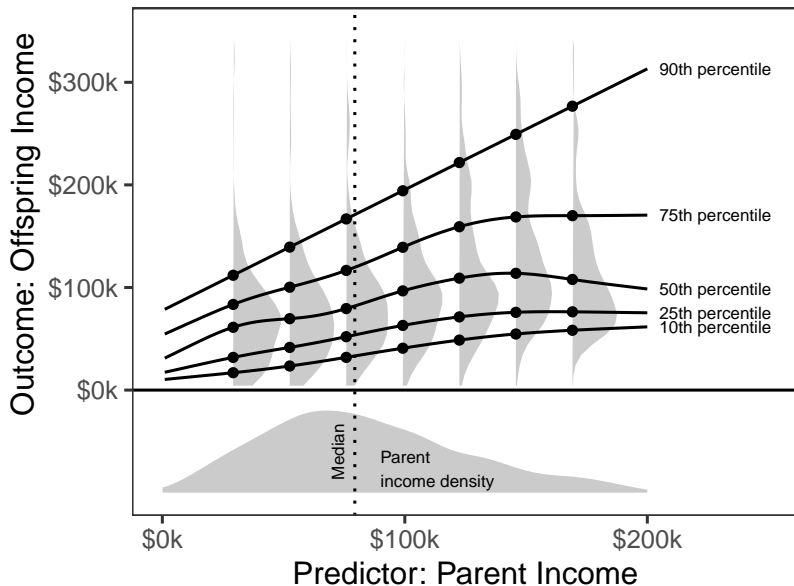
Single Summary Statistics Necessarily Mask Information



The Mean is a Normative Choice



A New Proposal



Single Number Summaries

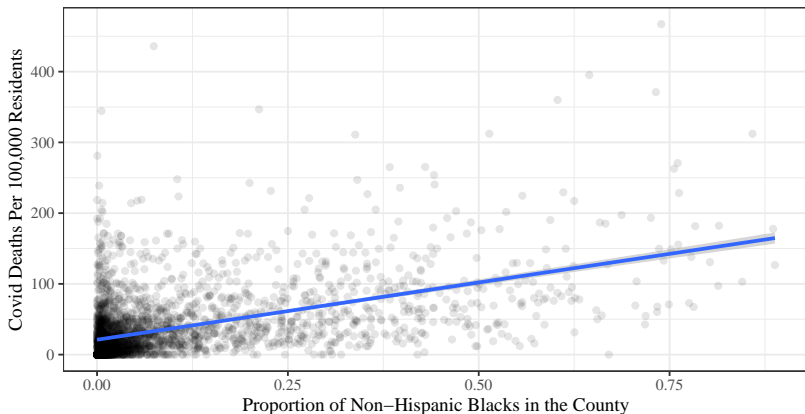
- A key selling point of the conventional IGE, the MG proposal and regression more broadly is the **single-number summary**.
- Any such summary necessitates a **loss of information**.
- Even with more complex functional forms, we can always calculate such a summary. For instance here, median is (on average) \$4k higher when parent income is \$10k higher.
- We obtain this by simply plugging in the 50th percentile at each offspring income, adding \$10k to each parent income and taking the average.
- If you are willing to commit to a **quantity of interest**, you can usually estimate it directly.
- At their best, single-number summaries are a way that the reader can calculate any approximation to a variety of quantities they are interested in. At their worst, they are a way for authors to abdicate responsibility for choosing a clear quantity of interest.

Broader Implications (Lee, Lundberg and Stewart)

Traditional Approach to Visualize Covid-19 Death Rates in US Counties

Covid data from NYTimes github as of 2020/09/07

Demographic data from American Community Survey 2014–2018 5–year estimate



Covid-19 Death Rates in US Counties

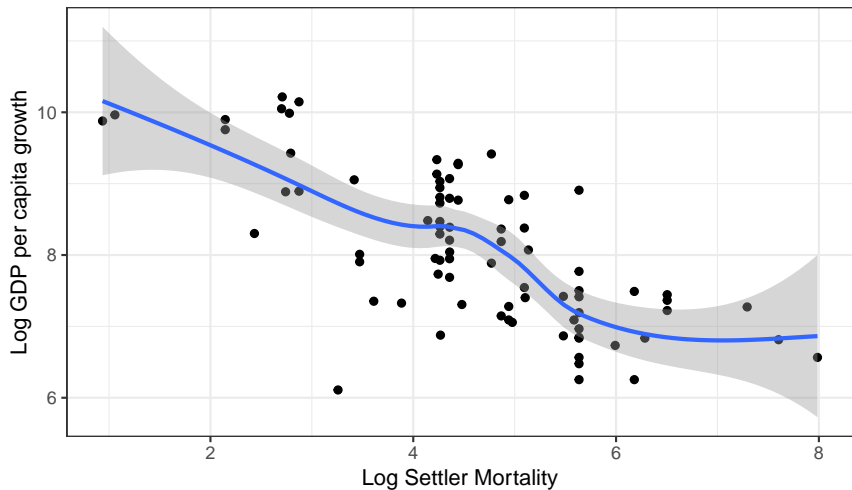
Covid data from NYTimes github as of 2020/09/07

Demographic data from American Community Survey 2014–2018 5–year estimate



- 1 Mechanics of OLS
- 2 Classical Perspective (Part 1, Unbiasedness)
 - Sampling Distributions
 - Classical Assumptions 1–4
- 3 Classical Perspective: Variance
 - Sampling Variance
 - Gauss-Markov
 - Large Samples
 - Small Samples
 - Agnostic Perspective
- 4 Inference
 - Hypothesis Tests
 - Confidence Intervals
 - Goodness of fit
 - Interpretation
- 5 Non-linearities
 - Log Transformations
 - Fun With Logs
 - LOESS

So what is ggplot2 doing?



LOESS

- We can combine the nonparametric kernel method idea of using only **local** data with a **parametric** model
- Idea: fit a linear regression within each band
- Locally weighted scatterplot smoothing (**LOWESS** or **LOESS**):
 - 1 Pick a subset of the data that falls in the interval $[x - h, x + h]$
 - 2 Fit a line to this subset of the data (= **local linear regression**), weighting the points by their distance to x using a kernel function
 - 3 Use the fitted regression line to predict the expected value of $E[Y|X = x_0]$

LOESS Example

We Covered

- Interpretation with logged independent and dependent variables
- The geometric mean!

This Week in Review

- OLS!
- Classical regression assumptions!
- Inference!
- Logs!

Going Deeper:

Aronow and Miller (2019) *Foundations of Agnostic Statistics*.
Cambridge University Press. Chapter 4.

Next week: Linear Regression with Two Variables!