Week 9: What Can Go Wrong and How To Fix It, Diagnostics and Solutions

Brandon Stewart[1]

Princeton

November 14, 16 and 21, 2016

[1]These slides are heavily influenced by Matt Blackwell, Adam Glynn, Jens Hainmueller and Kevin Quinn.

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

- Last Week
  - regression in the social sciences
- This "Week"
  - Monday (14):
    - unusual and influential data $\rightarrow$ robust estimation
  - Wednesday (16):
    - non-linearity $\rightarrow$ generalized additive models
  - Monday (21):
    - unusual errors $\rightarrow$ sandwich SEs and block bootstrap
- After Thanksgiving
  - causality with measured confounding
- Long Run
  - regression $\rightarrow$ diagnostics $\rightarrow$ causal inference

Questions?

Argument for Next Three Classes

Residuals are important. Look at them.

Review of the OLS assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

- 1-4 give us unbiasedness/consistency
- 1-5 are the Gauss-Markov, allow for large-sample inference
- 1-6 allow for small-sample inference

Review of the OLS Assumptions

- **Identification / Data Description**
  - No Perfect Collinearity
- **Unbiasedness / Consistency**
  - No Perfect Collinearity
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
- **Gauss-Markov (BLUE) / Asymptotic Inference (z and χ^2)**
  - No Perfect Collinearity
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
- **Classical LM (BUE) / Small-Sample Inference (t and F)**
  - No Perfect Collinearity
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
  - Normality of Errors

Violations of the assumptions

1. Nonlinearity
   - Result: biased/inconsistent estimates
   - Diagnose: scatterplots, added variable plots, component-plus-residual plots
   - Correct: transformations, polynomials, different model (next class)
2. iid/random sample
   - Result: no bias with appropriate alternative assumptions (structured dependence)
   - Result (ii): violations imply heteroskedasticity
   - Result (iii): outliers from different distributions can cause inefficiency/bias
   - Diagnose/Correct: various, this lecture and next week
3. Perfect collinearity
   - Result: can't run OLS
   - Diagnose/correct: drop one collinear term

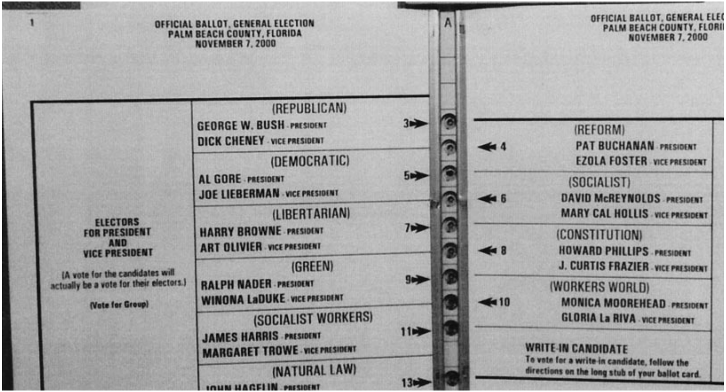
Violations of the assumptions (ii)

4. Zero conditional mean error
   - Result: biased/inconsistent estimates
   - Diagnose: very difficult
   - Correct: instrumental variables
5. Heteroskedasticity
   - Result: SEs are biased (usually downward)
   - Diagnose/correct: next week!
6. Non-Normality
   - Result: critical values for t and F tests wrong
   - Diagnose: checking the (studentized) residuals, QQ-plots, etc
   - Correct: transformations, add variables to $\mathbf{X}$, different model

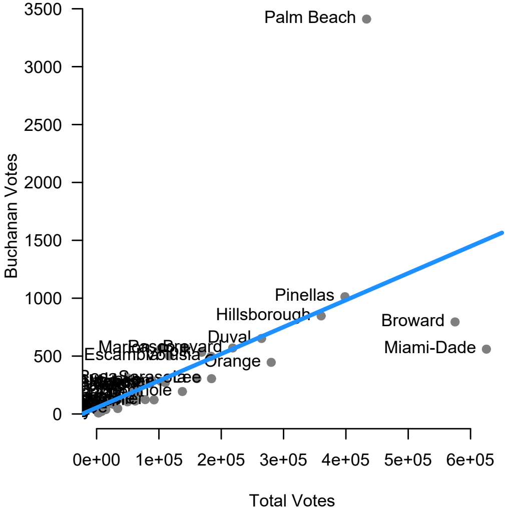
Example: Buchanan votes in Florida, 2000

Wand et al. show that the ballot caused 2,000 Democratic voters to vote by mistake for Buchanan, a number more than enough to have tipped the vote in FL from Bush to Gore, thus giving him FL's 25 electoral votes and the presidency.

FIGURE 1. The Palm Beach County Bufferfly Ballot



Example: Buchanan votes in Florida, 2000



Review of the Normality assumption

- In matrix notation:

$$\mathbf{u}|\mathbf{X} \sim \mathcal{N}(0, \sigma_u^2\mathbf{I})$$

- Equivalent to:

$$u_i|\mathbf{x}_i' \sim \mathcal{N}(0, \sigma_u^2)$$

- Fix $\mathbf{x}_i'$ and the distribution of errors should be Normal

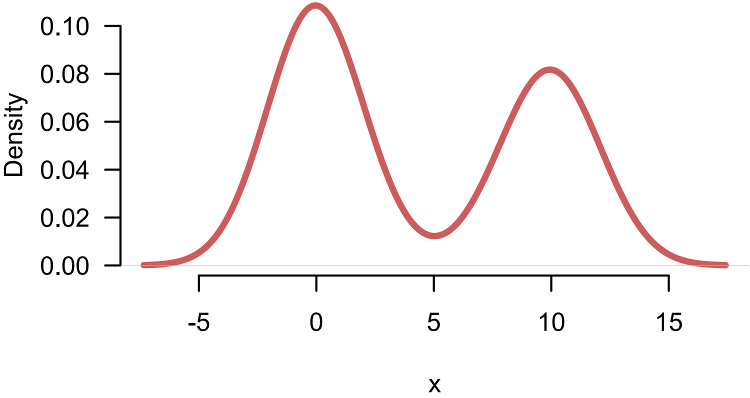
Consequences of non-Normal errors?

- In small samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}}$ will not be Normal
  - Test statistics will not have t or F distributions
  - Probability of Type I error will not be α
  - $1 - \alpha$ confidence interval will not have $1 - \alpha$ coverage
- In large samples:
  - Sampling distribution of $\widehat{\boldsymbol{\beta}} \approx$ Normal by the CLT
  - Test statistics will be $\approx t$ or F by the CLT
  - Probability of Type I error $\approx \alpha$
  - $1 - \alpha$ confidence interval will have $\approx 1 - \alpha$ coverage
- The sample size (n) needed for approximation to hold depends on how far the errors are from Normal.

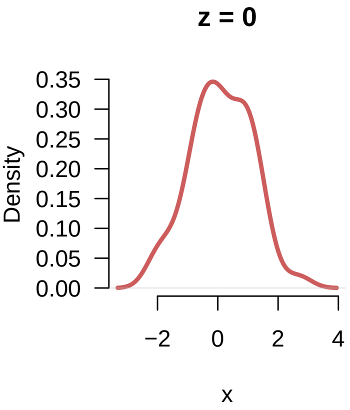
Marginal versus conditional

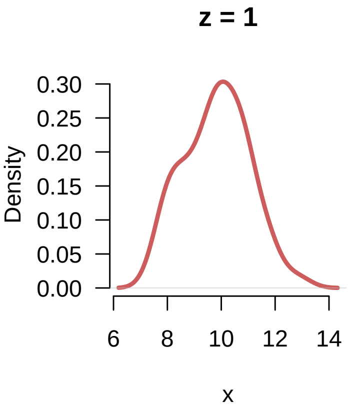
- Be careful with this assumption: distribution of the error ($u = y - X\beta$), not the distribution of the outcome y is the key assumption
- The marginal distribution of y can be non-Normal even if the conditional distribution is Normal!
- The plausibility depends on the X chosen by the researcher.

Example: Is this a violation?



Example: Is this a violation?





How to diagnose?

- Assumption is about unobserved $\mathbf{u} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta}$
- We can only observe residuals, $\widehat{\mathbf{u}} = \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$
- If distribution of residuals $\approx$ distribution of errors, we could check residuals
- But this is actually not true—the distribution of the residuals is complicated

Solution: Carefully investigate the residuals numerically and graphically.

To understand the relationship between residuals and errors, we need to derive the distribution of the residuals.

Hat matrix

- Define matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$

$$\begin{aligned}\widehat{\mathbf{u}} &= \mathbf{y} - \mathbf{X}\widehat{\beta} \\ &= \mathbf{y} - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} \\ &\equiv \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y}\end{aligned}$$

- $\mathbf{H}$ is the hat matrix because it puts the "hat" on $\mathbf{y}$:

$$\widehat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

  - $\mathbf{H}$ is an $n \times n$ symmetric matrix
  - $\mathbf{H}$ is idempotent: $\mathbf{H}\mathbf{H} = \mathbf{H}$

Relating the residuals to the errors

$$
\begin{aligned}
\widehat{\mathbf{u}} &= (\mathbf{I} - \mathbf{H})(y) \\
&= (\mathbf{I} - \mathbf{H})(\mathbf{X}\beta + \mathbf{u}) \\
&= (\mathbf{I} - \mathbf{H})\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{I}\mathbf{X}\beta - \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= \mathbf{X}\beta - \mathbf{X}\beta + (\mathbf{I} - \mathbf{H})\mathbf{u} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{u}
\end{aligned}
$$

- Residuals $\widehat{\mathbf{u}}$ are a linear function of the errors, $\mathbf{u}$
- For instance,

$$\widehat{u}_1 = (1 - h_{11})u_1 - \sum_{i=2}^{n} h_{1i} u_i$$

- Note that the residual is a function of all of the errors

Distribution of the residuals

$$\mathbb{E}[\widehat{\mathbf{u}}] = (\mathbf{I} - \mathbf{H})\mathbb{E}[\mathbf{u}] = \mathbf{0}$$

$$\text{Var}[\hat{\mathbf{u}}] = \sigma_u^2(\mathbf{I} - \mathbf{H})$$

The variance of the ith residual $\hat{u}_i$ is $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, where h_{ii} is the ith diagonal element of the matrix $\mathbf{H}$ (called the hat value).

Distribution of the Residuals

Notice in contrast to the unobserved errors, the estimated residuals

1. are not independent (because they must satisfy the two constraints $\sum_{i=1}^{n} \widehat{u}_i = 0$ and $\sum_{i=1}^{n} \widehat{u}_i x_i = 0$)
2. do not have the same variance. The variance of the residuals varies across data points $V[\hat{u}_i] = \sigma^2(1 - h_{ii})$, even though the unobserved errors all have the same variance σ^2

These properties can obscure the true patterns in the error distribution, and thus are inconvenient for our diagnostics.

Standardized Residuals

Let's address the second problem (unequal variances) by standardizing $\hat{u}_i$, i.e., dividing by their estimated standard deviations.

This produces standardized (or "internally studentized") residuals:

$$\hat{u}'_i = \frac{\hat{u}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}$$

where $\hat{\sigma}^2$ is our usual estimate of the error variance.

The standardized residuals are still not ideal, since the numerator and denominator of $\hat{u}'_i$ are not independent. This makes the distribution of $\hat{u}'_i$ nonstandard.

Studentized residuals

If we remove observation i from the estimation of σ, then we can eliminate the dependence and the result will have a standard distribution.

- estimate residual variance without residual i:

$$\widehat{\sigma}^2_{-i} = \frac{\mathbf{u}'\mathbf{u} - u_i^2/(1-h_{ii})}{n-k-2}$$

- Use this i-free estimate to standardize, which creates the studentized residuals:

$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1-h_{ii}}}$$

- If the errors are Normal, the studentized residuals follow a t distribution with $(n-k-2)$ degrees of freedom.
- Deviations from $t \implies$ violation of Normality

Example: Buchanan Votes in Florida

- Now that our studentized residuals follow a known standard distribution, we can proceed with diagnostic analysis for the nonnormal errors.
- We examine data from the 2000 presidential election in Florida used in Wand et al. (2001).
- Our analysis takes place at the county level and we will regress the number of Buchanan votes in each county on the total number of votes in each county.

Buchanan Votes and Total Votes

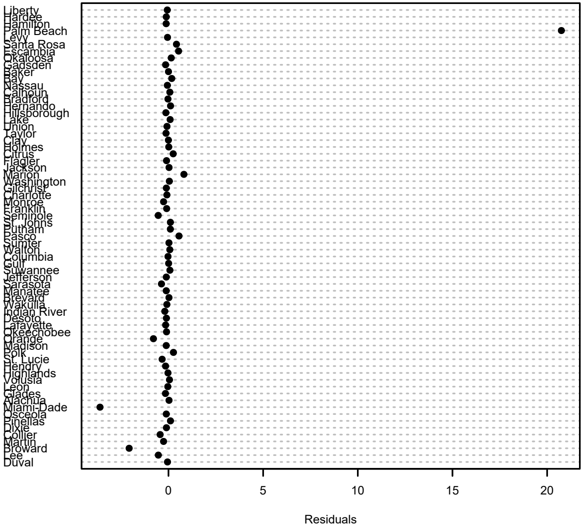
R Code

```
> mod1 <- lm(buchanan00~TotalVotes00,data=dta)
> summary(mod1)
Residuals:
    Min      1Q  Median      3Q     Max
-947.05  -41.74  -19.47   20.20 2350.54

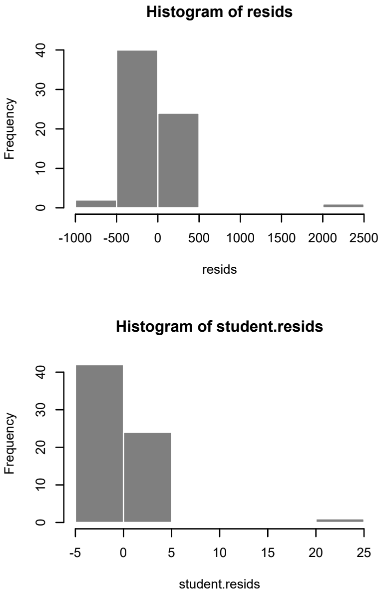
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.423e+01  4.914e+01   1.104    0.274
TotalVotes00 2.323e-03  3.104e-04   7.483 2.42e-10 ***
---
Residual standard error: 332.7 on 65 degrees of freedom
Multiple R-squared: 0.4628,	Adjusted R-squared: 0.4545
F-statistic:    56 on 1 and 65 DF,  p-value: 2.417e-10

> residuals              <- resid(mod1)
> standardized_residuals <- rstandard(mod1)
> studentized_residuals  <- rstudent(mod1)
> dotchart(residuals,dta$name,cex=.7,xlab="Residuals")
```

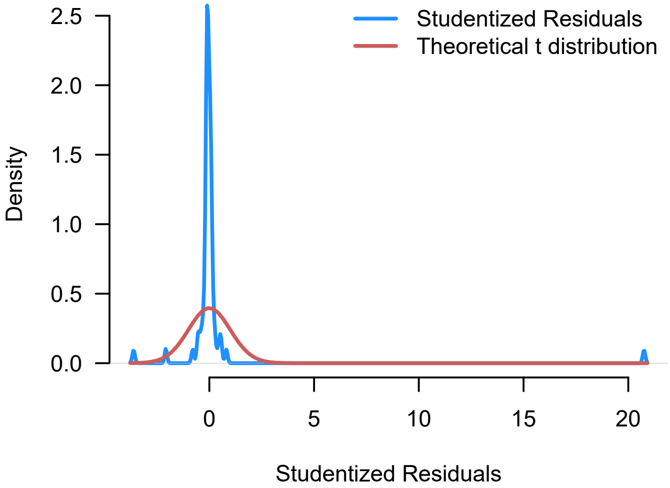
Plotting the residuals



Plotting the residuals



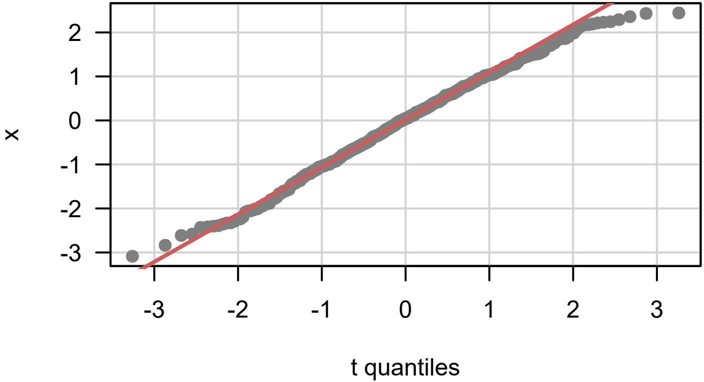
Plotting the residuals



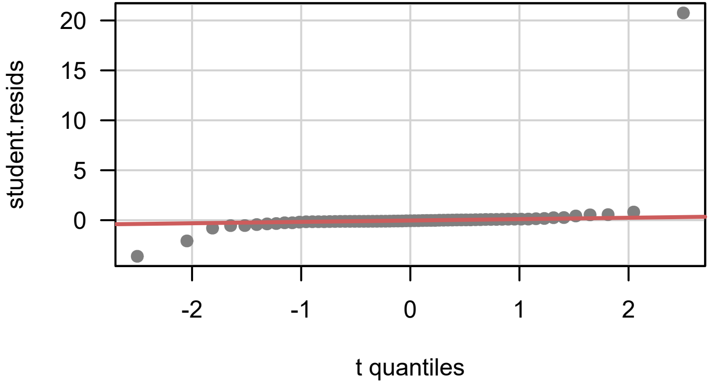
Quantile-Quantile plots

- Quantile-quantile plot or QQ-plot is useful for comparing distributions
- Plots the quantiles of one distribution against those of another distribution
- For example, one point is the (m_x, m_y) where m_x is the median of the x distribution and m_y is the median for the y distribution
- If distributions are equal $\implies$ 45 degree line

Good QQ-plot



Buchanan QQ-plot



How can we deal with nonnormal errors?

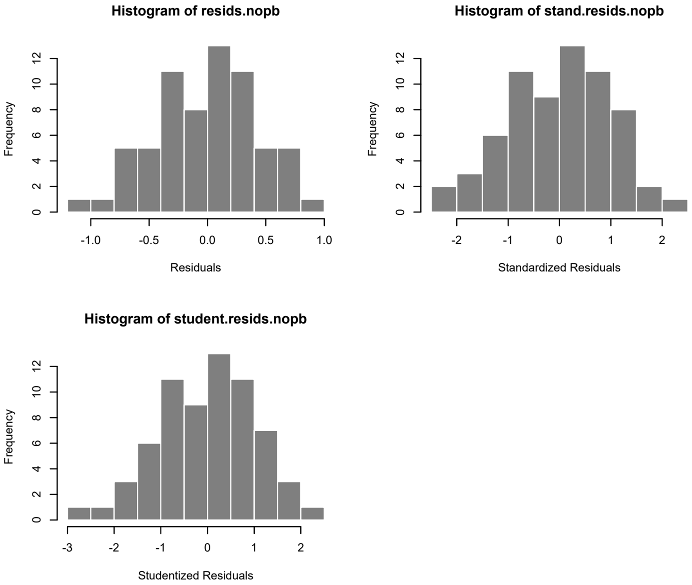
- Drop or change problematic observations (could be a bad idea unless you have some reason to believe the data are wrong or corrupted)
- Add variables to $\mathbf{X}$ (remember that the errors are defined in terms of explanatory variables)
- Use transformations (this may work, but a transformation affects all the assumptions of the model)
- Use estimators other than OLS that are robust to nonnormality (later this class)
- Consider other causes (next two classes)

Buchanan revisited

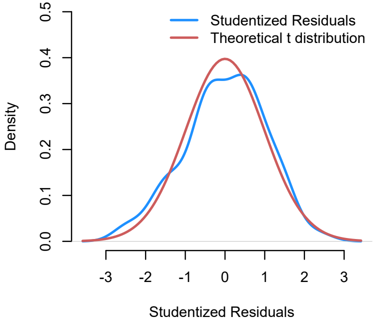
Let's delete Palm Beach and also use log transformations for both variables

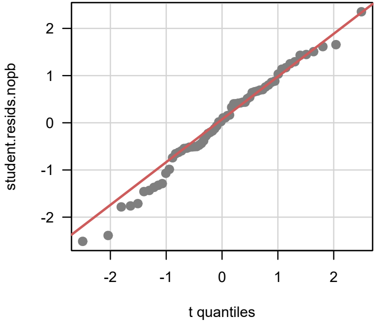
```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.48597    0.37889  -6.561 1.09e-08 ***
## log(edaytotal)  0.70311    0.03621  19.417  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4362 on 64 degrees of freedom
## Multiple R-squared:  0.8549, Adjusted R-squared:  0.8526
## F-statistic:   377 on 1 and 64 DF,  p-value: < 2.2e-16
```

Buchanan revisited



Buchanan revisited





A Note of Caution About Log Transformations

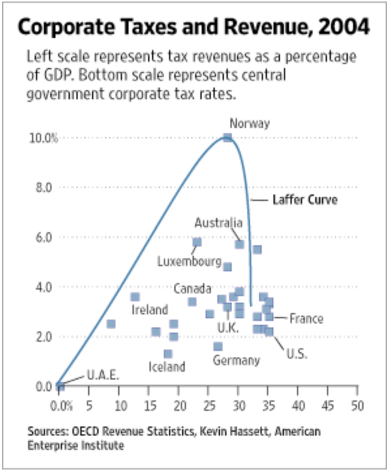
- Log transformations are a standard approach in the literature and intro regression classes
- They are extremely helpful for data that is skewed (e.g. a few very large positive values)
- Generally you want to convert these findings back to the original scale for interpretation
- You should know though that estimates of marginal effects in the untransformed states are not necessarily unbiased.
- Jensen's inequality gives us information on this relation: $f(E[X]) \leq E[f(X)]$ for any convex function $f()$
- The results will in general be consistent which ensures that the bias decreases in sample size.

The trouble with Norway

- Lange and Garrett (1985): organizational and political power of labor interact to improve economic growth
- Jackman (1987): relationship just due to North Sea Oil?
- Table guide:
  - x_1 = organizational power of labor
  - x_2 = political power of labor
  - Parentheses contain t-statistics

	Constant	x_1	x_2	$x_1 \cdot x_2$
Norway Obs Included	.814	-.192	-.278	.137
	(4.7)	(2.0)	(2.4)	(2.9)
Norway Obs Excluded	.641	-.068	-.138	.054
	(4.8)	(0.9)	(1.5)	(1.3)

Creative curve fitting with Norway



The Most Important Lesson: Check Your Data

"Do not attempt to build a model on a set of poor data! In human surveys, one often finds 14-inch men, 1000-pound women, students with 'no' lungs, and so on. In manufacturing data, one can find 10,000 pounds of material in a 100 pound capacity barrel, and similar obvious errors.

All the planning, and training in the world will not eliminate these sorts of problems. In our decades of experience with 'messy data,' we have yet to find a large data set completely free of such quality problems."

Draper and Smith (1981, p. 418)

Always Carefully Examine the Data First!!

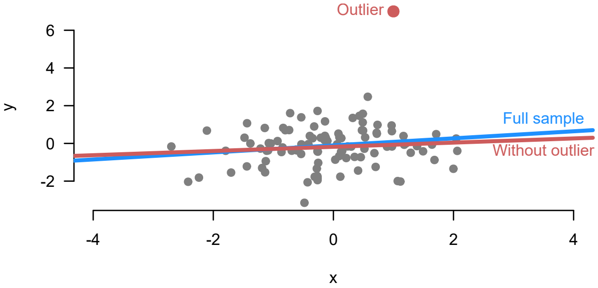
1. Examine summary statistics: `summary(data)`
2. Scatterplot matrix for densities and bivariate relationships: E.g. `scatterplotMatrix(data)` from `car` library.
3. Further conditional plots for multivariate data: E.g. use the `lattice` library or `ggplot2`

Three types of extreme values

1. Outlier: extreme in the y direction
2. Leverage point: extreme in one x direction
3. Influence point: extreme in both directions

- Not all of these are problematic
- If the data are truly "contaminated" (come from a different distribution), can cause inefficiency and possibly bias
- Can be a violation of iid (not identically distributed)

Outlier definition



- An **outlier** is a data point with very large regression errors, u_i
- Very distant from the rest of the data in the y-dimension
- Increases standard errors (by increasing $\widehat{\sigma}^2$)
- No bias if typical in the x's

Detecting outliers

- Look at standardized residuals, $\widehat{u}_i'$?
  - but $\widehat{\sigma}^2$ could be biased upwards by the large residual from the outlier
- Makes detecting residuals harder
- Possible solution: use studentized residuals

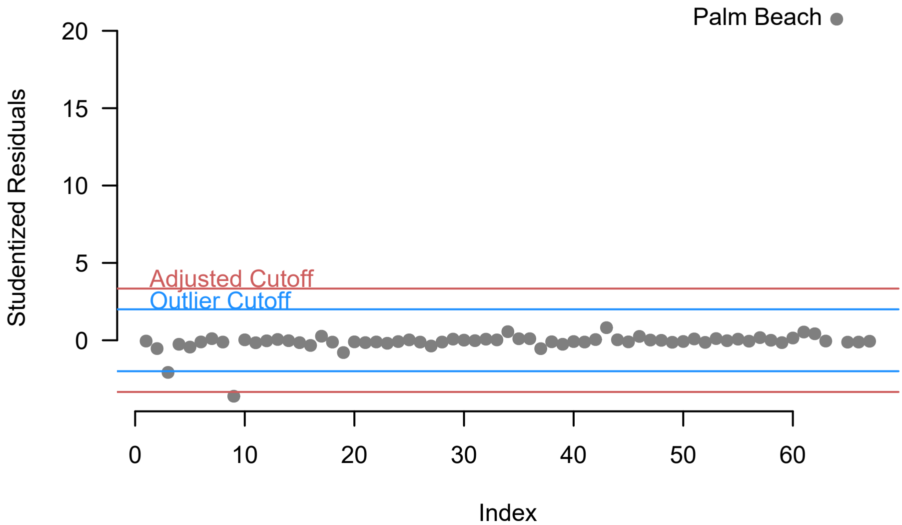
$$\widehat{u}_i^* = \frac{\widehat{u}_i}{\widehat{\sigma}_{-i}\sqrt{1 - h_i}}$$

- $\widehat{\sigma} > \widehat{\sigma}_{-i}$ because we drop the large residual from the outlier, and so $\widehat{u}_i' < \widehat{u}_i^*$

Cutoff rules for outliers

- The studentized residuals follow a t distribution, $u_i^* \sim t_{n-k-2}$, when $u_i \sim N(0, \sigma^2)$
- Rule of thumb: $|\widehat{u}_i^*| > 2$ will be relatively rare
- Extreme outliers, $|\widehat{u}_i^*| > 4 - 5$ are much less likely
- People usually adjust cutoff for multiple testing

Buchanan outliers



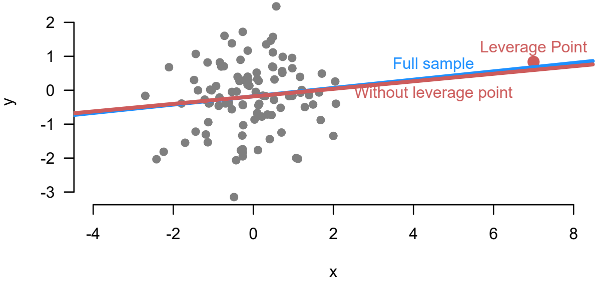
What to do about outliers

- Is the data corrupted?
  - Fix the observation (obvious data entry errors)
  - Remove the observation
  - Be transparent either way
- Is the outlier part of the data generating process?
  - Transform the dependent variable ($\log(y)$)
  - Use a method that is robust to outliers (robust regression)

A Cautionary Tale: The "Discovery" of the Ozone Hole

- In the late 70s, NASA used an automated data processing program on satellite measurements of atmospheric data to track changes in atmospheric variables such as ozone.
- This data "quality control" algorithm rejected abnormally low readings of ozone over the Antarctic as unreasonable.
- This delayed the detection of the ozone hole by several years until British Antarctic Survey scientists discovered it based on analysis of their own observations (*Nature*, May 1985).
- The ozone hole was detected in satellite data only when the raw data was reprocessed. When the software was rerun without the pre-processing flags, the ozone hole was seen as far back as 1976.

Leverage point definition



- Values that are extreme in the x direction
- That is, values far from the center of the covariate distribution
- Decrease SEs (more X variation)
- No bias if typical in y dimension

Leverage Points: Hat values

To measure leverage in multivariate data we will go back to the hat matrix $\mathbf{H}$:

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{y} = \mathbf{H}\mathbf{y}$$

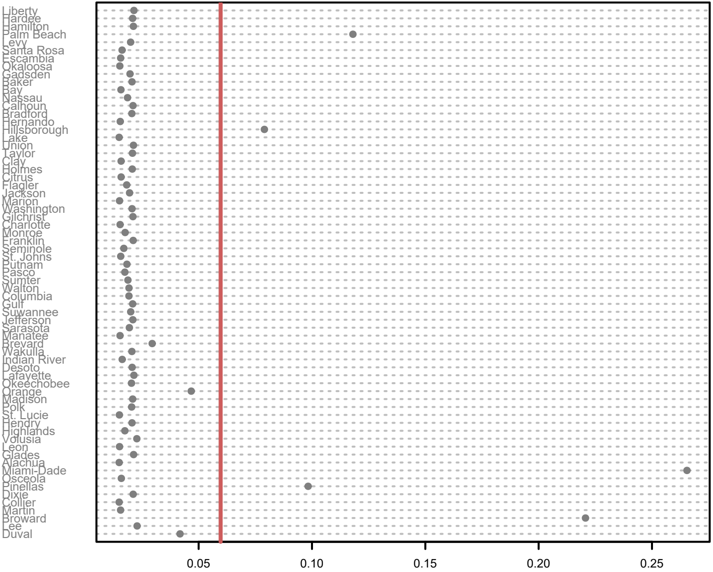
$\mathbf{H}$ is $n \times n$, symmetric, and idempotent. It generates fitted values as follows:

$$\hat{y}_i = \mathbf{h}_i'\mathbf{y} = \begin{bmatrix} h_{i,1} & h_{i,2} & \cdots & h_{i,n} \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \sum_{j=1}^{n} h_{i,j} y_j$$

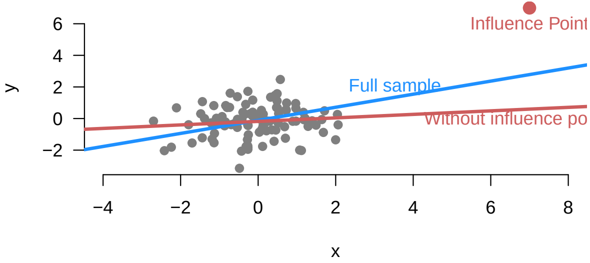
Therefore,

- h_{ij} dictates how important y_j is for the fitted value $\hat{y}_i$ (regardless of the actual value of y_j, since $\mathbf{H}$ depends only on $\mathbf{X}$)
- The diagonal entries $h_{ii} = \sum_{j=1}^{n} h_{ij}^2$, so they summarize how important y_i is for all the fitted values. We call them the hat values or leverages and a single subscript notation is used: $h_i = h_{ii}$
- Intuitively, the hat values measure how far a unit's vector of characteristics $\mathbf{x}_i$ is from the vector of means of $\mathbf{X}$
- Rule of thumb: examine hat values greater than $2(k+1)/n$

Buchanan hats



Influence points



- An influence point is one that is both an outlier (extreme in X) and a leverage point (extreme in Y).
- Causes the regression line to move toward it (bias?)

Detecting Influence Points/Bad Leverage Points

- Influence Points:
  Influence on coefficients = Leverage $\times$ Outlyingness

- More formally: Measure the change that occurs in the slope estimates when an observation is removed from the data set. Let

$$D_{ij} = \hat{\beta}_j - \hat{\beta}_{j(-i)}, \quad i = 1, \ldots, n, \;\; j = 0, \ldots, k$$

  where $\hat{\beta}_{j(-i)}$ is the estimate of the jth coefficient from the same regression once observation i has been removed from the data set.

- D_{ij} is called the DFbeta, which measures the influence of observation i on the estimated coefficient for the jth explanatory variable.

Standardized Influence

To make comparisons across coefficients, it is helpful to scale D_{ij} by the estimated standard error of the coefficients:

$$D_{ij}^* = \frac{\hat{\beta}_j - \hat{\beta}_{j(-i)}}{\hat{SE}_{-i}(\hat{\beta}_j)}$$

where D_{ij}^* is called DFbetaS.

- $D_{ij}^* > 0$ implies that removing observation i decreases the estimate of $\beta_j \rightarrow$ obs i has a positive influence on β_j.
- $D_{ij}^* < 0$ implies that removing observation i increases the estimate of $\beta_j \rightarrow$ obs i has a negative influence on β_j.
- Values of $|D_{ij}^*| > 2/\sqrt{n}$ are an indication of high influence.
- In R: `dfbetas(model)`

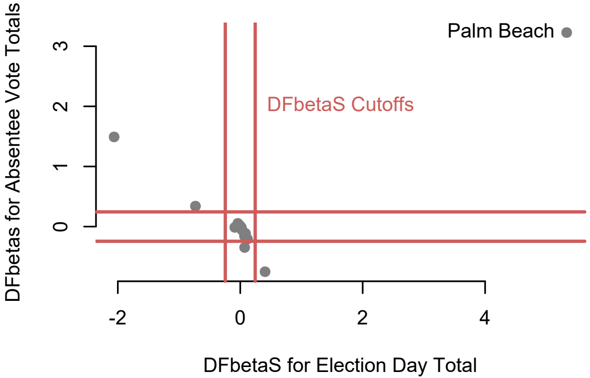
Buchanan influence

```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.935e+01  5.520e+01  -0.532  0.59686
## edaytotal     1.100e-03  4.797e-04   2.293  0.02529 *
## absnbuchanan  6.895e+00  2.129e+00   3.238  0.00195 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 317.2 on 61 degrees of freedom
##   (3 observations deleted due to missingness)
## Multiple R-squared:  0.5361, Adjusted R-squared:  0.5209
## F-statistic: 35.24 on 2 and 61 DF,  p-value: 6.711e-11
```

Buchanan influence

```
##     (Intercept)    edaytotal  absnbuchanan
## 1  0.3454475146  0.4050504921 -0.7505222758
## 2 -0.0234266617 -0.0241000045 -0.0131672181
## 3  0.0650795039 -0.7319311820  0.3401669862
## 4 -0.0333980968  0.0133802934 -0.0087505576
## 5 -0.0397626659 -0.0073746223  0.0096551713
## 6 -0.0009277798  0.0001505476  0.0002210247
```

Buchanan influence



- Palm Beach county moves each of the coefficients by more than 3 standard errors!

Summarizing Influence across All Coefficients

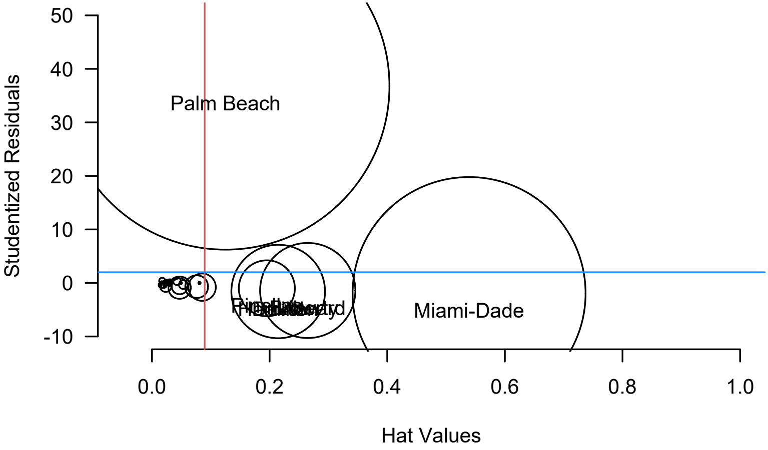
- Leverage tells us how much one data point affects a single coefficient.
- A number of summary measures exist for influence of data points across all coefficients, all involving both leverage and outlyingness.
- A popular measure is Cook's distance:

$$D_i = \frac{\hat{u}_i'^2}{k+1} \times \frac{h_i}{1-h_i}$$

  where $\hat{u}_i'$ is the standardized residual and h_i is the hat value.
  - It can be shown that D_i is a weighted sum of $k+1$ DFbetaS's for observation i
  - In R, `cooks.distance(model)`
  - $D > 4/(n-k-1)$ is commonly considered large

- The influence plot: the studentized residuals plotted against the hat values, size of points proportional to Cook's distance.

Influence Plot Buchanan



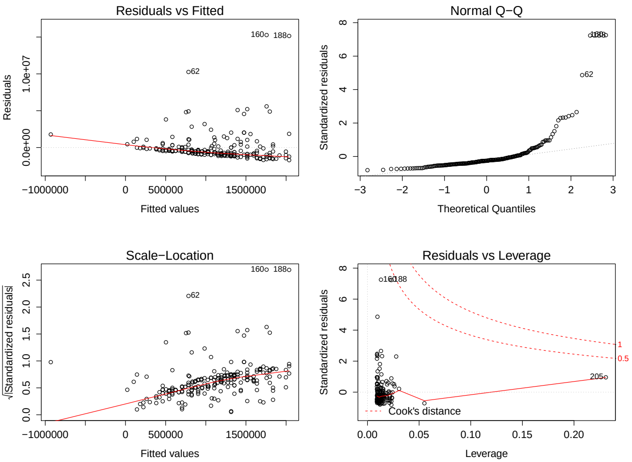
Code for Influence Plot

```
mod3 <- lm(edaybuchanan ~ edaytotal + absnbuchanan, data = flvote)
symbols(y = rstudent(mod3), x = hatvalues(mod3),
            circles = sqrt(cooks.distance(mod3)),
            ylab = "Studentized Residuals",
            xlab = "Hat Values", xlim = c(-0.05, 1),
            ylim = c(-10, 50), las = 1, bty = "n")
cutoffstud <- 2
cutoffhat <- 2 * (3)/nrow(flvote)
abline(v = cutoffhat, col = "indianred")
abline(h = cutoffstud, col = "dodgerblue")
filter <- rstudent(mod3) > cutoffstud | hatvalues(mod3) > cutoffhat
text(y = rstudent(mod3)[filter],
       x = hatvalues(mod3)[filter],
       flvote$county[filter], pos = 1)
```

A Quick Function for Standard Diagnostic Plots

R Code

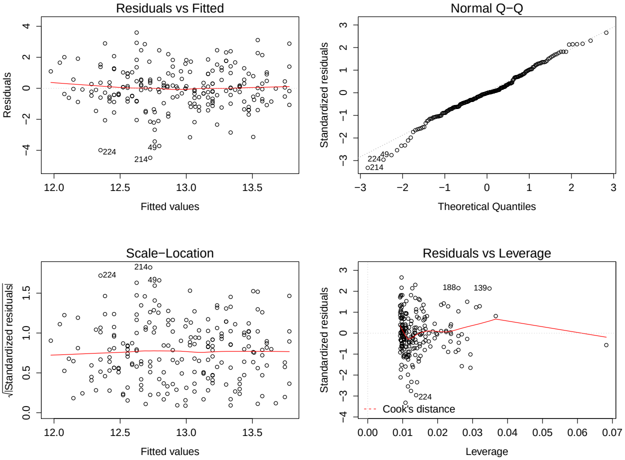
```
> par(mfrow=c(2,2))
> plot(mod1)
```



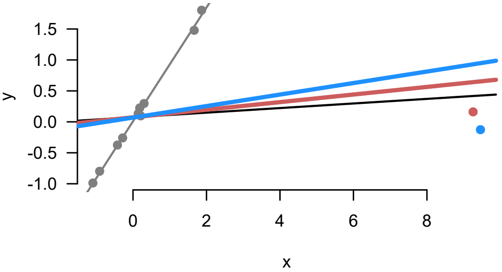
The Improved Model

R Code

```
> par(mfrow=c(2,2))
> plot(mod2)
```



Limitations of the standard tools



- What happens when there are two influence points?
- Red line drops the red influence point
- Blue line drops the blue influence point
- Neither of the “leave-one-out” approaches helps recover the line

The Idea of Robustness

- We will cover a few ideas in robust statistics over the next few days (much of which is due directly or indirectly to Peter Huber)
- Robust methods are procedures that are designed to continue to provide 'reasonable' answers in the presence of violation of some assumptions.
- A lot of social scientists use robust standard errors (we will discuss next week) but far fewer use robust regression tools.
- These methods used to be computationally prohibitive but haven't been for the last 10-15 years

But What About Gauss-Markov and BLUE?

- One argument here is that even without normality, we know that Gauss-Markov is the Best Linear Unbiased Estimator (BLUE)
- How comforting should this be? Not very.
- The Linear point is an artificial restriction. It means the estimator has to be of the form $\hat{\beta} = \mathbf{W}y$ but why only use those?
- With normality assumption we get Best Unbiased Estimator (BUE) which is quite comforting when $n \gg p$ (number of observations much larger than number of variables).

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

"[Even without normally distributed errors] OLS coefficient estimators remain unbiased and efficient." - Berry (1993)

Quotes from Rainey and Baissa (2015) presentation

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

> *"[The Gauss-Markov theorem] justifies the use of the OLS method rather than using a variety of competing estimators" - Wooldridge (2013)*

Quotes from Rainey and Baissa (2015) presentation

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

"We need not look for another linear unbiased estimator, for we will not find such an estimator whose variance is smaller than the OLS estimator"
- Gujarati (2004)

Quotes from Rainey and Baissa (2015) presentation

This Point is Not Obvious

This flies in the face of most conventional wisdom in textbooks.

"The Gauss-Markov theorem allows us to have considerable confidence in the least squares estimators." - Berry and Feldman (1993)

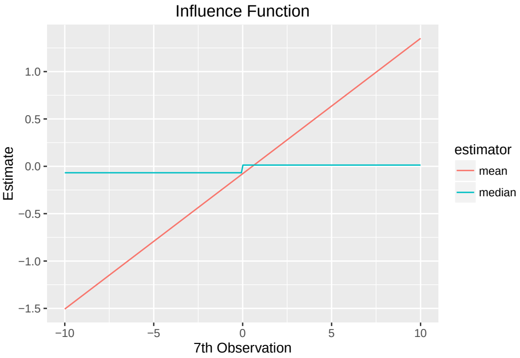
Quotes from Rainey and Baissa (2015) presentation

Robustly Estimating a Location

- Let's simplify- what if we want to estimate the center of a symmetric distribution.
- Two options (of many): mean and median
- Characteristics to consider: efficiency when assumptions hold, sensitivity to assumption violation.
- For normal data $y_i \sim \mathcal{N}(\mu, \sigma^2)$, median is less efficient:
  - $V(\hat{\mu}_{\text{mean}}) = \frac{\sigma^2}{n}$
  - $V(\hat{\mu}_{\text{median}}) = \frac{\pi\sigma^2}{2n}$
  - Median is $\frac{\pi}{2}$ times larger (i.e. less efficient)
- We can measure sensitivity with the influence function which measures change in estimator based on corruption in one datapoint.

Influence Function

- Imagine that we had a sample Y from a standard normal: -0.068, -1.282, 0.013, 0.141, -0.980, 1.63. $\bar{Y} = -1.52$
- Now imagine we add a contaminated 7th observation which could range from -10 to +10. How would the estimator change for the median and mean?



Example from Fox

Breakdown Point

- The influence function showed us how one aberrant point can change the resulting estimate.
- We also want to characterize the breakdown point which is the fraction of arbitrarily bad data that the estimator can tolerate without being affected to an arbitrarily large extent
- The breakdown point of the mean is 0 because (as we have seen) a single bad data point can change things a lot.
- The median has a breakdown point of 50% because half the data can be bad without causing the median to become completely unstuck.

M-estimators

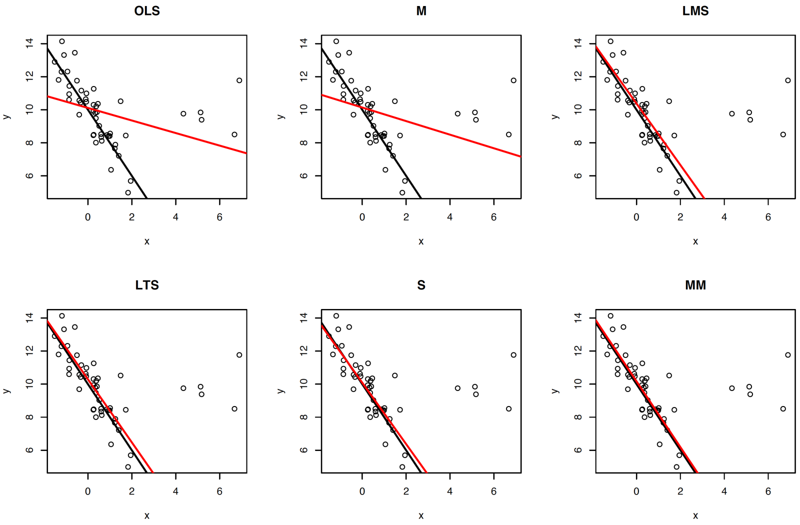
- We can phrase this more generally than the mean or the median which will allow us to extend the ideas to regression via M-estimation
- M-estimators minimize a sum over an objective function $\sum_i^n \rho(E)$ where E is $Y_i - \hat{\mu}$
  - The mean has $\sum_i \rho(E) = \sum_i (Y_i - \hat{\mu})^2$
  - The median has $\sum_i \rho(E) = \sum_i |(Y_i - \hat{\mu})|$
- The shape of the influence function is determined by the derivative of the objective function with respect to E.
- Other objectives include the Huber objective and Tukey's biweight objective which have different properties.
- Calculating robust M estimators often requires an iterative procedure and a careful initialization.

M-estimation for Regression

- We can apply this to regression fairly straightforwardly. In robust M-estimators we choose $\rho()$ so that observations with large residuals get less weight.
- Can be very robust to outliers in the Y space (less so in the X space usually)
- Some options:
  - Least Median Squares: choose $\hat{\beta}$ to minimize $\text{median}\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LMS}})^2\right\}_{i=1}^{n}$. Very high breakdown point, but very inefficient.
  - Least Trimmed Squares: choose $\hat{\beta}$ to minimize the sum of the p smallest elements of $\left\{(y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_{\text{LTS}})^2\right\}_{i=1}^{n}$. High breakdown point and more efficient, still not as efficient as some.
  - *MM*-estimator: with Huber's loss is what I recommend in practice (more in appendix)
- You can find an asymptotic covariance matrix for M-estimators but I would bootstrap it if possible as the asymptotics kick in slowly.

```
library(MASS)
set.seed(588)
n <- 50
x <- rnorm(n)
y <- 10 - 2*x + rnorm(n)
x[1:5] <- rnorm(5, mean=5)
y[1:5] <- 10 +  rnorm(5)
ols.out <- lm(y~x)
m.out <- rlm(y~x, method="M")
lms.out <- lqs(y~x, method="lms")
lts.out <- lqs(y~x, method="lts")
s.out <- lqs(y~x, method="S")
mm.out <- rlm(y~x, method="MM")
```

Simulation Results



Thoughts on Robust Estimators

- Robust estimators aren't commonly seen in applied social science work but perhaps they should be.
- Even though Gauss-Markov does not require normality, the L in BLUE is a fairly restrictive condition.
- In most cases I personally would start with OLS, do diagnostics and then consider a robust alternative. If I don't have time for diagnostics, maybe robust is better from the outset.
- I highly recommend Baissa and Rainey (2016) "When BLUE is Not Best: Non-Normal Errors and the Linear Model" for more on this topic.
- The Fox textbook Chapter 19 is also quite good on this and points out to the key references

Appendix: Characterizing Estimator Robustness (formally)

Definition (Breakdown Point)

The breakdown point of an estimator is the smallest fraction of the data that can be changed an arbitrary amount to produce an arbitrarily large change in the estimate (Seber and Lee 2003, pg 82)

Definition (Influence Function)

Let $F_p = (1-p)F + p\delta_{\mathbf{z}_0}$ where F is a probability measure, $\delta_{\mathbf{z}_0}$ is the point mass at $\mathbf{z}_0 \in \mathbb{R}^k$, and $p \in (0,1)$.
Let $T(\cdot)$ be a statistical functional. The influence function of T is

$$IF(\mathbf{z}_0; T, F) = \lim_{p \downarrow 0} \frac{T(F_p) - T(F)}{p}$$

The influence function is a function of $\mathbf{z}_0$ given T and F. It describes how T changes with small amounts of contamination at $\mathbf{z}_0$ (Hampel, Rousseeuw, Ronchetti, and Stahel, (1986), p. 84).

Appendix: S Estimators

To talk about $MM-$estimators we need a type of estimator called an S-estimator.

S-estimators work somewhat differently in that the goal is to minimize the scale estimate subject to a constraint.

An S-estimator for the regression model is defined as the values of $\hat{\boldsymbol{\beta}}_S$ and s that minimize s subject to the constraint:

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}_S}{s}\right) \geq K$$

where K is user-defined constant (typically set to 0.5) and $\rho : \mathbb{R} \rightarrow [0, 1]$ is a function with the following properties (Davies, 1990, p. 1653):

1. $\rho(0) = 1$
2. $\rho(u) = \rho(-u), u \in \mathbb{R}$
3. $\rho : \mathbb{R}_+ \rightarrow [0, 1]$ is nonincreasing, continuous at 0, and continuous on the left
4. for some $c > 0$, $\rho(u) > 0$ if $|u| < c$ and $\rho(u) = 0$ if $|u| > c$

Appendix: *MM*-Estimators

MM-estimators are, in some sense, the best of both worlds– very high breakdown point and good efficiency.

The work by first calculating S-estimates of the scale and coefficients and then using these as starting values for a particular M-estimator.

Good properties, but costly to compute (usually impossible to compute exactly).

Measurement Error

"It seems as if measurement error has been pushed into the role of the unwanted child whose existence we would rather deny. Maybe because measurement error is common, insipid, and unsophisticated. Unlike the hidden confounder challenging our intellect, to discover measurement error is a 'no-brainer' - it simply lurks everywhere. Our epidemiological fingerprints are contaminated with measurement error. Everything we observe, we observe with error. Since observation is our business, we would probably rather deny that what we observe is imprecise and maybe even inaccurate, but time has come to unveil the secret: measurement error is threatening our profession."

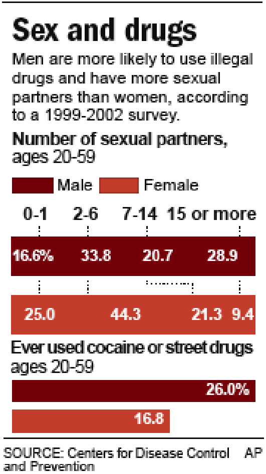
Karen Michals (2001)

Measuring Variables

Often, the variables that we use in our regression analysis are measured with error. For example:

- In cross-country data, variables are often measured by surveys within each country (e.g. perceived corruption)
- In individual level data, individuals may misreport information (sensitive questions about preferences, politics, income, etc.)
    - The Bradley effect (1982)
    - Shy Tory Factor (1992)
    - The Silent Trump Voter?
- In recall studys, people often can't remember (e.g. how many vegetables did you eat last week?)
- Many concepts are hard to measure (e.g. ability, IQ, religiosity, income, etc.)
- Other variables, like gender, number of children, may be measured with less error

US Survey Data



Measurement Error in the Dependent Variable

Consider the simple linear regression model where we observe Y_i^* instead of Y_i and the following relationships hold

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + U_i \\ U_i &\sim_{i.i.d} N(0, \sigma^2) \\ Y_i^* &= Y_i + E_i \end{aligned}$$

Therefore, the model for our observed data is

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 X_i + E_i + U_i \\ &= \beta_0 + \beta_1 X_i + V_i \end{aligned}$$

Let's assume that that the Gauss-Markov assumptions hold for the model with the true (but unobserved) variables so that OLS would be unbiased and consistent if we observed Y_i. Does the measurement error in Y_i^* cause any problems when fitting OLS to the observed data?

Measurement Error in the Dependent Variable

When we fit OLS to the observed data:

$$\begin{aligned} Y_i^* &= \beta_0 + \beta_1 X_i + E_i + U_i \\ &= \beta_0 + \beta_1 X_i + V_i \end{aligned}$$

- As long as the measurement error E is uncorrelated with X (eg. random reporting error), then under the usual assumptions V will be uncorrelated with X and OLS is unbiased and consistent
- If $E[E_i] > 0$ the constant will be biased upwards and vice versa, but our slope estimates remain unbiased $E[\hat{\beta}] = \beta$
- Since $V[U + E] = \sigma_U^2 + \sigma_E^2 > \sigma_U^2$ the estimates will be less precise
- If E is correlated with X then OLS is inconsistent and biased.

Measurement Error in the Independent Variable

Consider the simple linear regression model where all the traditional assumptions hold but we observe X_i^* instead of X_i

$$Y_i = \beta_0 + \beta_1 X_i + U_i$$
$$X_i^* = X_i + E_i$$

Therefore, the model for our observed data is

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 (X_i^* - E_i) + U_i \\ &= \beta_0 + \beta_1 X_i^* + (U_i - \beta_1 E_i) \end{aligned}$$

and we must state assumptions in terms of the new error term, $U_i - \beta_1 E_i$

Let's assume $E[E_i] = 0$ (the average error is zero) and the first four GM-assumptions hold (including $Cov[U, X] = 0$ and $Cov[U, X^*] = 0$) so that OLS would be unbiased and consistent if we observed X.

Does the measurement error cause any problems when fitting OLS to the observed data (i.e. using X^* instead of X)?

Measurement Error in the Independent Variable

The consequences depend on the correlation between E and X^* and or X.

Our model for our observed data is:

$$\begin{aligned} Y_i &= \beta_0 + \beta_1(X_i^* - E_i) + U_i \\ &= \beta_0 + \beta_1 X_i^* + (U_i - \beta_1 E_i) \end{aligned}$$

If the measurement error is uncorrelated with the observed variable $Cov[X^*, E] = 0$, then:

- since $(U_i - \beta_1 E_i)$ is uncorrelated with X_i^* if both E and U are uncorrelated with X_i^*, our estimator $\hat{\beta}_1$ is unbiased and consistent
- If E and U are uncorrelated, the overall error variance is $V[(U - \beta_1 E)] = \sigma_U^2 + \beta_1^2 \sigma_E^2$ which is $> \sigma_U^2$ unless $\beta_1 = 0$.
- Note: This has nothing to do with assumptions about errors U , we always maintain that $Cov[X^*, U] = 0$ and $Cov[X, U] = 0$

Measurement Error in the Independent Variable

Our model for the observed data is:

$$
\begin{aligned}
Y_i &= \beta_0 + \beta_1(X_i^* - E_i) + U_i \\
&= \beta_0 + \beta_1 X_i^* + (U_i - \beta_1 E_i)
\end{aligned}
$$

If the measurement error is uncorrelated with the unobserved variable, $Cov[X, E] = 0$, then:

- the observed variable $X^* = X + E$ and the measurement error E must be correlated:

  $$Cov[X^*, E] = E[X^* E] = E[(X + E)E] = E[XE + E^2] = 0 + \sigma_E^2 = \sigma_E^2$$

  so the covariance between the observed measure X^* and the measurement error E is equal to the variance of E (classical error-in-variables (CEV) assumption).

- This correlation causes problems since now: $Cov[X^*, U - \beta_1 E] = -\beta_1 Cov[X^*, E] = -\beta_1 \sigma_E^2$ and since the composite error is correlated with the observed measure OLS will be biased and inconsistent.

- Can we know the direction of the (asymptotic) bias?

Classical Errors in Variables (Appendix)

$$\begin{aligned}\hat{\beta}_1 &\xrightarrow{p} \beta_1 + \frac{Cov[X^*, U - \beta_1 E]}{V[X^*]} \\ &= \beta_1 - \frac{\beta_1 \sigma_E^2}{\sigma_X^2 + \sigma_E^2} \\ &= \beta_1 \left(1 - \frac{\sigma_E^2}{\sigma_X^2 + \sigma_E^2}\right) \\ &= \beta_1 \left(\frac{\sigma_X^2}{\sigma_X^2 + \sigma_E^2}\right) \\ &= \beta_1 \left(\frac{\sigma_X^2}{\sigma_{X^*}^2}\right)\end{aligned}$$

Notice that given our CEV assumption $\left(\frac{\sigma_X^2}{\sigma_X^2+\sigma_E^2}\right) < 1$ so that the probability limit of β_1 is always closer to zero than β_1 (attenuation bias).

Bias is small if variance of observed measure σ_X^2 is large relative to variance of error term σ_E^2 (high signal to noise ratio).

Measurement Error in Multiple Independent Variables

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + u$$

- If only X_1 is measured with CEV type error (but X_2 and X_3 are correct), then β_1 exhibits attenuation bias, and β_2 and β_3 will be inconsistent and biased unless X_1 is uncorrelated with X_2 and X_3
- the direction and magnitude of bias in X_2 and X_3 are not easy to derive and often unclear
- If we have CEV measurement error in multiple Xs then the size and direction of biases are unclear.

What can we do about measurement error?

- Improve our measures (e.g. pilot tests for surveys)
- Triangulate several measures (e.g. a battery of survey questions about the same issue)
- Triangulate several studies (replicate experiments several times with different subjects and at different times)
- Rely on variables that are less prone to bias
- Randomized response, list experiments etc.
- Modeling based approaches (next semester)

Summary for Measurement Error

- Measurement error in the outcome
  - does not cause bias unless measurement error correlated with X variables
  - does reduce efficiency
- Measurement error in the explanatory variable
  - does not cause bias if measurement error is uncorrelated with observed, mis-measured X (but increases variance)
  - does lead to attenuation bias even if measurement error is uncorrelated with unobserved true X

Note: This is true only under fairly strong assumptions including mean zero measurement error.

Concluding Thoughts for the Day

- Regression rests on a number of assumptions
- Easy to test some of these and hard to test others.
- Always check your data!
- Don't let regression be a magic black box for you- understand why it is giving the answers it gives.

References

- Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane Jr, Michael C. Herron, and Henry E. Brady. "The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* (2001): 793-810.
- Lange, Peter, and Geoffrey Garrett. "The politics of growth: Strategic interaction and economic performance in the advanced industrial democracies, 19741980." *The Journal of Politics* 47, no. 03 (1985): 791-827.
- Jackman, Robert W. "The Politics of Economic Growth in the Industrial Democracies, 197480: Leftist Strength or North Sea Oil?." The *Journal of Politics* 49, no. 01 (1987): 242-256.

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

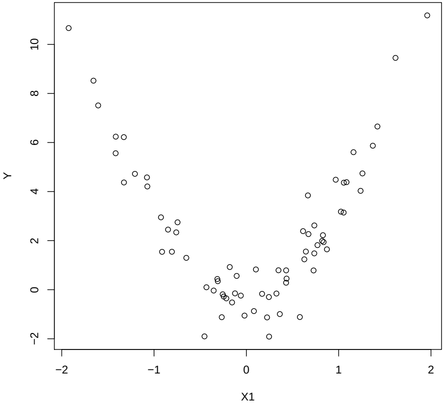
- Last Week
  - regression in the social sciences
- This "Week"
  - Monday (14):
    - ⋆ unusual and influential data → robust estimation
  - Wednesday (16):
    - ⋆ non-linearity → generalized additive models
  - Monday (21):
    - ⋆ unusual errors → sandwich SEs and block bootstrap
- After Thanksgiving
  - causality with measured confounding
- Long Run
  - regression → diagnostics → causal inference

Questions?

Residuals are still important. Look at them.

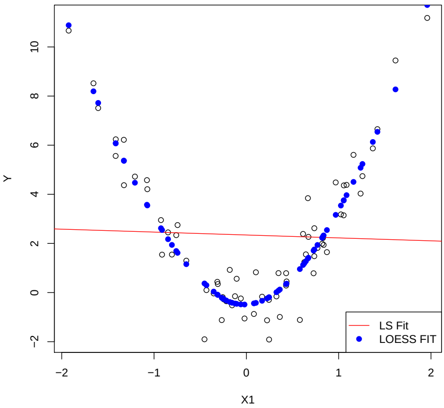
Nonlinearity

Linearity of the Conditional Expectation Function ($\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$) is a key assumption. Why?



Nonlinearity

Linearity of the Conditional Expectation Function ($\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$) is a key assumption. Why?



Nonlinearity

- If $E[Y|\mathbf{X}]$ is not linear in $\mathbf{X}$, $E[\mathbf{u}|\mathbf{X}] \neq 0$ for all values $\mathbf{X} = \mathbf{x}$ and $\hat{\beta}$ may be biased and inconsistent.
- Nonlinearities may be important but few social scientific theories offer any guidance as to functional form whatsoever.
  - Statements like "y increases with x" (monotonicity) are as specific as most social theories get.
  - Possible Exceptions: Returns to scale, constant elasticities, interactive effects, cyclical patterns in time series data, etc.
- Usually we employ "linearity by default" but we should try to make sure this is appropriate: detect non-linearities and model them accurately

Diagnosing Nonlinearity

- For marginal relationships Y and X
  - Scatterplots with loess lines
- For partial relationships Y and X_1, controlling for X_2, X_3,...,X_k the regression surface is high-dimensional. We need other diagnostic tools such as:
  - Added variables plots and component residual plots
  - Semi-parametric regression techniques like Generalized Additive Models (GAMs)
  - Non-parametric multiple regression techniques (beyond the scope of this course)

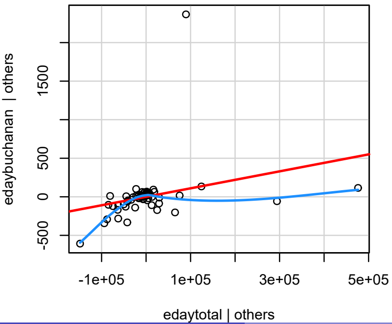
Added variable plot

- Need a way to visualize conditional relationship between Y and X_j
- How to construct an added variable plot:
  1. Get residuals from regression of Y on all covariates except X_j
  2. Get residuals from regression of X_j on all other covariates
  3. Plot residuals from (1) against residuals from (2)
- In R: `avPlots(model)` from the `car` package
- OLS fit to this plot will have exactly $\widehat{\beta}_j$ and 0 intercept (drawing on the partialing out interpretation we discussed before)
- Use local smoother (loess) to detect any non-linearity

Buchanan AV plot

R Code

```
par(mfrow = c(1,2))
out <- avPlots(mod3, "edaytotal")
lines(loess.smooth(x = out$edaytotal[,1],
 y= out$edaytotal[,2]), col = "dodgerblue", lwd = 2)
out2 <- avPlots(mod3, "absnbuchanan")
lines(loess.smooth(x = out2$absnbuchanan[,1],
y= out2$absnbuchanan[,2]), col = "dodgerblue", lwd = 2)
```



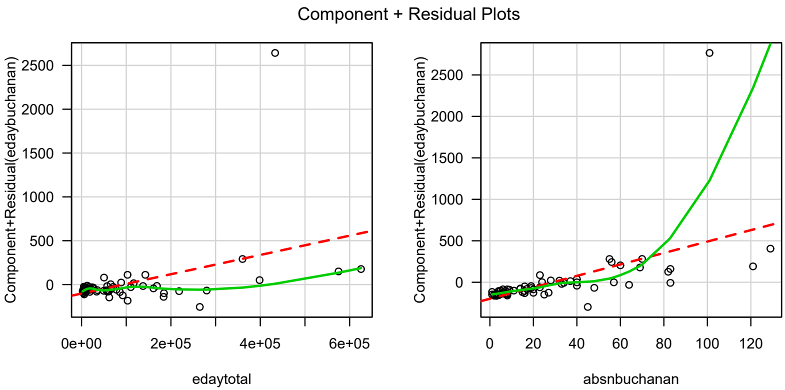
Component-Residual plots

- CR plots are a refinement of AV plots:
  1. Compute residuals from full regression:
  $$\widehat{u}_i = Y_i - \widehat{Y}_i$$
  2. Compute "linear component" of the partial relationship:
  $$C_i = \widehat{\beta}_j X_{ij}$$
  3. Add linear component to residual:
  $$\widehat{u}_i^j = \widehat{u}_i + C_i$$
  4. Plot partial residual $\widehat{u}_i^j$ against X_j
- Same slope as AV plots
- X-axis is the original scale of X_j, so slightly easier for diagnostics
- Use local smoother (loess) to detect non-linearity

Buchanan CR plot

R Code

```
crPlots(mod3, las = 1)
```



Limitations of CR Plots

- AV plots and CR plots can only reveal partial relationships
- Oftentimes, these two-dimensional displays fail to uncover structure in a higher-dimensional problem

- We may detect an interaction between X_1 and X_2 in a 3D scatterplot that we could miss in two scatterplots of Y on each X

- Cook (1993) shows that CR plots only work when either:
  1) The relationship between Y and X_j is linear
  2) Other explanatory variables $(X_1, ..., X_{j-1})$ are linearly related to X_j
  - This suggests that linearizing the relationship between the Xs through transformations can be helpful
  - Experience suggests weak non-linearities among Xs do not invalidate CR plots

How should we deal with nonlinearity?

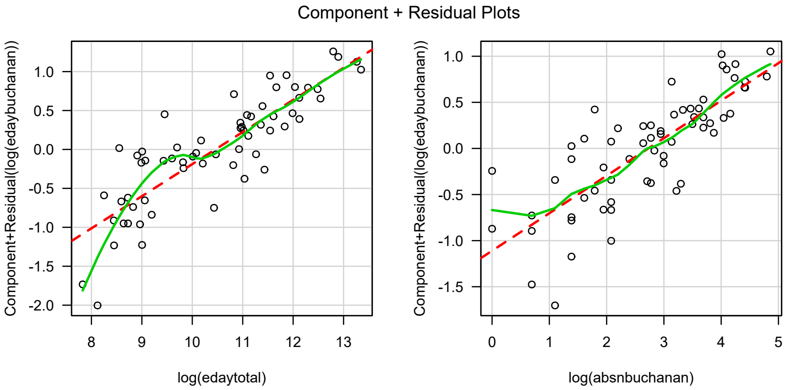
Given we have a linear regression model, our options are somewhat limited. However we can partially address nonlinearity by:

- Breaking categorical or continuous variables into dummy variables (e.g. education levels)
- Including interactions
- Including polynomial terms
- Transformations such as logs
- Generalized Additive Models (GAM)
- Many more flexible, nonlinear regression models exist beyond the scope of this course.

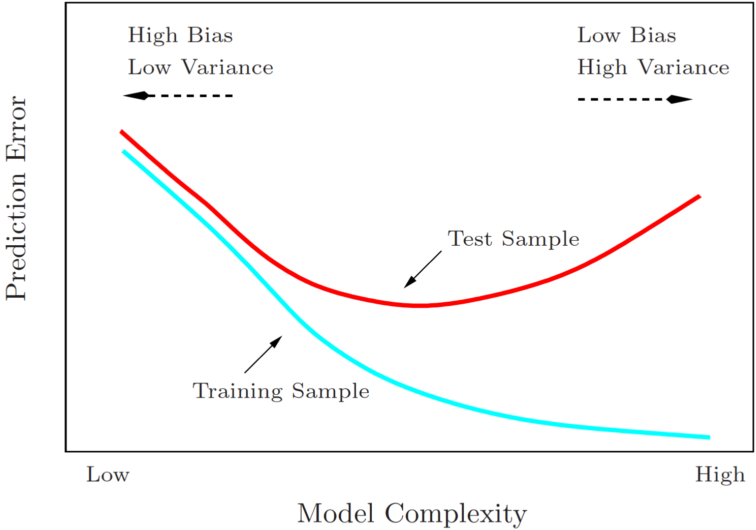
Transformed Buchanan regression

R Code

```
mod.nopb2 <- lm(log(edaybuchanan) ~ log(edaytotal) + log(absnbuchanan),
data = flvote, subset = county != "Palm Beach")
crPlots(mod.nopb2, las = 1)
```

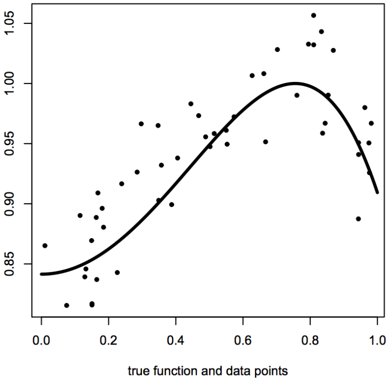


Bias-Variance Tradeoff



Example Synthetic Problem

$$y = \sin(1 + x^2) + \epsilon$$



This section adapted from slides by Radford Neal.

Linear Basis Function Models

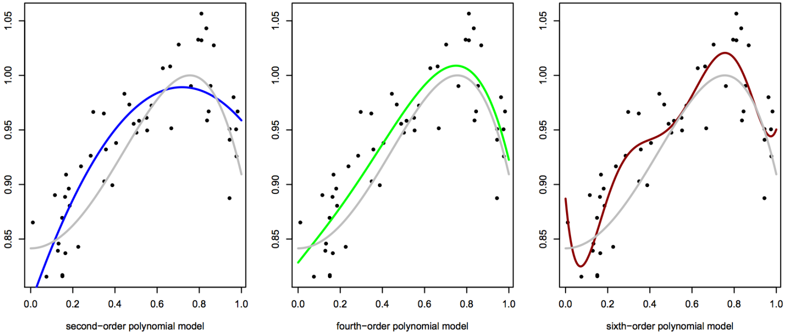
- We talked before about polynomials x^2, x^3, x^4 for modeling non-linearities, this is a linear basis function model.
- In general the idea is to do a linear regression of y on $\phi_1(x), \phi_2(x), \ldots, \phi_{m-1}(x)$ where ϕ_j are basis functions.
- The model is now:

$$y = f(x, \beta) + \epsilon$$

$$f(x, \beta) = \beta_0 + \sum_{j=1}^{m-1} \beta_j \phi_j(x) = \beta^T \phi(x)$$

Polynomial Basis Functions

We can look at OLS fits with polynomial basis functions of increasing order.



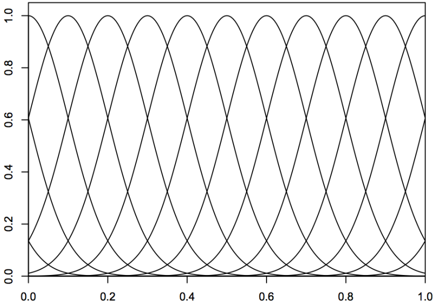
It appears that the last model is too complex and is overfitting a bit.

Local Basis Functions

Polynomials are global basis functions, each affecting the prediction over the whole input space. Often local basis functions are more appropriate.

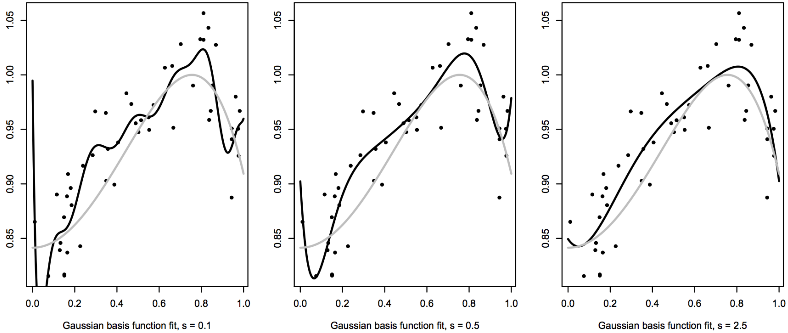
One choice is a Gaussian basis function

$$\phi_j(x) = \exp(-(x-\mu_j)^2)/2s^2)$$



Gaussian basis functions, s = 0.1

Gaussian Basis Fits

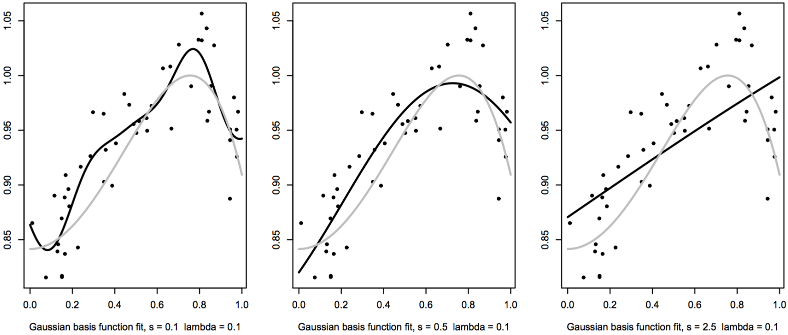


Regularization

- We've seen that flexible models can lead to overfitting
- Two ways to address: limit model flexibility or use a flexible model and regularize
- Regularization is a way of expressing a preference for smoothness in our function by adding a penalty term to our optimization function.
- Here we will consider a penalty of the form $\lambda \sum_{j=1}^{m-1} \beta_j^2$ where λ controls the strength of the penalty.
- The penalty trades off some bias for an improvement in variance
- The trick in general is how to set λ

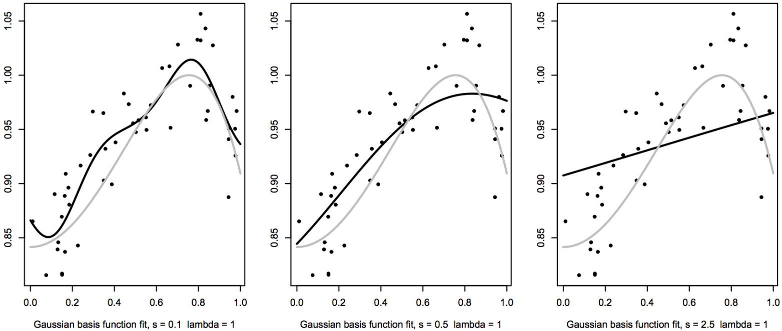
Results

Here are the results with $\lambda = 0.1$:



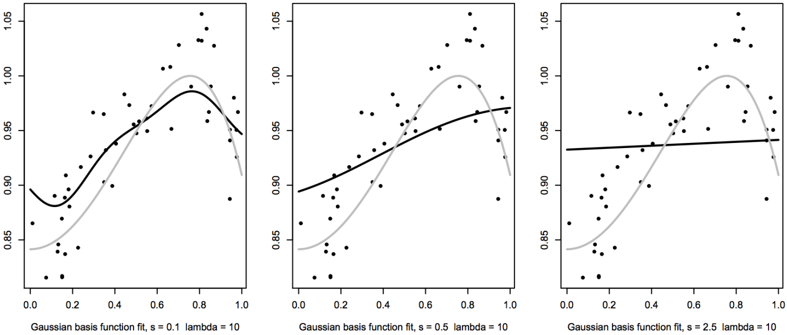
Results

Here are the results with $\lambda = 1$:



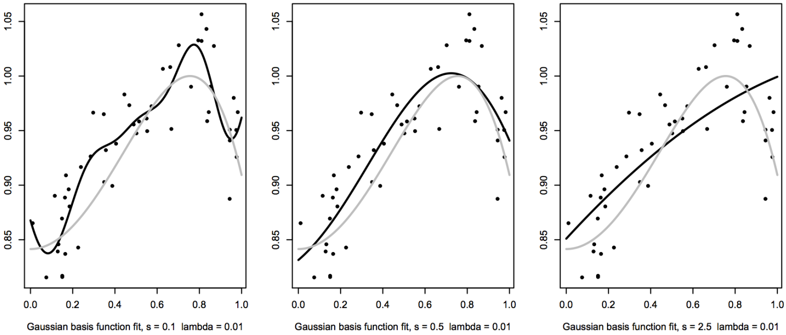
Results

Here are the results with $\lambda = 10$:



Results

Here are the results with $\lambda = 0.01$:



Conclusions from This Example

- we can control overfitting by modifying the width of the basis function s or with penalty
- we will need some way in general to tune these
- we will also need some way to handle multivariate functions.
- next up, Generalized Additive Models

Generalized Additive Models (GAM)

Recall the linear model,

$$y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + x_{3i}\beta_3 + u_i$$

For GAMs, we maintain additivity, but instead of imposing linearity we allow flexible functional forms for each explanatory variable, where $s_1(\cdot), s_2(\cdot)$, and $s_3(\cdot)$ are smooth functions that are estimated from the data:

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

Generalized Additive Models (GAM)

$$y_i = \beta_0 + s_1(x_{1i}) + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

- GAMS are semi-parametric, they strike a compromise between nonparametric methods and parametric regression
- $s_j(\cdot)$ are usually estimated with locally weighted regression smoothers or cubic smoothing splines (but many approaches are possible)
- They do NOT give you a set of regression parameters $\hat{\beta}$. Instead one obtains a graphical summary of how $E[Y|X_, X_2, ..., X_k]$ varies with X_1 (estimates of $s_j(\cdot)$ at every value of $X_{i,j}$)
- Theory and estimation are somewhat involved, but they are easy to use:
  - ```
    gam.out <- gam(y~s(x1)+s(x2)+x3)
    plot(gam.out)
    ```
  - Multiple functions but I recommend `mgcv` package
    ```

Generalized Additive Models (GAM)

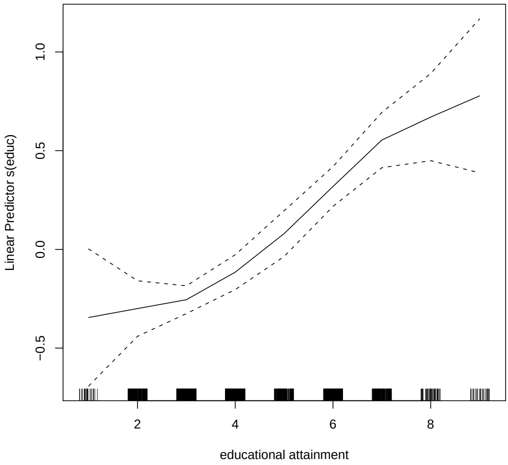
The GAM approach can be extended to allow interactions ($s_{12}(\cdot)$) between explanatory variables, but this eats up degrees of freedom so you need a lot of data.

$$y_i = \beta_0 + s_{12}(x_{1i}, x_{2i}) + s_3(x_{3i}) + u_i$$

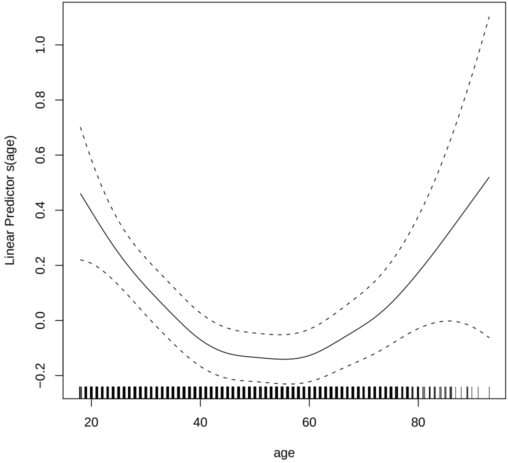
It can also be used for hybrid models where we model some variables as parametrically and other with a flexible function:

$$y_i = \beta_0 + \beta_1 x_{1i} + s_2(x_{2i}) + s_3(x_{3i}) + u_i$$

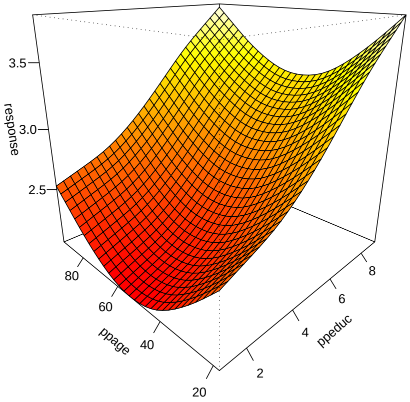
GAM Fit to Attitudes Toward Immigration



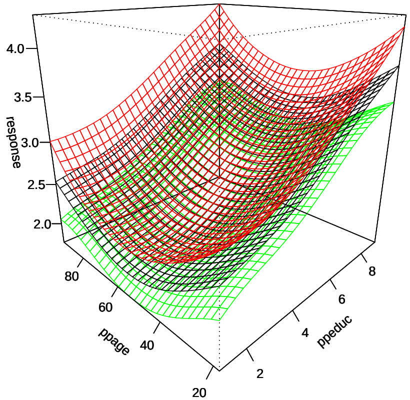
GAM Fit to Attitudes Toward Immigration



GAM Fit to Attitudes Toward Immigration

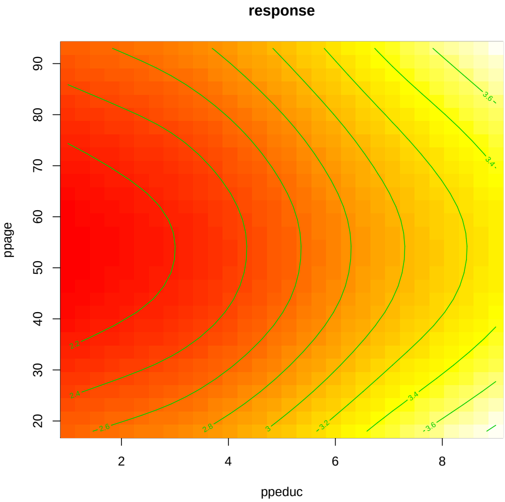


GAM Fit to Attitudes Toward Immigration

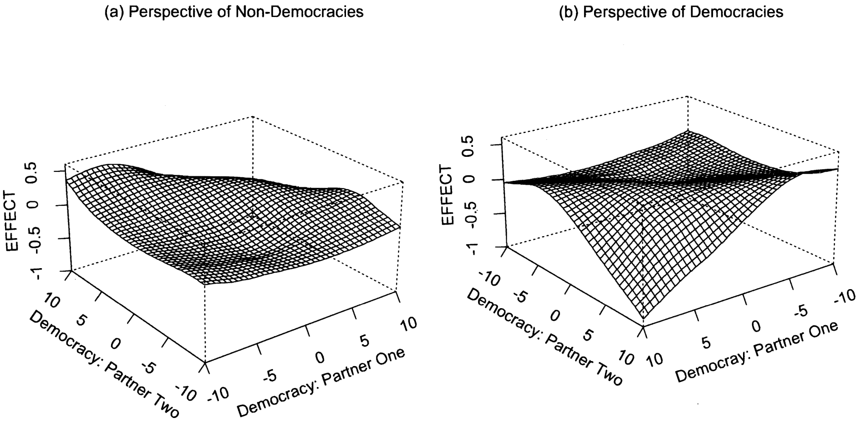


red/green are +/− 2 s.e.

GAM Fit to Attitudes Toward Immigration



GAM Fit to Dyadic Democracy and Militarized Disputes



Concluding Thoughts

- Non-linearity is pretty easy to detect and can substantially change our inferences
- GAMs are a great way to model/detect non-linearity but transformations are often simpler
- However, be wary of the global properties of transformations and polynomials
- Non-linearity concerns are most relevant for continuous covariates with a large range (age)

Fun With Kernels

Hainmueller and Hazlett (2013). "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach" *Political Analysis*.[2]

[2] I thank Chad Hazlett for sharing many of the slides that follow

Motivation: Misspecification Bias

Consider a data generating process such as:

```
> # Predictors
> GDP = runif(500)
> Polity  = .5*GDP^2 + .2*runif(200)
>
> # True Model
> Stability = log(GDP)+rnorm(500)
```

Regressing Stability on polity and GDP:

```
> # OLS
> lm(Stability ~ Polity + GDP)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.3000     0.1039 -22.145  < 2e-16 ***
Polity       -3.1983     0.7613  -4.201 3.15e-05 ***
GDP           4.3443     0.4237  10.252  < 2e-16 ***
```

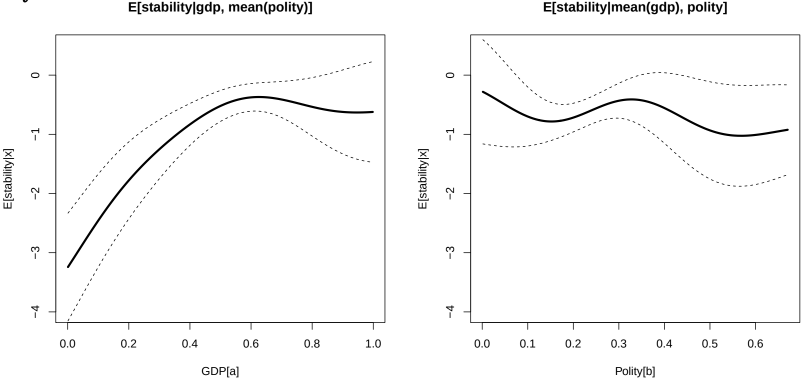
Entirely wrong conclusions!

Misspecification Bias

Try more flexible method that still reports marginal effects:

```
> krls(y=Stability,X=cbind(GDP,Polity))

Average Marginal Effects:
              Est Std. Error    t value     Pr(>|t|)
GDP      3.3855912  0.5217110  6.4893996 2.084441e-10
Polity -0.4143114  0.7826758 -0.5293525 5.967968e-01
```



Kernel Basics

Kernel

For now, a kernel is a function $\mathbb{R}^{\mathbb{P}} \times \mathbb{R}^{\mathbb{P}} \rightarrow \mathbb{R}$

$$k(x_i, x_j) \rightarrow \mathbb{R}$$

Some kernels are naturally interpretable as a distance metric, e.g. the Gaussian:

Gaussian Kernel

$$k(\cdot, \cdot) : \mathbb{R}^D \times \mathbb{R}^P \mapsto \mathbb{R}$$

$$k(x_j, x_i) = e^{-\frac{||x_j - x_i||^2}{\sigma^2}}$$

where $||X_j - X_i||$ is the Euclidean distance between X_j and X_i

Using the Kernel Trick for Regression

- A feature map, $\phi : \mathbb{R}^P \mapsto \mathbb{R}^{P'}$, such that: $k(X_i, X_j) = \langle \phi(X_i), \phi(X_j) \rangle$
- A linear model in the new features: $f(X_i) = \phi(X_i)^T \theta$, $\theta \in \mathbb{R}^{P'}$
- Regularized (ridge) regression:

$$\underset{\theta \in \mathbb{R}^{P'}}{\operatorname{argmin}} \sum_{i=1}^{N} (Y_i - \phi(X_i)^T \theta)^2 + \lambda \langle \theta, \theta \rangle$$

- Solve the F.O.C.s:

$$R(\theta, \lambda) = \sum_{i=1}^{N} (Y_i - \phi(X_i)^\top \theta)^2 + \lambda \theta^\top \theta$$

$$\frac{\partial R(\theta, \lambda)}{\partial \theta} = -2 \sum_{i=1}^{N} \phi(X_i)(Y_i - \phi(X_i)^T \theta) + 2\lambda\theta = 0$$

How would humans learn this?

Linear regression?

$$E[alt|lat, long] = \beta_0 + \beta_1 lat + \beta_2 long + \beta_3 lat \times long + \ldots$$

Similarity model:

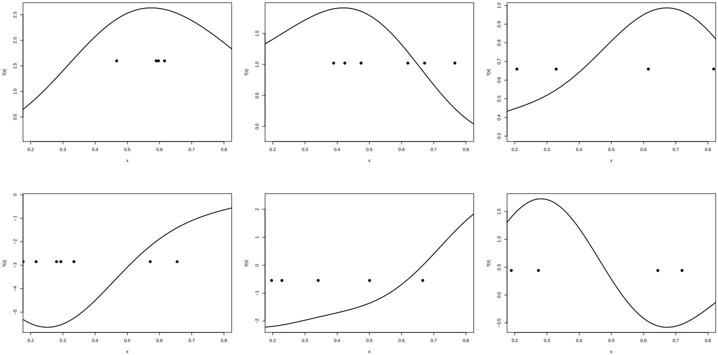
$$E[alt|lat, long] = c_1(\text{similarity to obs1}) + \ldots + c_5(\text{similarity to obs5})$$

Intuition: Similarity

Think of this function space as built on similarity:

$$f(X^\star) = \sum_{i=1}^{N} c_i k(X^\star, X_i)$$
$$= c_1(\text{similarity of } X^\star \text{ to } X_1) + \ldots + c_N(\text{similarity of } X^\star \text{ to } X_N)$$

Some random functions from this space:



A real example: Harff 2003

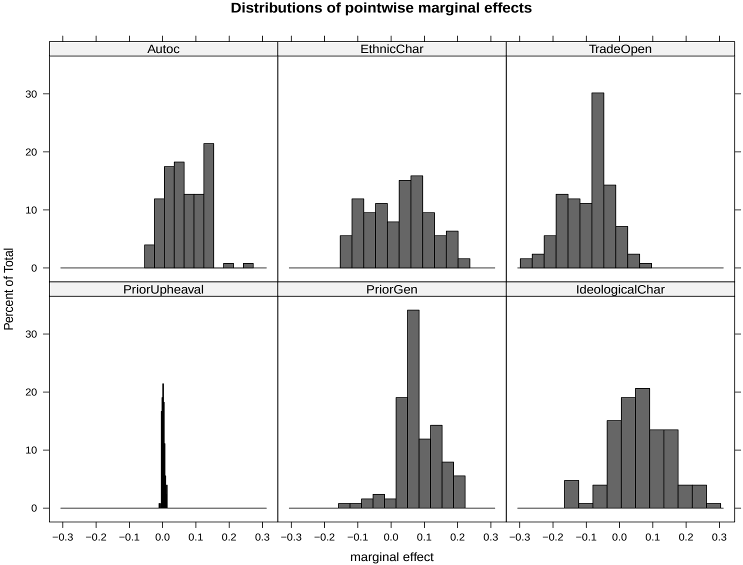
From `summary(krls(y,X))`

DV: Genocide onset

	β_{OLS}	$E[\frac{\hat{dy}}{dx_i}]$
Prior upheaval	0.009*	0.00
	(0.004)	0.00
Prior genocide	0.26*	0.19*
	(0.12)	(0.08)
Ideological char. elite	0.15*	0.13*
	(0.084)	(0.08)
Autocracy	0.16*	0.12*
	(0.077)	(0.07)
Ethnic char. elite	0.12	0.05
	(0.084)	(0.08)
log(trade openness)	-0.17*	-0.09*
	(0.057)	(0.03)

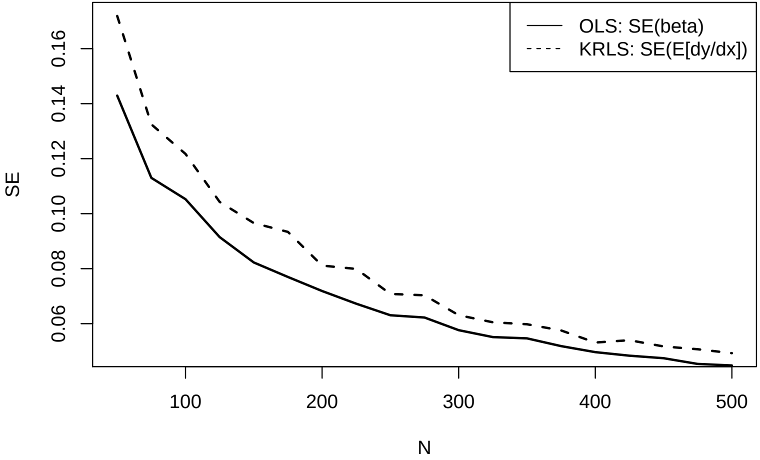
Behind the averages

`plot(krls(X,y))`

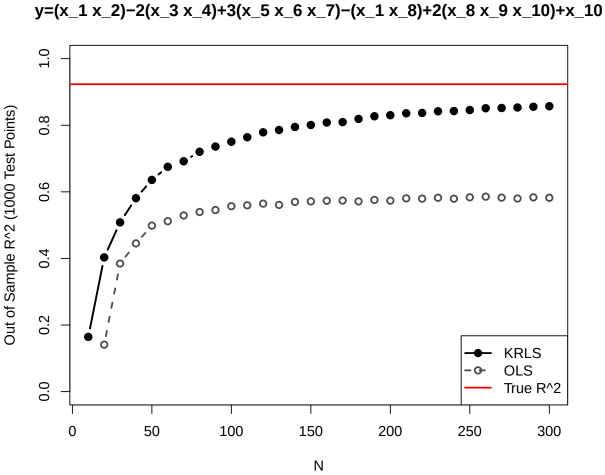


Efficiency Comparison

$$y = 2x + \epsilon,\ x \sim N(0,1), \epsilon \sim N(0,1)$$



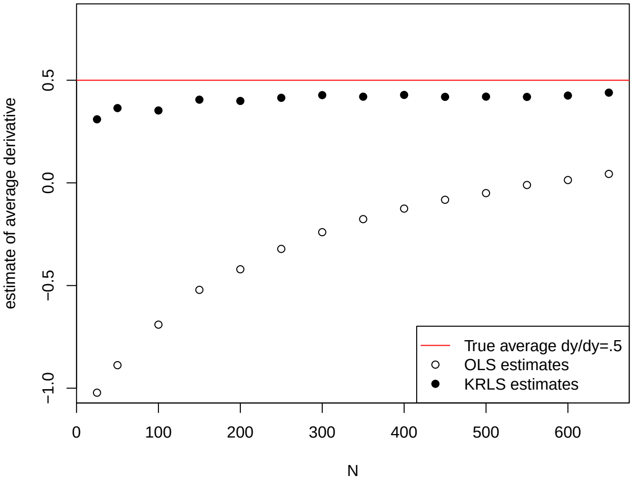
High-dimensional data with non-linearities



$y = (X_1X_2) - 2(X_3X_4) + 3(X_5X_6X_7) - (X_1X_8) + 2(X_8X_9X_{10}) + X_{10} + \epsilon$ where all X are i.i.d. *Bernoulli*(p) at varying p, $\varepsilon \sim N(0, .5)$. 1,000 test points.

Linear model with bad leverage points

- $y = .5x + \varepsilon$ where $\varepsilon \sim N(0, .3)$
- One bad point, $(y_i = -5, x_i = 5)$.



Interaction or non-linearity?

Truth: $y = 5x_1^2 + \varepsilon, \quad \rho(x_1, x_2) = .72$
$\varepsilon \sim (0, .44)$. $x_1 \sim Uniform(0, 2)$

OLS Model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 * x_2$

KRLS Model: krls(y,[x_1 x_2])

Estimator	OLS	KRLS			
$\partial y/\partial x_{ij}$	Average	Average	1st Qu.	Median	3rd Qu.
const	-1.50				
	(0.34)				
x_1	7.51	9.22	5.22	9.38	14.03
	(0.40)	(0.52)	(0.82)	(0.85)	(0.79)
x_2	-1.28	0.02	-0.08	0.00	0.10
	(0.21)	(0.13)	(0.19)	(0.16)	(0.20)
$(x_1 \cdot x_2)$	1.24				
	(0.15)				
N	250				

Concluding Thoughts

- Strengths
  - extremely powerful at detecting interactions
  - captures increasingly complex functions as data increases
  - great as a robustness check
- Difficulties/Future Work
  - computation scales in number of datapoints ($O(N^3)$) which means it doesn't work for more than about 5000 datapoints
  - it may model deep interactions but it is still hard to summarize deep interactions

References

- Hainmueller and Hazlett (2013). "Kernel Regularized Least Squares: Reducing Misspecification Bias with a Flexible and Interpretable Machine Learning Approach" *Political Analysis*.
- Beck, N. and Jackman, S. 1998. Beyond Linearity by Default: Generalized Additive Models. *American Journal of Political Science*.
- Wood (2003). "Thin plate regression splines." Journal of the Royal Statistical Society: Series B.
- Hastie, T.J. and Tibshirani, R.J. 1990. *General Additive Models*.
- Hastie, Tibshirani, and Friedman (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.
- Schölkopf and Smola (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. Cambridge, MA: MIT Press.

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

- Last Week
  - regression in the social sciences
- This "Week"
  - Monday (14):
    - unusual and influential data $\rightarrow$ robust estimation
  - Wednesday (16):
    - non-linearity $\rightarrow$ generalized additive models
  - Monday (21):
    - unusual errors $\rightarrow$ sandwich SEs and block bootstrap
- After Thanksgiving
  - causality with measured confounding
- Long Run
  - regression $\rightarrow$ diagnostics $\rightarrow$ causal inference

Questions?

A Quick Note of Thanks

Review of the OLS Assumptions

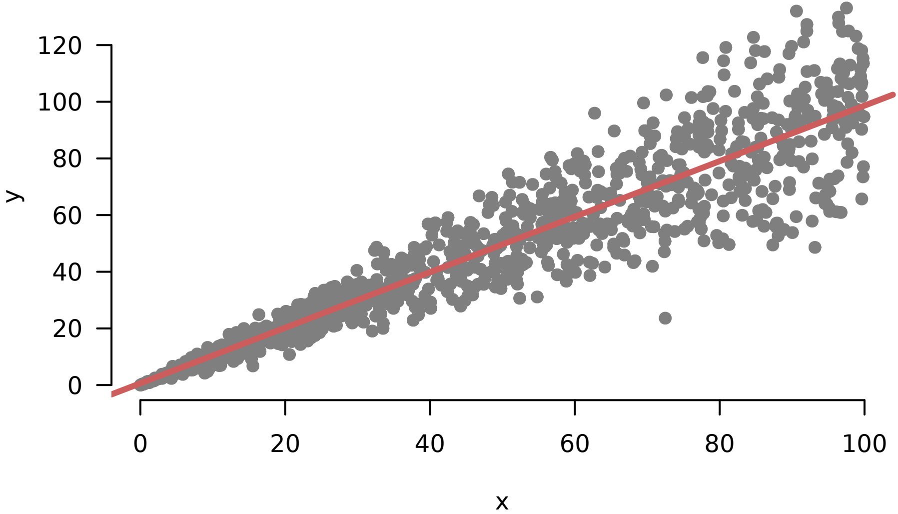
- **Identification / Data Description**
  - No Perfect Collinearity
- **Unbiasedness / Consistency**
  - No Perfect Collinearity
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
- **Gauss-Markov (BLUE) / Asymptotic Inference (z and χ^2)**
  - No Perfect Collinearity
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
- **Classical LM (BUE) / Small-Sample Inference (t and F)**
  - No Perfect Collinearity
  - Random Sampling
  - Linearity in Parameters
  - Zero Conditional Mean
  - Homoskedasticity
  - Normality of Errors

Review of the OLS Assumptions

1. Linearity: $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$
2. Random/iid sample: $(y_i, \mathbf{x}_i')$ are a iid sample from the population.
3. No perfect collinearity: $\mathbf{X}$ is an $n \times (K+1)$ matrix with rank $K+1$
4. Zero conditional mean: $\mathbb{E}[\mathbf{u}|\mathbf{X}] = \mathbf{0}$
5. Homoskedasticity: $\text{var}(\mathbf{u}|\mathbf{X}) = \sigma_u^2 \mathbf{I}_n$
6. Normality: $\mathbf{u}|\mathbf{X} \sim N(\mathbf{0}, \sigma_u^2 \mathbf{I}_n)$

- 1-4 give us unbiasedness/consistency
- 1-5 are the Gauss-Markov, allow for large-sample inference
- 1-6 allow for small-sample inference

How Do We Deal With This?



Plan for Today

Talk about different forms of error variance problems

1. Heteroskedasticity
2. Clustering
3. Optional: Serial correlation

Each is a violation of homoskedasticity, but each has its own diagnostics and corrections.

Then we will discuss a contrarian view

Review of Homoskedasticity

- Remember:

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

- Let $\text{Var}[\mathbf{u}|\mathbf{X}] = \boldsymbol{\Sigma}$
- Using assumptions 1 and 4, we can show that we have the following (derivation in the appendix):

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

- With homoskedasticity, $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$

$$\begin{aligned}\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \text{ (by homoskedasticity)} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

- Replace σ^2 with estimate $\widehat{\sigma}^2$ will give us our estimate of the covariance matrix

Non-constant Error Variance

- Homoskedastic:

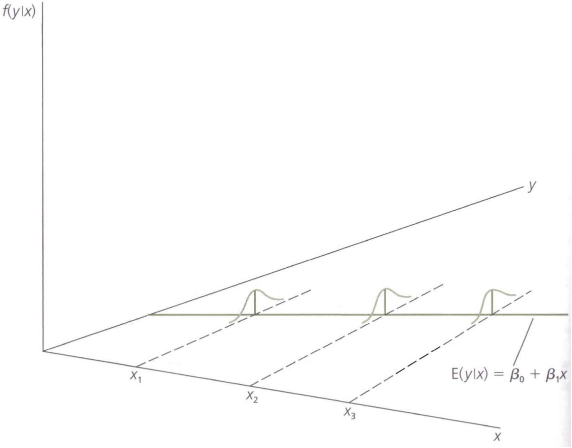
$$V[\mathbf{u}|\mathbf{X}] = \sigma^2\mathbf{I} = \begin{bmatrix} \sigma^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma^2 \end{bmatrix}$$

- Heteroskedastic:

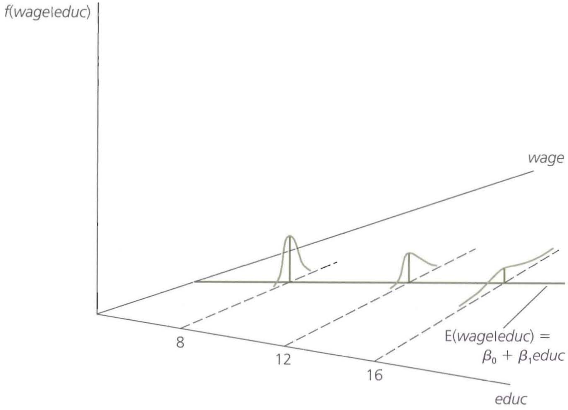
$$V[\mathbf{u}|\mathbf{X}] = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

- Independent, not identical
- $\text{Cov}(u_i, u_j|\mathbf{X}) = 0$
- $\text{Var}(u_i|\mathbf{X}) = \sigma_i^2$

Example: $V[\mathbf{u}|\mathbf{X}] = \sigma^2$ Homoskedasticity



Example: $V[\mathbf{u}|\mathbf{X}] = \sigma_i^2$ Heteroskedasticity



Consequences of Heteroskedasticity

- Standard $\widehat{\sigma}^2$ is biased and inconsistent for σ^2
- Standard error estimates incorrect:

$$\widehat{SE}[\widehat{\beta}_1] = \frac{\widehat{\sigma}^2}{\sum_i (X_i - \overline{X})^2}$$

- Test statistics won't have t or F distributions
- α-level tests, the probability of Type I error $\neq \alpha$
- Coverage of $1 - \alpha$ CIs $\neq 1 - \alpha$
- OLS is not BLUE
- However:
    - $\widehat{\boldsymbol{\beta}}$ still unbiased and consistent for $\boldsymbol{\beta}$
    - degree of the problem depends on how serious the heteroskedasticity is

Visual diagnostics

1. Plot of residuals versus fitted values
   - In R, `plot(mod, which = 1)`
2. Spread location plots
   - y-axis: Square-root of the absolute value of the residuals (folds the plot in half)
   - x-axis: Fitted values
   - Usually has loess trend curve to check if variance varies with fitted values
   - In R, `plot(mod, which = 3)`

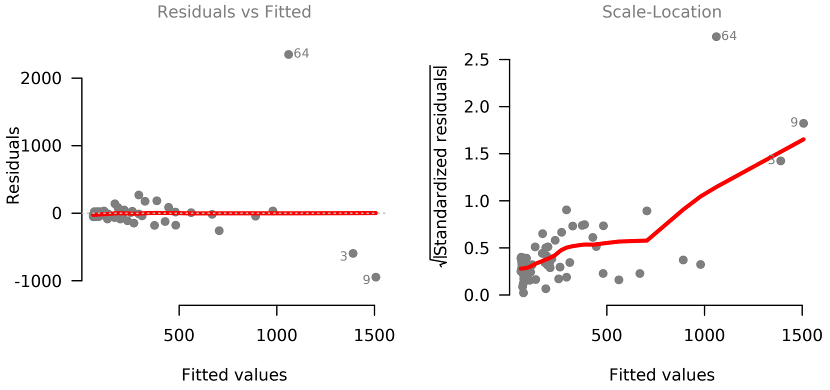
Example: Buchanan votes

```
flvote <- foreign::read.dta("flbuchan.dta")
mod <- lm(edaybuchanan ~ edaytotal, data = flvote)
summary(mod)

##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 5.423e+01  4.914e+01   1.104    0.274
## edaytotal   2.323e-03  3.104e-04   7.483 2.42e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 332.7 on 65 degrees of freedom
## Multiple R-squared:  0.4628, Adjusted R-squared:  0.4545
## F-statistic:    56 on 1 and 65 DF,  p-value: 2.417e-10
```

Diagnostics

```
par(mfrow = c(1,2), pch = 19, las = 1, col = "grey50", bty = "n")
plot(mod, which = 1, lwd = 3)
plot(mod, which = 3, lwd = 3)
```



Formal Tests for Non-constant Error Variances

- Plots are usually sufficient, but formal tests for heteroskedasticity exist (e.g. Breusch-Pagan, Cook-Weisberg, White, etc.).
- They are all roughly based on the same idea:
  - H_0: $V[u_i|\mathbf{X}] = \sigma^2$
  - Under the zero conditional mean assumption, this is equivalent to H_0: $E[u_i^2|\mathbf{X}] = E[u_i^2] = \sigma^2$, a constant unrelated to $\mathbf{X}$
  - This implies that, under H_0, the squared residuals should also be unrelated to the explanatory variables
- The Breusch-Pagan test:
  1. Regression y_i on $\mathbf{x}_i'$ and store residuals, $\widehat{u}_i$
  2. Regress $\widehat{u}_i^2$ on $\mathbf{x}_i'$
  3. Run F-test against null that all slope coefficients are 0

  - In R, `bptest` in the `lmtest` package

Breush-Pagan Example

```
library(lmtest)
bptest(mod)

##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 12.59, df = 1, p-value = 0.0003878
```

Dealing with Non-Constant Error Variance

1. Transform the dependent variable
   (this will affect other model assumptions)
2. Adjust for the heteroskedasticity using known weights and Weighted Least Squares (WLS)
3. Use an estimator of $\text{Var}[\widehat{\beta}]$ that is robust to heteroskedasticity
4. Admit we have the wrong model and use a different approach

Variance Stabilizing Transformations

If the variance for each error (σ_i^2) is proportional to some function of the mean ($\mathbf{x}_i\boldsymbol{\beta}$), then a variance stabilizing transformation may be appropriate.

Note: Transformations will affect the other regression assumptions, as well as interpretation of the regression coefficients.

Examples:

Transformation	Mean/Variance Relationship
$\sqrt{Y}$	$\sigma_i^2 \propto \mathbf{x}_i\boldsymbol{\beta}$
$\log Y$	$\sigma_i^2 \propto (\mathbf{x}_i\boldsymbol{\beta})^2$
$1/Y$	$\sigma_i^2 \propto (\mathbf{x}_i\boldsymbol{\beta})^4$

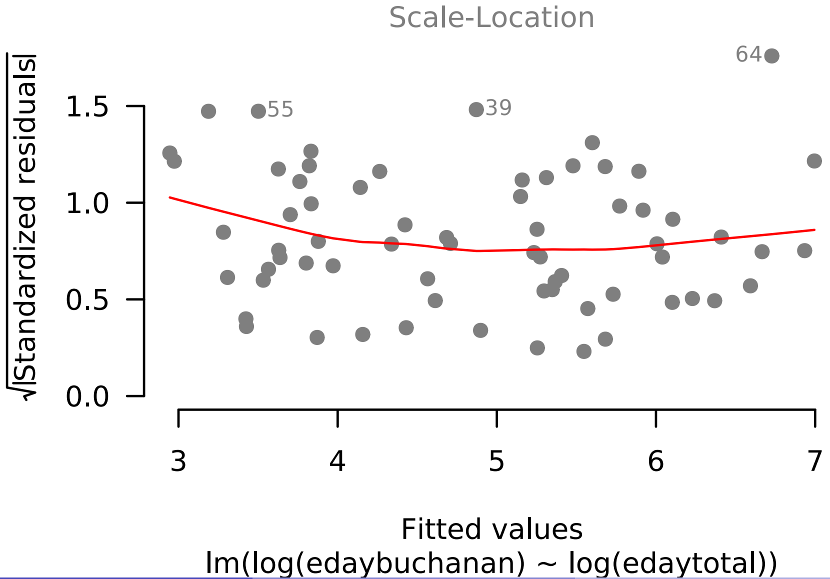
Example: Transforming Buchanan Votes

```
mod2 <- lm(log(edaybuchanan) ~ log(edaytotal), data = flvote)
summary(mod2)

##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2.72789    0.39956  -6.827  3.5e-09 ***
## log(edaytotal)  0.72853    0.03803  19.154  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4688 on 65 degrees of freedom
## Multiple R-squared:  0.8495, Adjusted R-squared:  0.8472
## F-statistic: 366.9 on 1 and 65 DF,  p-value: < 2.2e-16
```

Example: Transformed Scale-Location Plot

```
plot(mod2, which=3)
```



Example: Transformed

```
bptest(mod, studentize=FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  mod
## BP = 250.07, df = 1, p-value < 2.2e-16
```

```
bptest(mod2, studentize=FALSE)
```

```
##
##  Breusch-Pagan test
##
## data:  mod2
## BP = 0.01105, df = 1, p-value = 0.9163
```

Appendix: Weighted Least Squares

- Suppose that the heteroskedasticity is known up to a multiplicative constant:

$$\text{Var}[u_i|\mathbf{X}] = a_i\sigma^2$$

where $a_i = a_i(\mathbf{x}_i')$ is a positive and known function of $\mathbf{x}_i'$

- WLS: multiply y_i by $1/\sqrt{a_i}$:

$$y_i/\sqrt{a_i} = \beta_0/\sqrt{a_i} + \beta_1 x_{i1}/\sqrt{a_i} + \cdots + \beta_k x_{ik}/\sqrt{a_i} + u_i/\sqrt{a_i}$$

Appendix: Weighted Least Squares Intuition

- Rescales errors to $u_i/\sqrt{a_i}$, which maintains zero mean error
- But makes the error variance constant again:

$$\begin{aligned}\text{Var}\left[\frac{1}{\sqrt{a_i}}u_i|\mathbf{X}\right] &= \frac{1}{a_i}\text{Var}\left[u_i|\mathbf{X}\right] \\ &= \frac{1}{a_i}a_i\sigma^2 \\ &= \sigma^2\end{aligned}$$

- If you know a_i, then you can use this approach to makes the model homoskedastic and, thus, BLUE again
- When do we know a_i?

Appendix: Weighted Least Squares procedure

- Define the weighting matrix:

$$\mathbf{W} = \begin{bmatrix} 1/\sqrt{a_1} & 0 & 0 & 0 \\ 0 & 1/\sqrt{a_2} & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & 1/\sqrt{a_n} \end{bmatrix}$$

- Run the following regression:

$$\begin{aligned} \mathbf{Wy} &= \mathbf{WX}\beta + \mathbf{Wu} \\ \mathbf{y}^* &= \mathbf{X}^*\beta + \mathbf{u}^* \end{aligned}$$

- Run regression of $\mathbf{y}^* = \mathbf{Wy}$ on $\mathbf{X}^* = \mathbf{WX}$ and all Gauss-Markov assumptions are satisfied
- Plugging into the usual formula for $\widehat{\beta}$:

$$\widehat{\beta}_W = (\mathbf{X}'\mathbf{W}'\mathbf{WX})^{-1}\mathbf{X}'\mathbf{W}'\mathbf{Wy}$$

Appendix: WLS Example

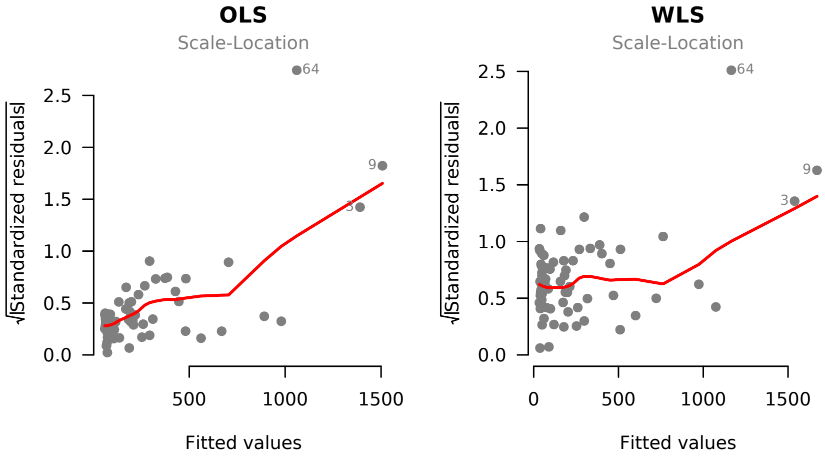
- In R, use `weights` = argument to `lm` and give the weights squared: $1/a_i$
- With the Buchanan data, maybe we think that the variance is proportional to the total number of ballots cast:

```
mod.wls <- lm(edaybuchanan ~ edaytotal, weights = 1/edaytotal,
                     data = flvote)
summary(mod.wls)

##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.707e+01  8.507e+00   3.182  0.00225 **
## edaytotal   2.628e-03  2.502e-04  10.503 1.22e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5645 on 65 degrees of freedom
## Multiple R-squared:  0.6292, Adjusted R-squared:  0.6235
## F-statistic: 110.3 on 1 and 65 DF,  p-value: 1.22e-15
```

Appendix: Comparing WLS to OLS

```
par(mfrow=c(1,2), pch = 19, las = 1, col = "grey50", bty = "n")
plot(mod, which = 3, main = "OLS", lwd = 2)
plot(mod.wls, which = 3, main = "WLS", lwd = 2)
```



Heteroskedasticity Consistent Estimator

- Under non-constant error variance:

$$\text{Var}[\mathbf{u}] = \mathbf{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \ldots & 0 \\ 0 & \sigma_2^2 & 0 & \ldots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \ldots & \sigma_n^2 \end{bmatrix}$$

- When $\mathbf{\Sigma} \neq \sigma^2 \mathbf{I}$, we are stuck with this expression:

$$\text{Var}[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\Sigma\mathbf{X} \left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- Idea: If we can consistently estimate the components of $\mathbf{\Sigma}$, we could directly use this expression by replacing $\mathbf{\Sigma}$ with its estimate, $\hat{\mathbf{\Sigma}}$.

White's Heteroskedasticity Consistent Estimator

Suppose we have heteroskedasticity of unknown form:

$$V[\mathbf{u}] = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \sigma_n^2 \end{bmatrix}$$

then $V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$ and White (1980) shows that

$$\widehat{V[\hat{\beta}|\mathbf{X}]} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \begin{bmatrix} \hat{\mathbf{u}}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{\mathbf{u}}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{\mathbf{u}}_n^2 \end{bmatrix} \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$$

is a consistent estimator of $V[\hat{\beta}|\mathbf{X}]$ under any form of heteroskedasticity consistent with $V[\mathbf{u}]$ above.

The estimate based on the above is called the heteroskedasticity consistent (HC) or robust standard errors.

White's Heteroskedasticity Consistent Estimator

Robust standard errors are easily computed with the "sandwich" formula:

1. Fit the regression and obtain the residuals $\hat{\mathbf{u}}$
2. Construct the "meat" matrix $\widehat{\boldsymbol{\Sigma}}$ with squared residuals in diagonal:

$$\widehat{\boldsymbol{\Sigma}} = \begin{bmatrix} \hat{\mathbf{u}}_1^2 & 0 & 0 & \dots & 0 \\ 0 & \hat{\mathbf{u}}_2^2 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & 0 & 0 & \dots & \hat{\mathbf{u}}_n^2 \end{bmatrix}$$

3. Plug $\widehat{\boldsymbol{\Sigma}}$ into the sandwich formula to obtain the robust estimator of the variance-covariance matrix

$$V[\hat{\beta}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- There are various small sample corrections to improve performance when sample size is small. The most common variant (sometimes labeled HC1) is:

$$V[\hat{\beta}|\mathbf{X}] \; = \; \frac{n}{n-k-1} \cdot \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Regular & Robust Standard Errors in Florida Example

R Code

```
> library(sandwich)
> library(lmtest)
> coeftest(mod1) # homoskedasticity
t test of coefficients:
              Estimate Std. Error t value  Pr(>|t|)
(Intercept)  5.4231e+01 4.9141e+01  1.1036    0.2738
TotalVotes00 2.3229e-03 3.1041e-04  7.4831 2.417e-10 ***

> coeftest(mod1,vcov = vcovHC(mod1, type = "HC0")) # classic White
t test of coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.4231e+01 4.0612e+01  1.3353  0.18642
TotalVotes00 2.3229e-03 8.7047e-04  2.6685  0.00961 **

> coeftest(mod1,vcov = vcovHC(mod1, type = "HC1")) # small sample correction
t test of coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.4231e+01 4.1232e+01  1.3153  0.19304
TotalVotes00 2.3229e-03 8.8376e-04  2.6284  0.01069 *
```

WLS vs. White's Estimator

- WLS:
  - With known weights, WLS is efficient
  - and $\widehat{SE}[\widehat{\beta}_{WLS}]$ is unbiased and consistent
  - but weights usually aren't known
- White's Estimator:
  - Doesn't change estimate $\widehat{\boldsymbol{\beta}}$
  - Consistent for Var$[\widehat{\beta}]$ under any form of heteroskedasticity
  - Because it relies on consistency, it is a large sample result, best with large n
  - For small n, performance might be poor (correction factors exist but are often insufficient)

Clustered Dependence: Intuition

- Think back to the Gerber, Green, and Larimer (2008) social pressure mailer example.
- Their design: randomly sample households and randomly assign them to different treatment conditions
- But the measurement of turnout is at the individual level
- Violation of iid/random sampling:
  - errors of individuals within the same household are correlated
  - $\rightsquigarrow$ violation of homoskedasticity
- Called clustering or clustered dependence

Clustered Dependence: notation

- Clusters: $j = 1, \ldots, m$
- Units: $i = 1, \ldots, n_j$
- n_j is the number of units in cluster j
- $n = \sum_j n_j$ is the total number of units
- Units (usually) belong to a single cluster:
  - voters in households
  - individuals in states
  - students in classes
  - rulings in judges
- Especially important when outcome varies at the unit-level, y_{ij} and the main independent variable varies at the cluster level, x_j.
- Ignoring clustering is "cheating": units not independent

Clustered Dependence: Example Model

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + \varepsilon_{ij} \\ &= \beta_0 + \beta_1 x_{ij} + v_j + u_{ij} \end{aligned}$$

- $v_j \overset{iid}{\sim} N(0, \rho\sigma^2)$ cluster error component
- $u_{ij} \overset{iid}{\sim} N(0, (1-\rho)\sigma^2)$ unit error component
- v_j and u_{ij} are assumed to be independent of each other
- $\rho \in (0, 1)$ is called the within-cluster correlation.
- What if we ignore this structure and just use ε_{ij} as the error?
- Variance of the composite error is σ^2:

$$\begin{aligned} \text{Var}[\varepsilon_{ij}] &= \text{Var}[v_j + u_{ij}] \\ &= \text{Var}[v_j] + \text{Var}[u_{ij}] \\ &= \rho\sigma^2 + (1-\rho)\sigma^2 = \sigma^2 \end{aligned}$$

Lack of Independence

- Covariance between two units i and s in the same cluster is $\rho\sigma^2$:

$$\text{Cov}[\varepsilon_{ij}, \varepsilon_{sj}] = \rho\sigma^2$$

- Correlation between units in the same group is just ρ:

$$\text{Cor}[\varepsilon_{ij}, \varepsilon_{sj}] = \rho$$

- Zero covariance of two units i and s in different clusters j and k:

$$\text{Cov}[\varepsilon_{ij}, \varepsilon_{sk}] = 0$$

Example Covariance Matrix

$$\boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_{1,1} & \varepsilon_{2,1} & \varepsilon_{3,1} & \varepsilon_{4,2} & \varepsilon_{5,2} & \varepsilon_{6,2} \end{bmatrix}'$$

$$\text{Var}[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma} = \begin{bmatrix} \sigma^2 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & 0 & 0 & 0 \\ \sigma^2 \cdot \rho & \sigma^2 & \sigma^2 \cdot \rho & 0 & 0 & 0 \\ \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho \\ 0 & 0 & 0 & \sigma^2 \cdot \rho & \sigma^2 & \sigma^2 \cdot \rho \\ 0 & 0 & 0 & \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \sigma^2 \end{bmatrix}$$

Appendix: Example 6 Units, 2 Clusters

$$\boldsymbol{\varepsilon} = [\ \varepsilon_{1,1} \quad \varepsilon_{2,1} \quad \varepsilon_{3,1} \quad \varepsilon_{4,2} \quad \varepsilon_{5,2} \quad \varepsilon_{6,2}\ ]'$$

$$V[\boldsymbol{\varepsilon}] = \boldsymbol{\Sigma} = \begin{bmatrix} V[\varepsilon_{1,1}] & Cov[\varepsilon_{2,1},\varepsilon_{1,1}] & Cov[\varepsilon_{3,1},\varepsilon_{1,1}] & . & . & . \\ Cov[\varepsilon_{1,1},\varepsilon_{2,1}] & V[\varepsilon_{2,1}] & Cov[\varepsilon_{3,1},\varepsilon_{2,1}] & . & . & . \\ Cov[\varepsilon_{1,1},\varepsilon_{3,1}] & Cov[\varepsilon_{2,1},\varepsilon_{3,1}] & V[\varepsilon_{3,1}] & . & . & . \\ Cov[\varepsilon_{1,1},\varepsilon_{4,2}] & Cov[\varepsilon_{2,1},\varepsilon_{4,2}] & Cov[\varepsilon_{3,1},\varepsilon_{4,2}] & V[\varepsilon_{4,2}] & . & . \\ Cov[\varepsilon_{1,1},\varepsilon_{5,2}] & Cov[\varepsilon_{2,1},\varepsilon_{5,2}] & Cov[\varepsilon_{3,1},\varepsilon_{5,2}] & Cov[\varepsilon_{4,2},\varepsilon_{5,2}] & V[\varepsilon_{5,2}] & . \\ Cov[\varepsilon_{1,1},\varepsilon_{6,2}] & Cov[\varepsilon_{2,1},\varepsilon_{6,2}] & Cov[\varepsilon_{3,1},\varepsilon_{6,2}] & Cov[\varepsilon_{4,2},\varepsilon_{6,2}] & Cov[\varepsilon_{5,2},\varepsilon_{6,2}] & V[\varepsilon_{6,2}] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma^2 & \sigma^2\cdot\rho & \sigma^2\cdot\rho & 0 & 0 & 0 \\ \sigma^2\cdot\rho & \sigma^2 & \sigma^2\cdot\rho & 0 & 0 & 0 \\ \sigma^2\cdot\rho & \sigma^2\cdot\rho & \sigma^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma^2 & \sigma^2\cdot\rho & \sigma^2\cdot\rho \\ 0 & 0 & 0 & \sigma^2\cdot\rho & \sigma^2 & \sigma^2\cdot\rho \\ 0 & 0 & 0 & \sigma^2\cdot\rho & \sigma^2\cdot\rho & \sigma^2 \end{bmatrix}$$

which can be verified as follows:

- $V[\varepsilon_{ij}] = V[v_j + u_{ij}] = V[v_j] + V[u_{ij}] = \rho\sigma^2 + (1-\rho)\sigma^2 = \sigma^2$
- $Cov[\varepsilon_{ij},\varepsilon_{lj}] = E[\varepsilon_{ij}\varepsilon_{lj}] - E[\varepsilon_{ij}]E[\varepsilon_{lj}] = E[\varepsilon_{ij}\varepsilon_{lj}] = E[(v_j + u_{ij})(v_j + u_{lj})]$
  $= E[{v_j}^2] + E[v_j u_{ij}] + E[v_j u_{lj}] + E[u_{ij} u_{lj}]$
  $= E[{v_j}^2] + E[v_j]E[u_{ij}] + E[v_j]E[u_{lj}] + E[u_{ij}]E[u_{lj}]$
  $= E[{v_j}^2] = V[v_j] + (E[v_j])^2 = V[v_j] = \rho\sigma^2$
- $Cov[\varepsilon_{ij},\varepsilon_{lk}] = E[\varepsilon_{ij}\varepsilon_{lk}] - E[\varepsilon_{ij}]E[\varepsilon_{lk}] = E[\varepsilon_{ij}\varepsilon_{lk}] = E[(v_j + u_{ij})(v_k + u_{lk})]$
  $= E[v_j v_k] + E[v_j u_{lk}] + E[v_k u_{ij}] + E[u_{ij} u_{lk}]$
  $= E[v_j]E[v_k] + E[v_j]E[u_{lk}] + E[v_k]E[u_{ij}] + E[u_{ij}]E[u_{lk}] = 0$

Error Variance Matrix with Cluster Dependence

The variance-covariance matrix of the error, $\boldsymbol{\Sigma}$, is block diagonal:

- By independence, the errors are uncorrelated across clusters:

$$V[\varepsilon] = \boldsymbol{\Sigma} = \left[\begin{array}{c|c|c|c} \boldsymbol{\Sigma}_1 & 0 & \dots & 0 \\ \hline \mathbf{0} & \boldsymbol{\Sigma}_2 & \dots & \mathbf{0} \\ \hline & & \ddots & \\ \hline \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_M \end{array}\right]$$

- But the errors may be correlated for units within the same cluster:

$$\boldsymbol{\Sigma}_j = \begin{bmatrix} \sigma^2 & \sigma^2 \cdot \rho & \dots & \sigma^2 \cdot \rho \\ \sigma^2 \cdot \rho & \sigma^2 & \dots & \sigma^2 \cdot \rho \\ & & \ddots & \\ \sigma^2 \cdot \rho & \sigma^2 \cdot \rho & \dots & \sigma^2 \end{bmatrix}$$

Correcting for Clustering

1. Including a dummy variable for each cluster (fixed effects)
2. "Random effects" models (take above model as true and estimate ρ and σ^2)
3. Cluster-robust ("clustered") standard errors
4. Aggregate data to the cluster-level and use OLS $\overline{y}_j = \frac{1}{n_j}\sum_i y_{ij}$
    - If n_j varies by cluster, then cluster-level errors will have heteroskedasticity
    - Can use WLS with cluster size as the weights

Cluster-Robust SEs

- First, let's write the within-cluster regressions like so:

$$\mathbf{y}_j = \mathbf{X}_j\boldsymbol{\beta} + \boldsymbol{\varepsilon}_j$$

- $\mathbf{y}_j$ is the vector of responses for cluster j, and so on
- We assume that respondents are independent across clusters, but possibly dependent within clusters. Thus, we have

$$\text{Var}[\boldsymbol{\varepsilon}_j|\mathbf{X}_j] = \boldsymbol{\Sigma}_j$$

- Remember our sandwich expression:

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\boldsymbol{\Sigma}\mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

- Under this clustered dependence, we can write this as:

$$\text{Var}[\hat{\boldsymbol{\beta}}|\mathbf{X}] = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\left(\sum_{j=1}^{m}\mathbf{X}_j'\boldsymbol{\Sigma}_j\mathbf{X}_j\right)\left(\mathbf{X}'\mathbf{X}\right)^{-1}$$

Estimating the Variance Components: ρ and σ^2

The overall error variance σ^2 is easily estimated using our usual estimator based on the regression residuals: $\widehat{\sigma}^2 = \frac{\hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}}{N-k-1}$

The within-cluster correlation can be estimated as follows:

1. Subtract from each residual $\hat{\varepsilon}_{ij}$ the mean residual within its cluster. Call this vector of demeaned residuals $\tilde{\boldsymbol{\varepsilon}}$, which estimates the unit error component $\boldsymbol{u}$
2. Compute the variance of the demeaned residuals as: $\widehat{\tilde{\sigma}}^2 = \frac{\tilde{\boldsymbol{\varepsilon}}'\tilde{\boldsymbol{\varepsilon}}}{N-M-k-1}$, which estimates $(1-\rho)\sigma^2$
3. The within cluster correlation is then estimated as: $\hat{\rho} = \frac{\widehat{\sigma}^2 - \widehat{\tilde{\sigma}}^2}{\widehat{\sigma}^2}$

Estimating Cluster Robust Standard Errors

We can now compute the CRSEs using our sandwich formula:

1. Take your estimates of $\widehat{\sigma^2}$ and $\widehat{\rho}$ and construct the block diagonal variance-covariance matrix $\widehat{\boldsymbol{\Sigma}}$:

$$\widehat{\boldsymbol{\Sigma}} = \left[\begin{array}{c|c|c|c} \widehat{\boldsymbol{\Sigma}_1} & 0 & \dots & 0 \\ \hline \mathbf{0} & \widehat{\boldsymbol{\Sigma}_2} & \dots & \mathbf{0} \\ \hline & & \ddots & \\ \hline \mathbf{0} & \mathbf{0} & \dots & \widehat{\boldsymbol{\Sigma}_M} \end{array}\right] \text{ with } \widehat{\boldsymbol{\Sigma}_j} = \begin{bmatrix} \widehat{\sigma^2} & \widehat{\sigma^2}\cdot\widehat{\rho} & \dots & \widehat{\sigma^2}\cdot\widehat{\rho} \\ \widehat{\sigma^2}\cdot\widehat{\rho} & \widehat{\sigma^2} & \dots & \widehat{\sigma^2}\cdot\widehat{\rho} \\ & & \ddots & \\ \widehat{\sigma^2}\cdot\widehat{\rho} & \widehat{\sigma^2}\cdot\widehat{\rho} & \dots & \widehat{\sigma^2} \end{bmatrix}$$

2. Plug $\widehat{\boldsymbol{\Sigma}}$ into the sandwich estimator to obtain the cluster "corrected" estimator of the variance-covariance matrix

$$V[\hat{\beta}|\mathbf{X}] = (\mathbf{X}'\mathbf{X})^{-1}\,\mathbf{X}'\widehat{\boldsymbol{\Sigma}}\mathbf{X}\,(\mathbf{X}'\mathbf{X})^{-1}$$

- No canned function for CRSE in R; use our custom function posted on the course website

```
> source("vcovCluster.r")
> coeftest(model, vcov = vcovCluster(model, cluster = clusterID))
```

Example: Gerber, Green, Larimer

Dear Registered Voter:

WHAT IF YOUR NEIGHBORS KNEW WHETHER YOU VOTED?

Why do so many people fail to vote? We've been talking about the problem for years, but it only seems to get worse. This year, we're taking a new approach. We're sending this mailing to you and your neighbors to publicize who does and does not vote.

The chart shows the names of some of your neighbors, showing which have voted in the past. After the August 8 election, we intend to mail an updated chart. You and your neighbors will all know who voted and who did not.

DO YOUR CIVIC DUTY — VOTE!

MAPLE DR	Aug 04	Nov 04	Aug 06
9995 JOSEPH JAMES SMITH	Voted	Voted	______
9995 JENNIFER KAY SMITH		Voted	______
9997 RICHARD B JACKSON		Voted	______
9999 KATHY MARIE JACKSON		Voted	______

Social Pressure Model

```
load("gerber_green_larimer.RData")
social$voted <- 1 * (social$voted == "Yes")
social$treatment <- factor(social$treatment,
   levels = c("Control", "Hawthorne", "Civic Duty",
              "Neighbors", "Self"))
mod1 <- lm(voted ~ treatment, data = social)
coeftest(mod1)
```

```
##
## t test of coefficients:
##
##                        Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)           0.2966383  0.0010612 279.5250 < 2.2e-16 ***
## treatmentHawthorne    0.0257363  0.0026007   9.8958 < 2.2e-16 ***
## treatmentCivic Duty 0.0178993  0.0026003   6.8835 5.849e-12 ***
## treatmentNeighbors    0.0813099  0.0026008  31.2634 < 2.2e-16 ***
## treatmentSelf         0.0485132  0.0026003  18.6566 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Social Pressure Model, CRSEs

Again no canned CRSE in R, so we use our own.

```
source("vcovCluster.R")
coeftest(mod1, vcov = vcovCluster(mod1, "hh_id"))
```

```
##
## t test of coefficients:
##
##                         Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)            0.2966383  0.0013096 226.5172 < 2.2e-16 ***
## treatmentHawthorne     0.0257363  0.0032579   7.8997 2.804e-15 ***
## treatmentCivic Duty    0.0178993  0.0032366   5.5302 3.200e-08 ***
## treatmentNeighbors     0.0813099  0.0033696  24.1308 < 2.2e-16 ***
## treatmentSelf          0.0485132  0.0033000  14.7009 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Cluster-Robust Standard Errors

- CRSE do not change our estimates $\widehat{\beta}$, cannot fix bias
- CRSE is consistent estimator of Var$[\widehat{\beta}]$ given clustered dependence
  - Relies on independence between clusters, dependence within clusters
  - Doesn't depend on the model we present
  - CRSEs usually $>$ conventional SEs—use when you suspect clustering
- Consistency of the CRSE are in the number of groups, not the number of individuals
  - CRSEs can be incorrect with a small ($<$ 50 maybe) number of clusters (often biased downward)
  - Block bootstrap can be a useful alternative (key idea: bootstrap by resampling the clusters)

Time Dependence: Serial Correlation

- Sometimes we deal with data that is measured over time, $t = 1, \ldots, T$
- Examples: a country over several years or a person over weeks/months
- Often have serially correlated: errors in one time period are correlated with errors in other time periods
- Many different ways for this to happen, but we often assume a very limited type of dependence called AR(1).

Time Dependence: Serial Correlation

Suppose we observe a unit at multiple times $t = 1, ..., T$ (e.g. a country over several years, an individual over several month, etc.).

Such observations are often serially correlated (not independent across time). We can model this with the following AR(1) model:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

where the autoregressive error is

$$u_t = \rho\, u_{t-1} + e_t \quad \text{where} \quad |\rho| < 1$$

- $e_t \sim N(0, \sigma_e^2)$
- ρ is an unknown autoregressive coefficient (note if $\rho = 0$ we have classic errors used before)
- Typically assume stationarity meaning that $V[u_t]$ and $Cov[u_t, u_{t+h}]$ are independent of t
- Generalizes to higher order serial correlation (e.g. an AR(2) model is given by $u_t = \rho u_{t-1} + \delta u_{t-2} + e_t$).

The Error Structure for the AR(1) Model

We have $\mathbf{u} = \begin{bmatrix} u_1 & u_2 & \dots & u_T \end{bmatrix}'$ and the AR(1) model implies the following error structure (derivation in appendix):

$$V[\mathbf{u}] = \boldsymbol{\Sigma} = \frac{\sigma^2}{(1-\rho^2)} \begin{bmatrix} 1 & \rho & \rho^2 & \dots & \rho^{T-1} \\ \rho & 1 & \rho & \dots & \rho^{T-2} \\ \rho^2 & \rho & 1 & \dots & \rho^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \rho^{T-1} & \rho^{T-2} & \rho^{T-3} & \dots & 1 \end{bmatrix}$$

That is, the covariance between errors in $t = 1$ and $t = 2$ is $\frac{\sigma^2}{(1-\rho^2)}\rho$, between errors in $t = 1$ and $t = 3$ is $\frac{\sigma^2}{(1-\rho^2)}\rho^2$, etc.

This implies that the correlation between the errors decays exponentially with the number of periods separating them.

ρ is usually positive, which implies that we underestimate the variance if we ignore serial correlation.

How to Detect and Fix Serial Correlated Errors

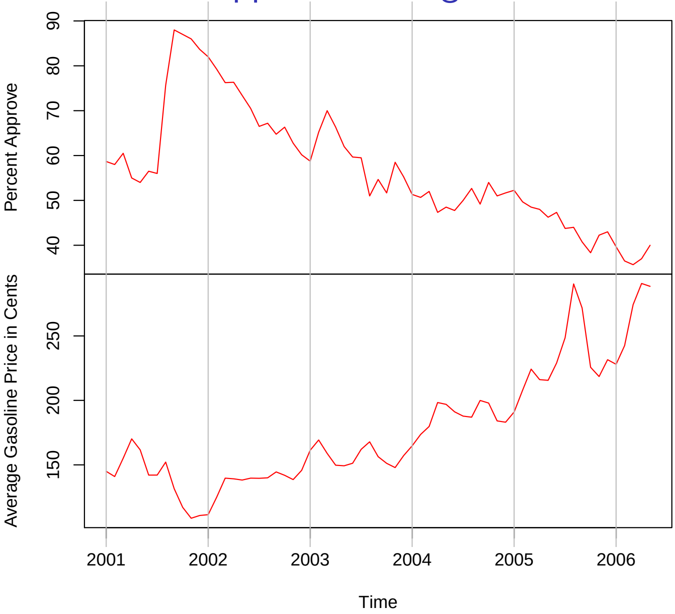
Detection:

- Plot residuals over time (or more fancy "autocorrelation" plots)
- Formal tests (e.g. Durbin-Watson statistics)

Possible Corrections:

- Use standard errors that are robust to serial correlation (e.g. Newey-West)
- AR corrections (e.g. Prais-Winston, Cochrane-Orcutt, etc.)
- Lagged dependent variables or other dynamic panel models
- First-differencing the data

Monthly Presidential Approval Ratings and Gas Prices



Monthly Presidential Approval Ratings and Gas Prices

R Code

```
> library(Zelig)
> data(approval)
> mod1 <- lm(approve ~ avg.price, data=approval)
> coeftest(mod1)

t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 100.472076   3.567277  28.165 < 2.2e-16 ***
avg.price    -0.243885   0.019465 -12.529 < 2.2e-16 ***
```

Tests for Serial Correlation: Durbin-Watson

Recall our AR(1) model is:

$$y_t = \beta_0 + \beta_1 x_t + u_t$$

where $u_t = \rho u_{t-1} + e_t$, $e_t \sim N(0, \sigma^2)$, and ρ is our unknown autoregressive coefficient (with $|\rho| < 1$).

The null hypothesis (no serial correlation) is: H_0: $\rho = 0$
The alternative (positive serial correlation): H_1: $\rho > 0$

One common test for serial correlation is the Durbin-Watson statistic:

$$DW = \frac{\sum_{t=2}^{n} \hat{u}_t - \hat{u}_{t-1}}{\sum_{t=1}^{n} \hat{u}_t^2} \quad \text{where} \quad DW \approx 2(1 - \widehat{\rho})$$

- If $DW \approx 2$ then $\widehat{\rho} \approx 0$ (Note that $0 \leq DW \leq 4$)
- If $DW < 1$ we have serious positive serial correlation
- If $DW > 3$ we have serious negative serial correlation

Monthly Presidential Approval Ratings and Gas Prices

R Code

```
> library(lmtest)
> dwtest(approve ~ avg.price, data=approval)

        Durbin-Watson test

data:  approve ~ avg.price
DW = 0.4863, p-value = 1.326e-14
alternative hypothesis: true autocorrelation is greater than 0
```

The test suggests strong positive serial correlation. Standard errors are severely downward biased.

Corrections: HAC Standard Errors

- A common way to correct for serial correlation is to use OLS but to estimate the variances using an estimator that is heteroskedasticity and autocorrelation consistent (HAC) (Newey and West (1987)).

- The theory behind the HAC variance estimator is somewhat complicated, but the interpretation is similar to our usual OLS robust standard errors.
  - HAC standard errors leave estimate of $\hat{\beta}$ unchanged and do not fix potential bias in $\hat{\beta}$
  - HAC are consistent estimator for $V[\hat{\beta}]$ in the presence of heteroskedasticity and or autocorrelation
  - The `sandwich` package in R implements a variety of HAC estimators
  - A common option is `NeweyWest`

Monthly Presidential Approval Ratings and Gas Prices

R Code

```
> mod1 <- lm(approve~avg.price,data=approval)
> coeftest(mod1) # homoskedastic errors
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 100.472076   3.567277  28.165 < 2.2e-16 ***
avg.price    -0.243885   0.019465 -12.529 < 2.2e-16 ***

> coeftest(mod1, vcov = NeweyWest) # HAC errors
t test of coefficients:

              Estimate Std. Error t value  Pr(>|t|)
(Intercept) 100.472076  14.499337  6.9294 2.652e-09 ***
avg.price    -0.243885   0.071733 -3.3999  0.001174 **
```

Once we correct for autocorrelation, standard errors increase dramatically.

Review of Standard Errors

- Violations of homoskedasticity can come in many forms
  - Non-constant error variance
  - Clustered dependence
  - Serial dependence
- Use plots or formal tests to detect heteroskedasticity
- "Robust SEs" of various forms are consistent even when these problems are present
  - White HC standard errors
  - Cluster-robust standard errors
  - Newey-West HAC standard errors

Appendix: Derivation of Error Structure for the AR(1) Model

We have

$$V[u_t] = V[\rho u_{t-1} + e_t] = \rho^2 V[u_{t-1}] + \sigma^2$$

with stationarity, $V[u_t] = V[u_{t-1}]$, and so

$$V[u_t](1-\rho^2) = \sigma^2 \Rightarrow V[u_t] = \frac{\sigma^2}{(1-\rho^2)}$$

also

$$Cov[u_t, u_{t-1}] = E[u_t u_{t-1}] = E[(\rho u_{t-1} + e_t)e_{t-1}] = \rho V[e_{t-1}] = \rho \frac{\sigma^2}{(1-\rho^2)}$$

or generally

$$Cov[u_t, u_{t-h}] = \rho^h \frac{\sigma^2}{(1-\rho^2)}$$

A Contrarian View of Robust Standard Errors

King, Gary and Margaret E. Roberts. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It" *Political Analysis* (2015) 23: 159-179.[3]

[3] I thank Gary and Molly for the slides that follow.

Robust Standard Errors: Used Everywhere

- Robust standard errors: a widely used technique to fix SEs under model misspecification
  - In Political Science:
    - APSR (2009-2012): 66% of articles using regression
    - IO (2009-2012): 73% of articles using regression
    - AJPS (2009-2012): 45% of articles using regression
  - Everywhere else, too:
    - All of Google Scholar: 53,900 mentions
    - And going up: 1,000 new per month

Robust Standard Errors are a Bright, Red Flag

- People think robust se's will inoculate them from criticism
- They are wrong
- Instead, they are a bright, reg flag saying:
  "My model is misspecified!"

RSEs: Two Possibilities

RSEs and SEs differ

- In the best case scenario:
  - Some coefficients: unbiased but inefficient
  - Other quantities of interest: Biased
- In the worst case scenario:
  - The functional form, variance, or dependence specification is wrong
  - All quantities of interest will be biased.

RSEs and SEs are the same

- Consistent with a correctly specified model
- RSEs are not useful, as a "fix"

Their Alternative Procedure

Robust standard errors:

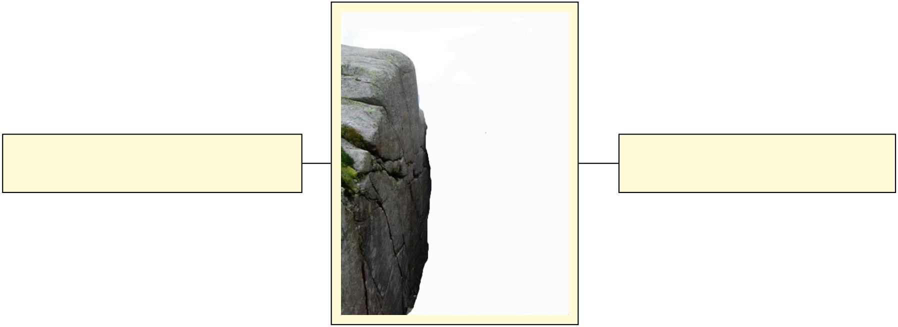
- What they are not: an elixir
- What they are: Extremely sensitive misspecification detectors!

We should use them to: Test misspecification!

1. Do RSEs and SEs differ?
2. If they do:
   - Use model diagnostics (e.g. residual plots, qq-plots, misspecification tests)
   - Evaluate misspecification
   - Respecify the model
3. Keeping going, until they don't differ.

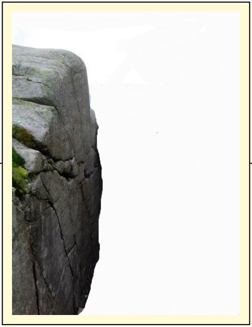
For RSEs to help: Everything has to be Juuuussttt Right

For RSEs to help: Everything has to be Juuuussttt Right



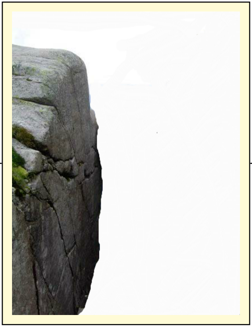
For RSEs to help: Everything has to be Juuuussttt Right

Model Correct



For RSEs to help: Everything has to be Juuuussttt Right

Model Correct



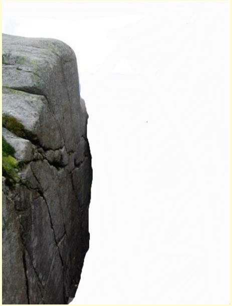
RSEs same as SEs

For RSEs to help: Everything has to be Juuuussttt Right

Model Correct

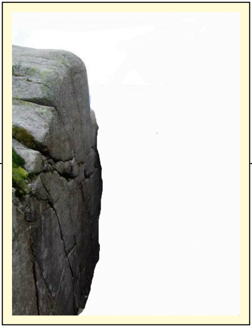
RSEs same as SEs

Point estimates correct



For RSEs to help: Everything has to be Juuuussttt Right

Model Correct



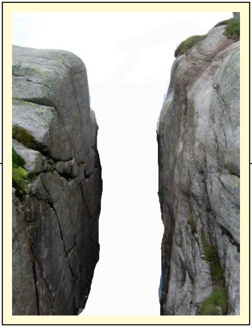
RSEs same as SEs

Point estimates correct

Awesome!

For RSEs to help: Everything has to be Juuuussttt Right

Model Correct



Model Misspecified

RSEs same as SEs

Point estimates correct

Awesome!

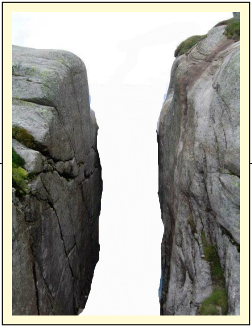
For RSEs to help: Everything has to be Juuuussttt Right

Model Correct

RSEs same as SEs

Point estimates correct

Awesome!



Model Misspecified

RSEs differ from SEs

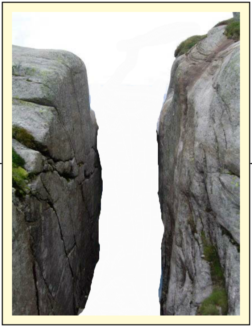
For RSEs to help: Everything has to be Juuuussttt Right

Model Correct

RSEs same as SEs

Point estimates correct

Awesome!



Model Misspecified

RSEs differ from SEs

Point estimates biased

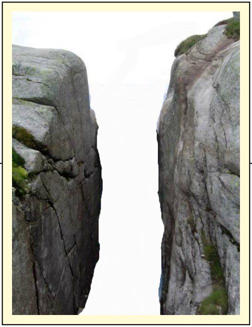
For RSEs to help: Everything has to be Juuuussttt Right

Model Correct

RSEs same as SEs

Point estimates correct

Awesome!



Model Misspecified

RSEs differ from SEs

Point estimates biased

Respecify!

For RSEs to help: Everything has to be Juuuussttt Right

Goldilocks Region

Model Correct

RSEs same as SEs

Point estimates correct

Awesome!

Model Misspecified

RSEs differ from SEs

Point estimates biased

Respecify!

For RSEs to help: Everything has to be Juuuussttt Right

Goldilocks Region

Model Correct

RSEs same as SEs

Point estimates correct

Awesome!

Model Misspecified

RSEs differ from SEs

Point estimates biased

Respecify!

For RSEs to help: Everything has to be Juuuussttt Right

Biased just enough to make RSEs useful,

Goldilocks Region

Model Correct

RSEs same as SEs

Point estimates correct

Awesome!

Model Misspecified

RSEs differ from SEs

Point estimates biased

Respecify!

For RSEs to help: Everything has to be Juuuussttt Right

Biased just enough to make RSEs useful,

Goldilocks Region

but not so much as to bias everything else

Model Correct

RSEs same as SEs

Point estimates correct

Awesome!

Model Misspecified

RSEs differ from SEs

Point estimates biased

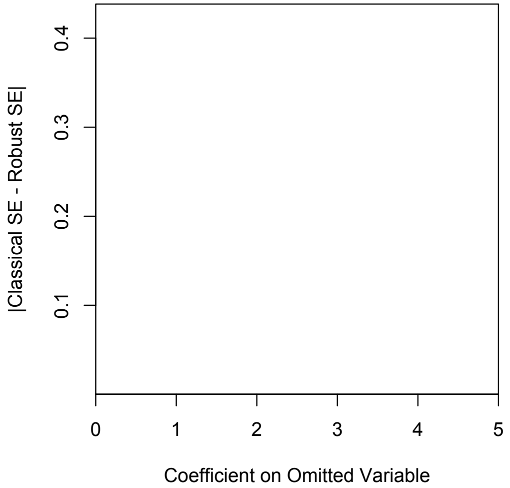
Respecify!

The Goldilocks Region is not Idyllic

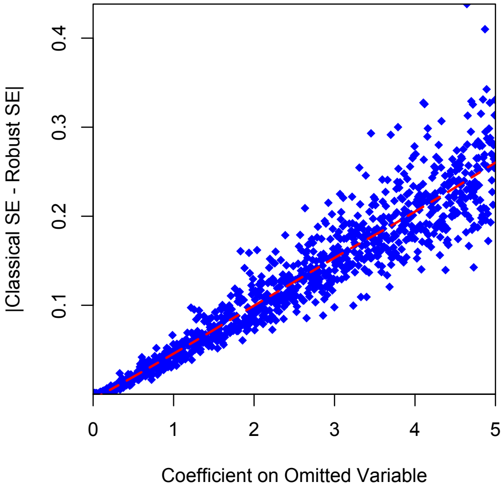
In the Goldilocks region,

- No fully specified model
- Only a few QOI's can be estimated.
  - Suppose DV: Democrat proportion of two-party vote.
  - We can estimate β, but not:
    - the probability the Democrat wins,
    - the variation in vote outcome,
    - or vote predictions with confidence intervals.
  - We can't check: whether model implications are realistic.
- Parts of the model are wrong; why do we think the rest are right?

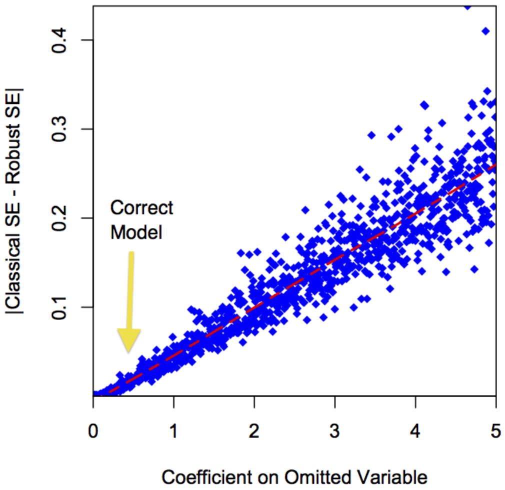
Difference Between SE and RSE Exposes Misspecification



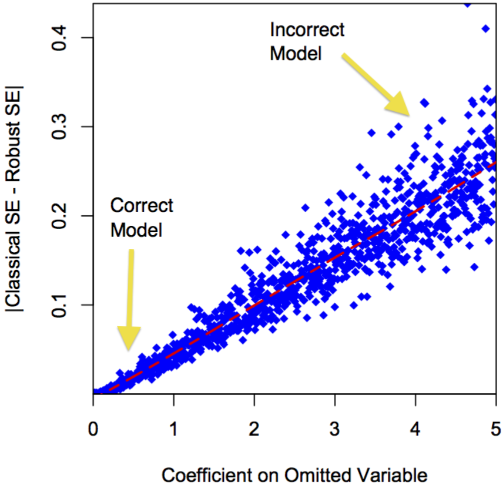
Difference Between SE and RSE Exposes Misspecification



Difference Between SE and RSE Exposes Misspecification



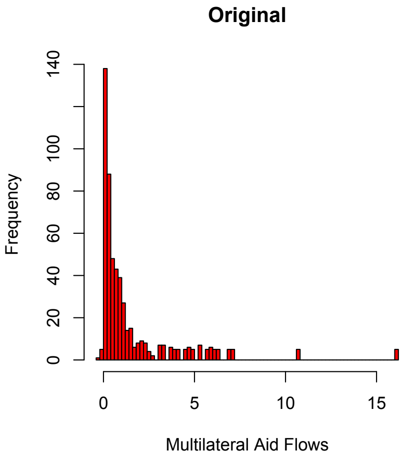
Difference Between SE and RSE Exposes Misspecification

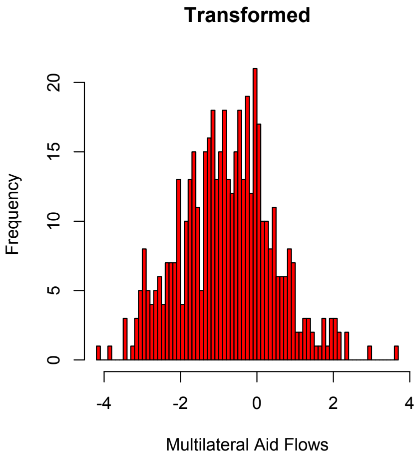


Example: RSEs Expose Non-normality

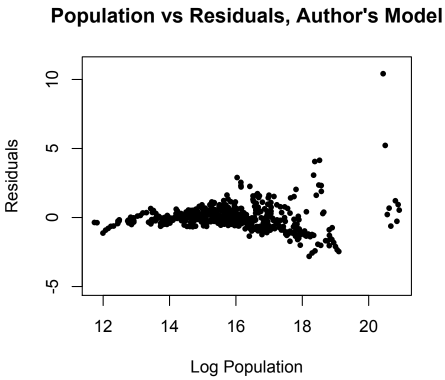
- Replication Neumayer, ISQ 2003
- Dependent variable: multilateral aid flows (as percentage of GDP)
- Treatment of interest: population
- Controls: GDP, former colony status, distance from Western world, etc ...
- Conclusion: Aid favors less populous countries.
- Difference between RSEs and SEs: Large.
  - Robust SE: 0.72, SE: 0.37
  - $\Rightarrow$ indicates model misspecification

Problem: Highly Skewed Dependent Variable

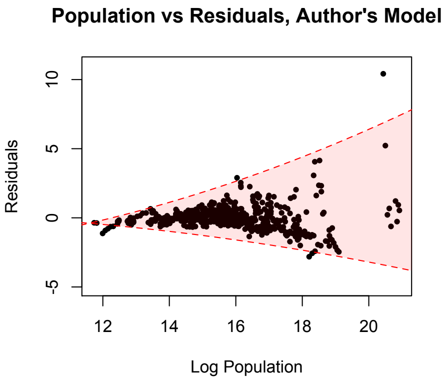


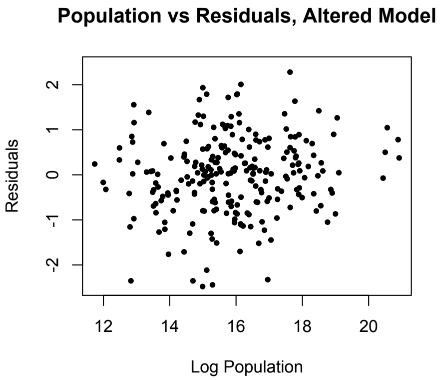


Diagnostics: Reveal Misspecification



Diagnostics: Reveal Misspecification

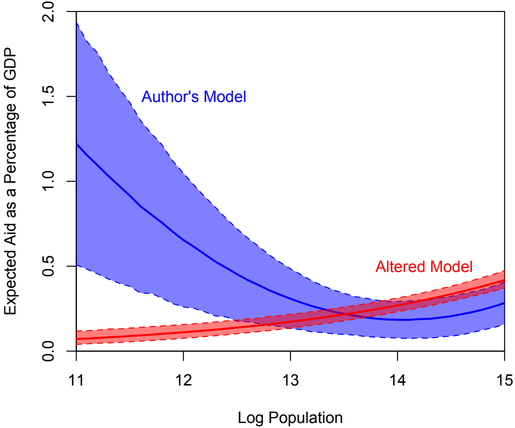




Textbook case of heteroskedasticity

Textbook case of homoskedasticity

After Fix: Different Conclusion



Concluding Contrarian Thoughts

Their advice:

- RSEs: not an elixir. Should not be used as a patch.
- Instead: a sensitive detector of misspecification.
- Evaluate misspecification, by conducting diagnostic tests.
- Respecify the model, until robust and classical SE's coincide

Their Examples:

- Robust SEs indicate fundamental modelling problems
- Easily identified with diagnostics
- Fixing these problems $\Rightarrow$ hugely different substantive conclusions

Concluding Thoughts on Diagnostics

Residuals are important. Look at them.

Next Week

- Causality with Measured Confounding
- Reading:
  - Angrist and Pishke Chapter 2 (The Experimental Ideal) Chapter 3.2 (Regression and Causality)
  - Morgan and Winship Chapters 3-4 (Causal Graphs and Conditioning Estimators)
  - Optional: Elwert and Winship (2014) "Endogenous selection bias: The problem of conditioning on a collider variable" *Annual Review of Sociology*
  - Optional: Morgan and Winship Chapter 11 Repeated Observations and the Estimation of Causal Effects
- As a side note: if you want to read the argument against the contrarian response: Aronow (2016) "A Note on 'How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It.'" It is an interesting piece- feel free to come talk to me about this debate!

Appendix: Derivation of Variance under Homoskedasticity

$$\begin{aligned}\hat{\boldsymbol{\beta}} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}+\mathbf{u}) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}\end{aligned}$$

$$\begin{aligned}V[\hat{\boldsymbol{\beta}}|\mathbf{X}] &= V[\boldsymbol{\beta}|\mathbf{X}] + V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\ &= V[(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{u}|\mathbf{X}] \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \text{ (note: } \mathbf{X} \text{ nonrandom } |\mathbf{X}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'V[\mathbf{u}|\mathbf{X}]\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \text{ (by homoskedasticity)} \\ &= \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\end{aligned}$$

Replacing σ^2 with our estimator $\widehat{\sigma}^2$ gives us our estimator for the $(k+1)\times(k+1)$ variance-covariance matrix for the vector of regression coefficients:

$$\widehat{V[\hat{\boldsymbol{\beta}}|\mathbf{X}]} = \widehat{\sigma}^2(\mathbf{X}'\mathbf{X})^{-1}$$

Fun With Neighbors

- We talked about error dependence induced by time and by cluster.
- An alternative process is spatial dependence.
- Just as with the other types of models we have to specify what it means to be close to a neighbor, but this choice is often more influential than anticipated.

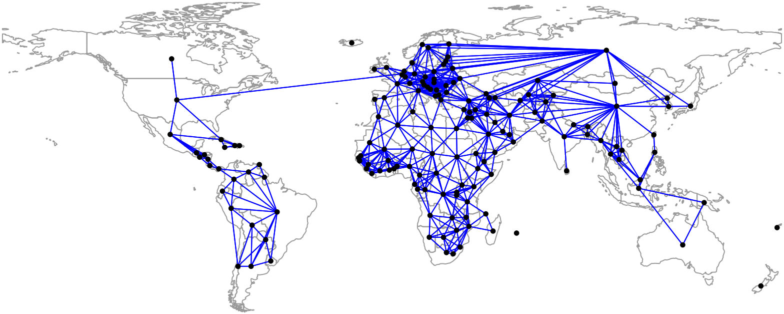
Zhukov, Yuri M. and Brandon M. Stewart. "Choosing Your Neighbors: Networks of Diffusion in International Relations" *International Studies Quarterly* 2013; 57: 271-287.

Our Main Questions

How Do We Generally Choose Neighbors?

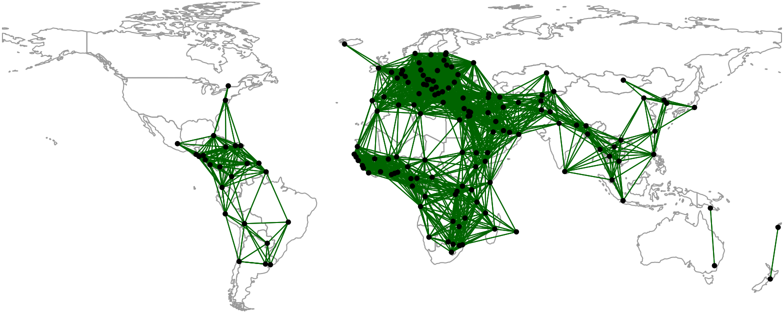
Visualization of Connections: Contiguity

Figure: Contiguity neighbors with 500 km snap distance



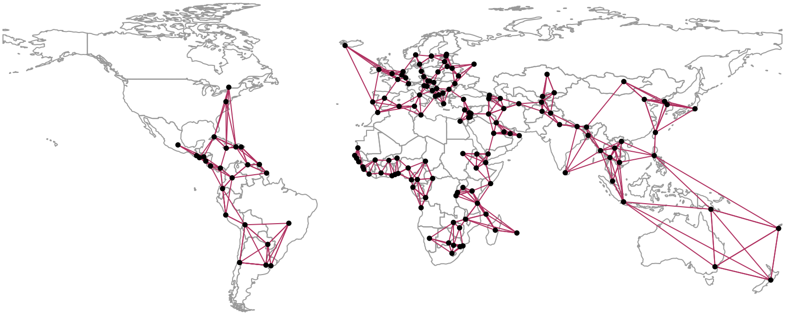
Visualization of Connections: Minimum Distance

Figure: Minimum distance neighbors (capital cities)



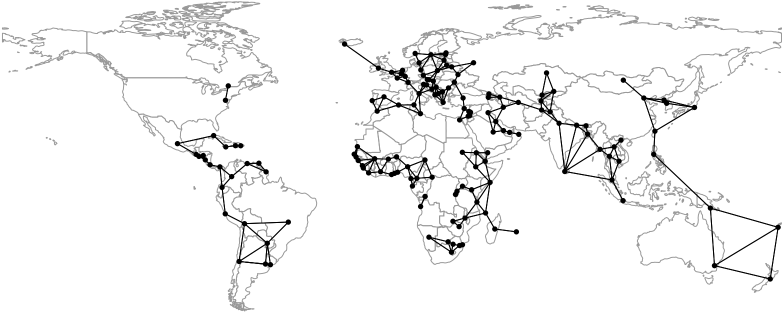
Visualization of Connections: K-Nearest Neighbors

Figure: $k = 4$ Nearest Neighbors (capital cities)



Visualization of Connections: Graph-based Neighbors

Figure: Sphere of Influence Neighbors (capital cities)



Application: Democratic Diffusion

Gleditsch and Ward (2006)
Changes of political regime modeled as a first-order Markov chain process with the transition matrix

$$\mathbf{K} = \begin{bmatrix} Pr(y_{i,t} = 0 | y_{i,t-1} = 0) & Pr(y_{i,t} = 1 | y_{i,t-1} = 0) \\ Pr(y_{i,t} = 0 | y_{i,t-1} = 1) & Pr(y_{i,t} = 1 | y_{i,t-1} = 1) \end{bmatrix}$$

where $y_{i,t} = 1$ if an (A)utocratic regime exists in country i at time t, and $y_{i,t} = 0$ if the regime is (D)emocratic.
. . . in other words:

$$\mathbf{K} = \begin{bmatrix} Pr(D \to D) & Pr(D \to A) \\ Pr(A \to D) & Pr(A \to A) \end{bmatrix}$$

Equilibrium Effects of Democratic Transition

If a regime transition takes place in country i, what is the change in predicted probability of a regime transition in country j (country i's neighbor)?

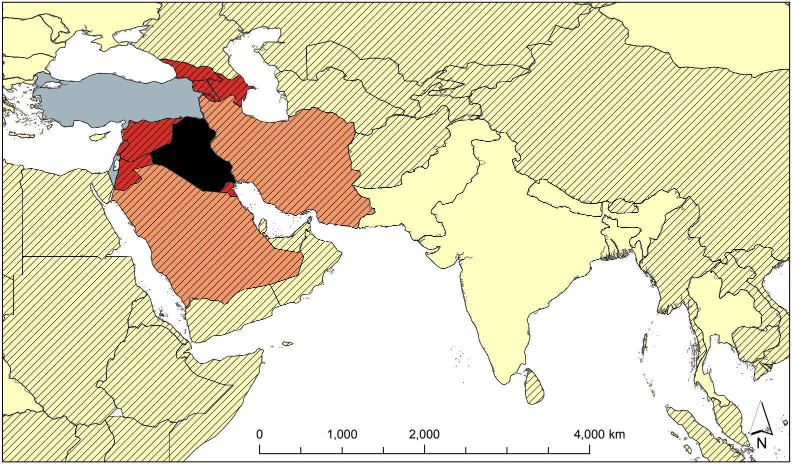
$$\text{QI} = Pr(y_{j,t}|y_{i,t} = y_{i,t-1}) - Pr(y_{j,t}|y_{i,t} \neq y_{i,t-1})$$

where $y_{i,t} = 0$ if country i is a democracy at time t and $y_{i,t} = 1$ if it is an autocracy. All other covariates are held constant.

Illustrative cases

- Iraq transitions from autocracy to democracy.
- Russia transitions from democracy to autocracy.

Iraq's democratization and regional regime stability



Contiguity + 500 km

Iraq transitions from autocracy to democracy (1998 data)

Monte Carlo simulation (1,000 runs)

Iraq

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

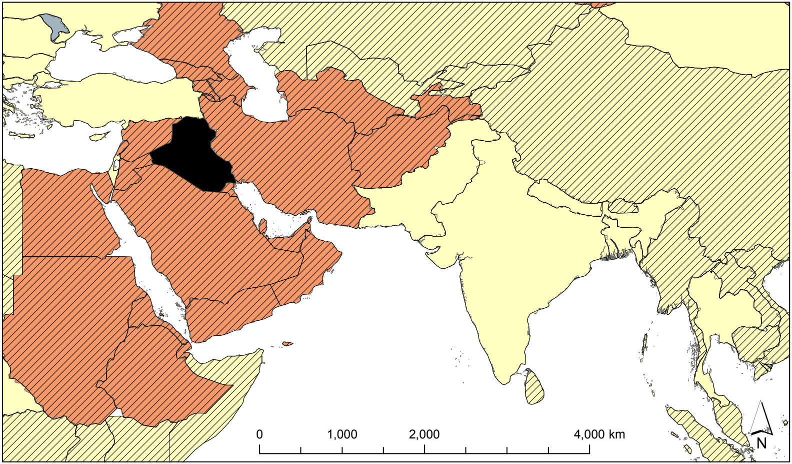
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Iraq's democratization and regional regime stability



Minimum Distance

Iraq transitions from autocracy to democracy (1998 data)

Monte Carlo simulation (1,000 runs)

Iraq

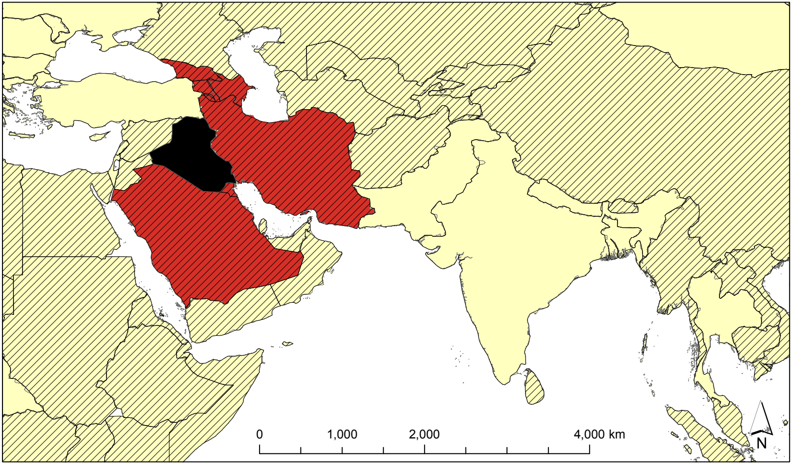
Regime Type

- Democracy
- Autocracy

Change in Transition Probability

- -0.05- -0.025
- -0.025 - -0.001
- 0
- 0.001 - 0.025
- 0.025 - 0.05

Iraq's democratization and regional regime stability



k = 4 Nearest Neighbors

Iraq transitions from autocracy to democracy (1998 data)

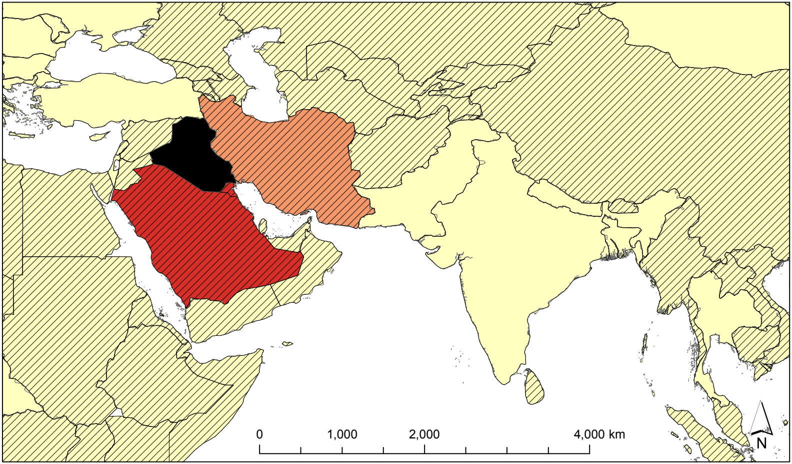
Monte Carlo simulation (1,000 runs)

Iraq

Regime Type
- Democracy
- Autocracy

Change in Transition Probability
- -0.05- -0.025
- -0.025 - -0.001
- 0
- 0.001 - 0.025
- 0.025 - 0.05

Iraq's democratization and regional regime stability



Sphere of Influence

Iraq transitions from autocracy to democracy (1998 data)

Monte Carlo simulation (1,000 runs)

Iraq

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

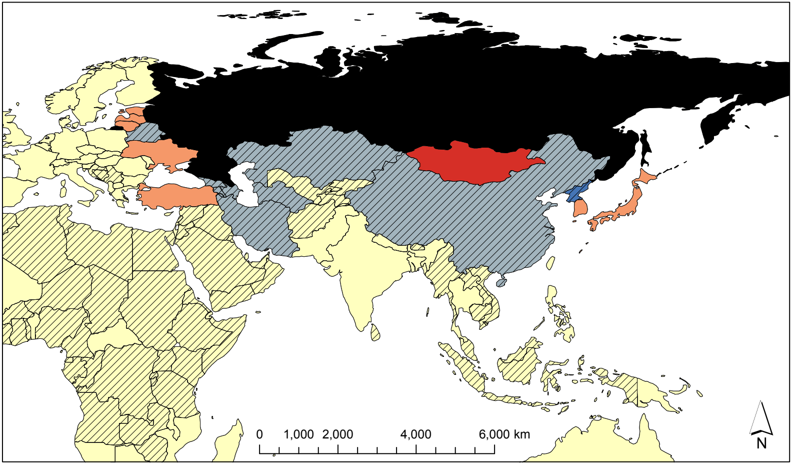
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Russia's autocratization and regional regime stability



Contiguity + 500 km

Russia transitions from democracy to autocracy (1998 data)

Monte Carlo simulation (1,000 runs)

Russia

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

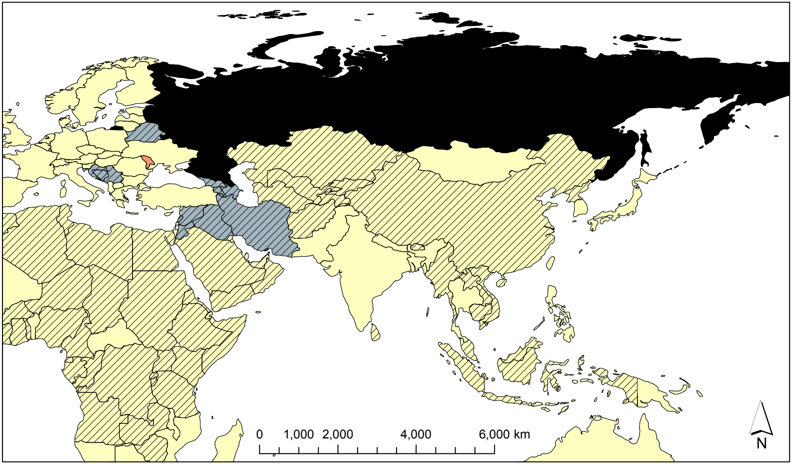
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Russia's autocratization and regional regime stability



Minimum Distance

Russia transitions from democracy to autocracy
(1998 data)

Monte Carlo simulation (1,000 runs)

Russia

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

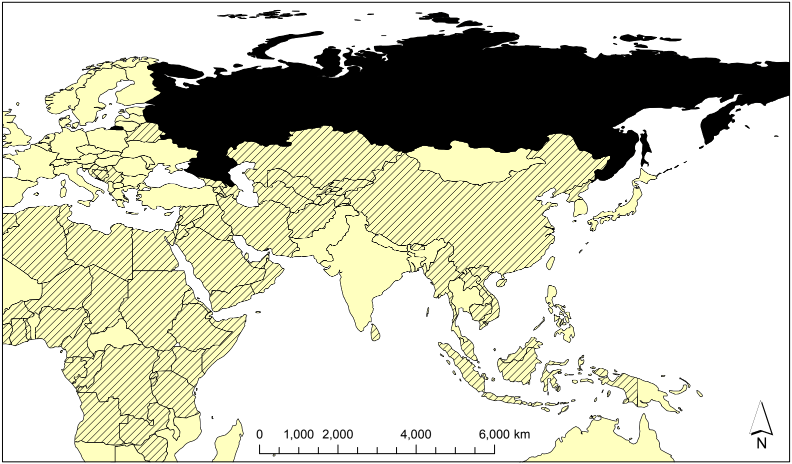
-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Russia's autocratization and regional regime stability



k = 4 Nearest Neighbors

Russia transitions from democracy to autocracy (1998 data)

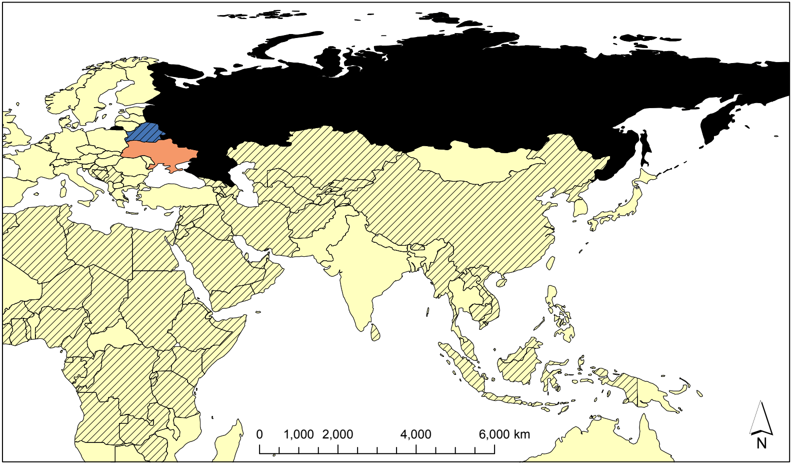
Monte Carlo simulation (1,000 runs)

Russia	
Regime Type	
Democracy	
Autocracy	

Change in Transition Probability

- -0.05- -0.025
- -0.025 - -0.001
- 0
- 0.001 - 0.025
- 0.025 - 0.05

Russia's autocratization and regional regime stability



Sphere of Influence

Russia transitions from democracy to autocracy (1998 data)

Monte Carlo simulation (1,000 runs)

Russia

Regime Type

Democracy

Autocracy

Change in Transition Probability

-0.05- -0.025

-0.025 - -0.001

0

0.001 - 0.025

0.025 - 0.05

Kitten Wars


kittenwar
Warm, Flaky, Delicious
Crowd Pleaser.
Pillsbury™ Crescent Rolls
Save $1 Now
Crescent
Alfredo vs. Amy
Home
Winningest Kittens
Losingest Kittens
Newest Kittens
Add your Kitten
The Daily Kitten
facebook group
faq
Privacy Policy
e-mail us
kitten search:
Go
Click on the cutest to decide the winner!!!
Can't decide? Refresh the page for a draw.

Kitten Wars

Winningest Kittens!

more

1. 

Bitsy
has won 76% of 8642 battles.

2. 

Freddie
has won 76% of 3768 battles.

Losingest Kittens!

more

1. 

Scary Cat
has lost 79% of 11211 battles.

2. 

Beitsim
has lost 79% of 1919 battles.

Kitten Wars for Ideas (Salganik and Levy)

RESEARCH ARTICLE

Wiki Surveys: Open and Quantifiable Social Data Collection

Matthew J. Salganik, Karen E. C. Levy

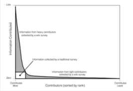
Published: May 20, 2015 • DOI: 10.1371/journal.pone.0123483

Article	Authors	Metrics	Comments	Related Content

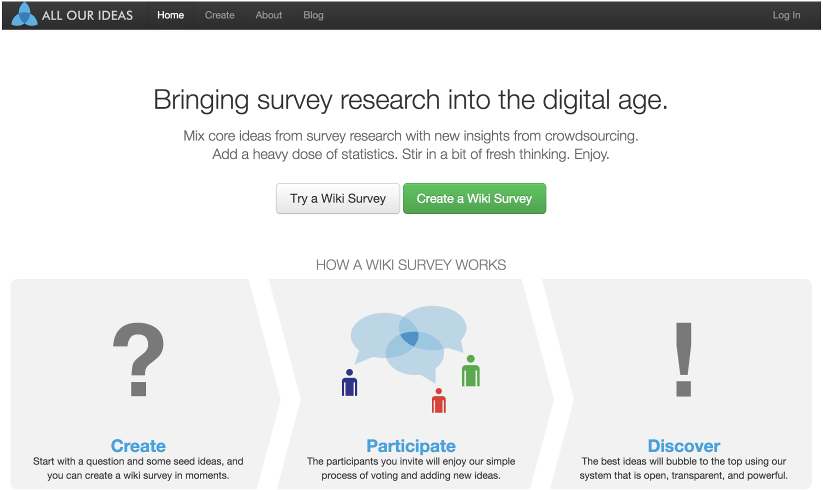
Abstract

In the social sciences, there is a longstanding tension between data collection methods that facilitate quantification and those that are open to unanticipated information. Advances in technology now enable new, hybrid methods that combine some of the benefits of both approaches. Drawing inspiration from online information aggregation systems like Wikipedia and from traditional survey research, we propose a new class of research instruments called *wiki surveys*. Just as Wikipedia evolves over time based on contributions from participants, we envision an evolving survey driven by contributions from respondents. We develop three general principles that underlie wiki surveys: they should be greedy, collaborative, and adaptive. Building on these principles, we develop methods for data collection and data analysis for one type of wiki survey, a pairwise wiki survey. Using two proof-of-concept case studies involving our free and open-source website www.allourideas.org, we show that pairwise wiki surveys can yield insights that would be difficult to obtain with other methods.

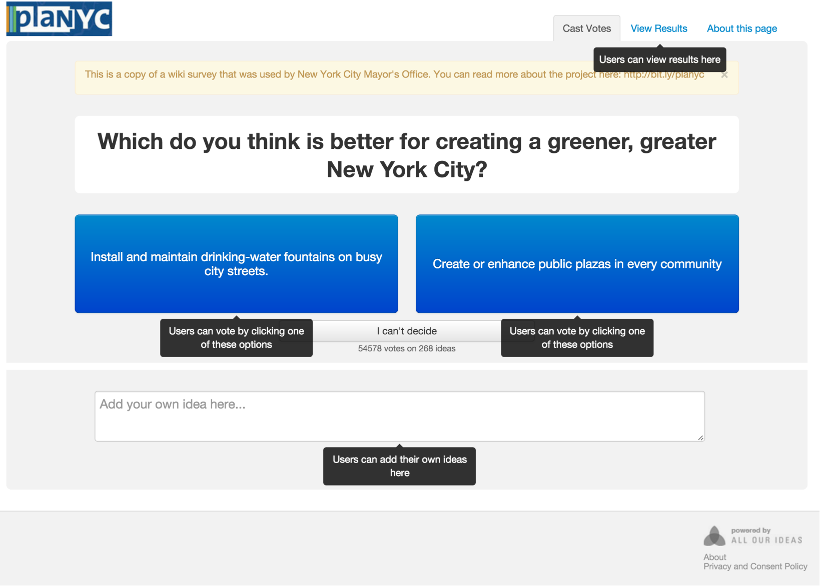
Figures



Kitten Wars for Ideas (Salganik and Levy)



Kitten Wars for Ideas (Salganik and Levy)



Kitten Wars for Ideas (Salganik and Levy)

Which do you think is better for creating a greener, greater New York City?

Ideas	Score (0 - 100)
Require all big buildings to make certain energy efficiency upgrades	67
Promote cycling by installing safe bike lanes	65
Promote the use of solar energy using the latest technology on all high-rise buildings.	65
Invest in multiple modes of transportation and provide both improved infrastructure and improved safety	65
Continue enhancing bike lane network, to finally connect separated bike lane systems to each other across all five boroughs.	65
Replace sodium vapor street lights with LED or other energy-saving lights.	64
Utilize NYC Rooftops to install Solar PV panels	63
Plant more trees	62
Create a network of protected bike paths throughout the entire city	62
Add improvements to the bike lanes in the inner city. This will encourage exercise and reduce city's carbon footprint.	62

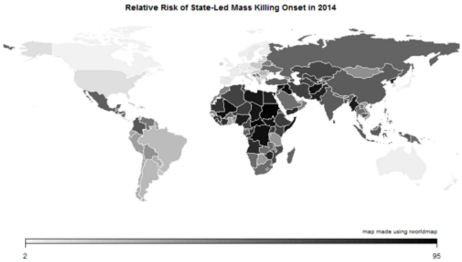
The Power of Releasing Software

The Governor asks ... again

Governor Tarso Genro of the state of Rio Grande do Sul, Brazil has done it again. The Governor and his team completed a second round of their amazing open government project called Governador Pergunta (The Governor Asks), which collects public feedback on important policy challenges using a customized version of allourideas.org.

The Power of Releasing Software

wiki surveys to assess risks of state-led mass killings



As part of their work with the Holocaust Museum's Center for the Prevention of Genocide, Jay Ulfelder and Ben Valentino launched a wiki survey to help assess the risks of state-led mass killing onsets in 2014. You can read about their results on this interesting blog post.

The Power of Releasing Software

UN Global Sustainability Report 2013

We are happy to announce that the United Nations Division for Sustainable Development is using allourideas.org to solicit ideas from scientists around the world for the 2013 UN Global Sustainability Report. The report will

Powered By Research (You Can Do It Too!)

Backed by research

All Our Ideas is a research project based at Princeton University that is dedicated to creating new ways of collecting social data. You can learn more about the theory and methods behind our project by reading our paper or watching our talk. Thanks to Google, the National Science Foundation, and Princeton for supporting this research.

$$z_i \sim \begin{cases} N(\dot{\boldsymbol{x}}_i^T \boldsymbol{\theta}_{\boldsymbol{v}}, 1) I(z_i^* > 0) & \text{if } y_i = 1 \\ N(\dot{\boldsymbol{x}}_i^T \boldsymbol{\theta}_{\boldsymbol{v}}, 1) I(z_i^* < 0) & \text{if } y_i = 0 \end{cases}$$

Powered By Research (You Can Do It Too!)

Bradley–Terry model

From Wikipedia, the free encyclopedia

The **Bradley–Terry model** is a probability model that can predict the outcome of a comparison. Given a pair of individuals i and j drawn from some population, it estimates the probability that the pairwise comparison $i > j$ turns out true, as

$$P(i > j) = \frac{p_i}{p_i + p_j}$$

where p_i is a positive real-valued score assigned to individual i. The comparison $i > j$ can be read as "i is preferred to j", "i ranks higher than j", or "i beats j", depending on the application.

For example, p_i may represent the skill of a team in a sports tournament, estimated from the number of times i has won a match. $P(i > j)$ then represents the probability that i will win a match against j.[1][2] Another example used to explain the model's purpose is that of scoring products in a certain category by quality. While it's hard for a person to draft a direct ranking of (many) brands of wine, it may be feasible to compare a sample of pairs of wines and say, for each pair, which one is better. The Bradley–Terry model can then be used to derive a full ranking.[2]

Powered By Research (You Can Do It Too!)

References

- Angrist, Joshua D., and Jrn-Steffen Pischke. *Mostly harmless econometrics: An empiricist's companion*. Princeton university press, 2008.
- Breusch, Trevor S., and Adrian R. Pagan. "A simple test for heteroscedasticity and random coefficient variation." *Econometrica: Journal of the Econometric Society* (1979): 1287-1294.
- Durbin, James, and Geoffrey S. Watson. "Testing for serial correlation in least squares regression. II." *Biometrika* (1951): 159-177.
- Freedman, David A. "On the so-called Huber sandwich estimator and robust standard errors." *The American Statistician* 60.4 (2006).
- King, Gary and Margaret E. Roberts. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do About It" *Political Analysis* (2015) 23: 159-179.
- Newey, Whitney K., and Kenneth D. West. "A simple, positive semi-definite, heteroskedasticity and autocorrelationconsistent covariance matrix." (1986).
- Salganik MJ, Levy KEC (2015) Wiki Surveys: Open and Quantifiable Social Data Collection. *PLoS ONE* 10(5): e0123483.
- Wand, Jonathan N., Kenneth W. Shotts, Jasjeet S. Sekhon, Walter R. Mebane Jr, Michael C. Herron, and Henry E. Brady. "The butterfly did it: The aberrant vote for Buchanan in Palm Beach County, Florida." *American Political Science Review* (2001): 793-810.