

14. Algebra of Least Squares

Spring 2021

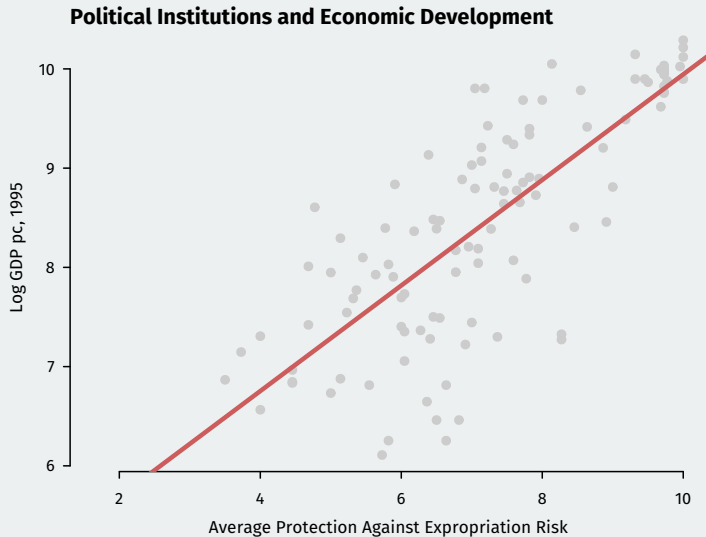
Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

- We saw how the population linear projection works.
- How can we estimate the parameters of the linear projection or CEF?
- Now: least squares estimator and its algebraic properties.
- After that: the statistical properties of least squares.

Acemoglu, Johnson, and Robinson (2001)



Samples vs population

Assumption

The variables $\{(Y_1, \mathbf{X}_1), \dots, (Y_i, \mathbf{X}_i), \dots, (Y_n, \mathbf{X}_n)\}$ are i.i.d. draws from a common distribution F .

- F is the **population distribution** or **DGP**.
 - Without i subscripts, (Y, \mathbf{X}) are r.v.s and draws from F .
- $\{(Y_i, \mathbf{X}_i) : i = 1, \dots, n\}$ is the **sample** and can be seen in two ways:
 - Numbers in your data matrix, fixed to the analyst.
 - From a statistical POV, they are realizations of a random process.
- Violations include time-series data and clustered sampling.
 - Weakening i.i.d. usually complicates notation but can be done.

Quantity of interest

- Population linear projection model:

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

- Here $\boldsymbol{\beta}$ minimizes the **population** expected squared error:

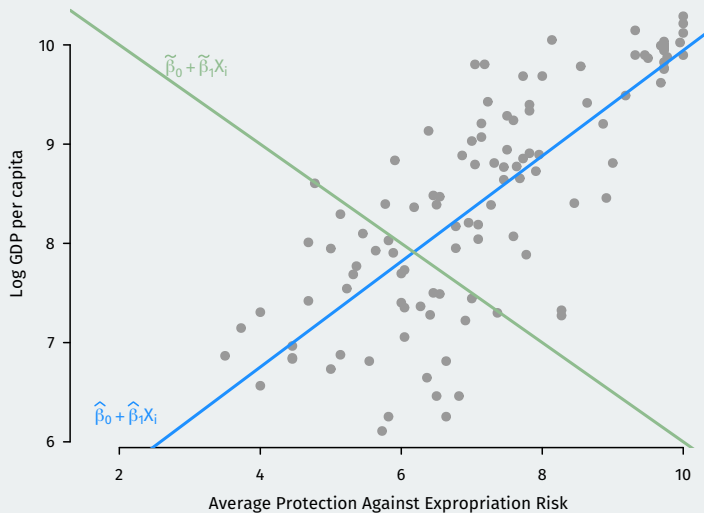
$$\boldsymbol{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^k} S(\mathbf{b}), \quad S(\mathbf{b}) = \mathbb{E} \left[(Y - \mathbf{X}'\mathbf{b})^2 \right]$$

- Last time we saw that this can be written:

$$\boldsymbol{\beta} = (\mathbb{E}[\mathbf{X}\mathbf{X}'])^{-1} \mathbb{E}[\mathbf{X}Y]$$

- How do we estimate $\boldsymbol{\beta}$?

Which line is better?



Plug-in principle returns!

- **Plug-in estimator:** solve the sample version of the population goal.
- Replace projection errors with observed errors, or **residuals:** $Y_i - \mathbf{X}_i' \mathbf{b}$
 - **Sum of squared residuals**, $SSR(\mathbf{b}) = \sum_{i=1}^n (Y_i - \mathbf{X}_i' \mathbf{b})^2$.
 - Total prediction error using \mathbf{b} as our estimated coefficient.
- We can use these residuals to get a sample average prediction error:

$$\hat{S}(\mathbf{b}) = \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{X}_i' \mathbf{b})^2 = \frac{1}{n} SSR(\mathbf{b})$$

- $\hat{S}(\mathbf{b})$ is an estimator of the expected squared error, $S(\mathbf{b})$.

Least squares estimator

- **Ordinary least squares estimator** minimizes \hat{S} in place of S .

$$\boldsymbol{\beta} = \arg \min_{\mathbf{b} \in \mathbb{R}^k} \mathbb{E} \left[(Y - \mathbf{X}'\mathbf{b})^2 \right]$$

$$\hat{\boldsymbol{\beta}} = \arg \min_{\mathbf{b} \in \mathbb{R}^k} \frac{1}{n} \sum_{i=1}^n (Y_i - \mathbf{x}_i'\mathbf{b})^2$$

- In words: find the coefficients that minimize the sum/average of the squared residuals.
- After some calculus, we can write this as a plug-in estimator:

$$\hat{\boldsymbol{\beta}} = \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i Y_i \right)$$

- $n^{-1} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i'$ is the sample version of $\mathbb{E}[\mathbf{X}\mathbf{X}']$
- $n^{-1} \sum_{i=1}^n \mathbf{x}_i Y_i$ is the sample version of $\mathbb{E}[\mathbf{X}Y]$

Bivariate regressions

- **Bivariate regression** is the linear projection model with $\mathbf{X} = (1, X)$:

$$Y = \beta_0 + X\beta_1 + e$$

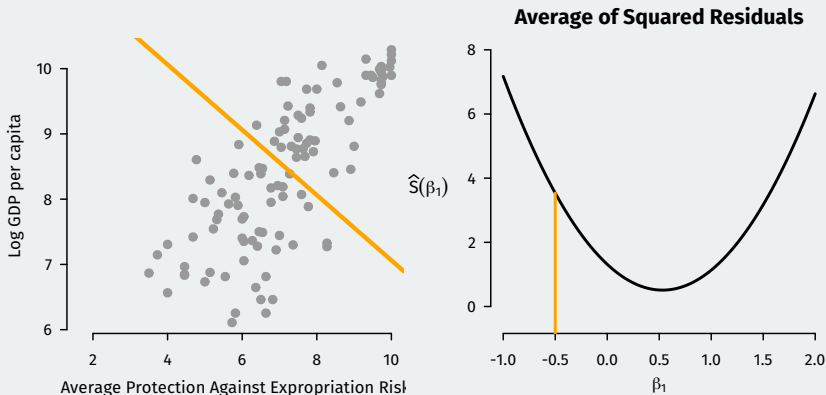
- Linear projection slope in the population from last times:

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

- We can show the OLS estimator of the slope is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\mathbb{V}}[X]}$$

Visualizing OLS



Residuals

- **Fitted value** $\widehat{Y}_i = \mathbf{X}_i' \widehat{\boldsymbol{\beta}}$ is what the model predicts at \mathbf{X}_i
 - Not really a prediction for Y_i since that was used to generate $\widehat{\boldsymbol{\beta}}$
- **Residuals** are the difference between observed and fitted values:

$$\widehat{e}_i = Y_i - \widehat{Y}_i = Y_i - \mathbf{X}_i' \widehat{\boldsymbol{\beta}}$$

- We can write $Y_i = \mathbf{X}_i' \boldsymbol{\beta} + e_i$.
 - \widehat{e}_i are not the true errors e_i
- Key **mechanical properties** of OLS residuals:

$$\sum_{i=1}^n \mathbf{X}_i \widehat{e}_i = 0$$

- Sample covariance between \mathbf{X}_i and \widehat{e}_i is 0.
 - If \mathbf{X}_i has a constant, then $n^{-1} \sum_{i=1}^n \widehat{e}_i = 0$

Prediction error

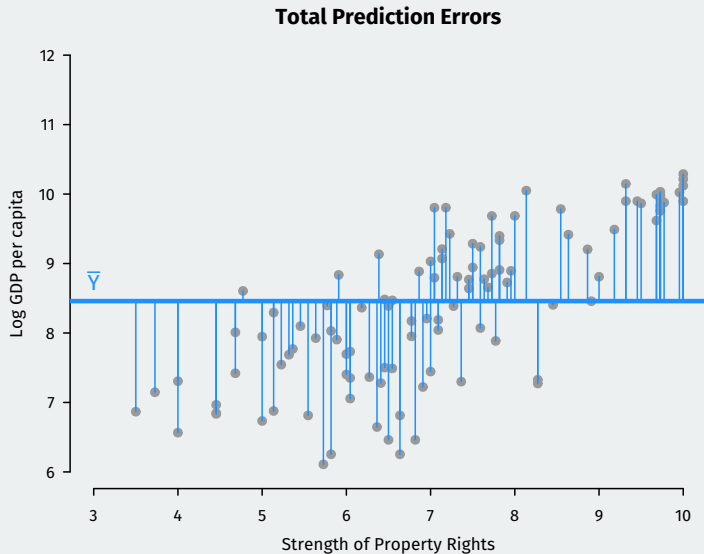
- How do we judge how well a regression fits the data?
- How much does \mathbf{X}_i help us predict Y_i ?
- **Prediction errors without \mathbf{X}_i :**
 - Best prediction is the mean, \bar{Y}
 - Prediction error is called the total sum of squares (TSS) would be:

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

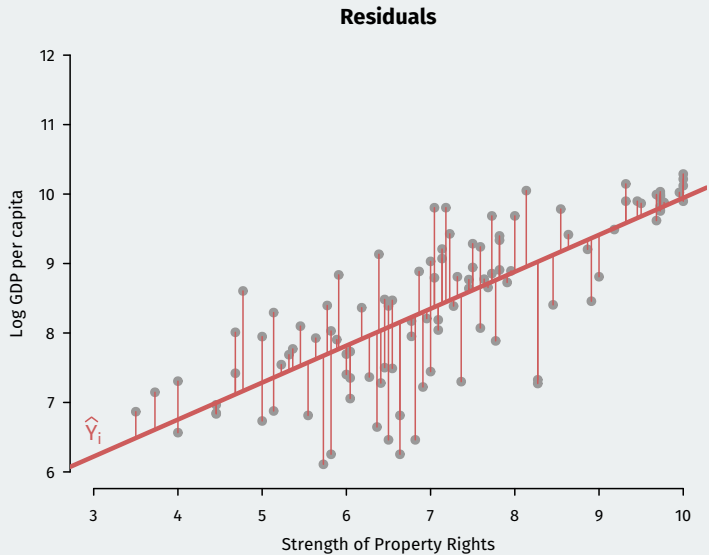
- **Prediction errors with \mathbf{X}_i :**
 - Best predictions are the fitted values, \hat{Y}_i .
 - Prediction error is the sum of the squared residuals or SSR :

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

Total SS vs SSR



Total SS vs SSR



R-squared

- Regression will always improve in-sample fit: $TSS > SSR$
- How much better does using \mathbf{X}_i do? **Coefficient of determination** or R^2 :

$$R^2 = \frac{TSS - SSR}{TSS} = 1 - \frac{SSR}{TSS}$$

- R^2 = fraction of the total prediction error eliminated by using \mathbf{X}_i .
- **Common interpretation:** R^2 is the fraction of the variation in Y_i is “explained by” \mathbf{X}_i .
 - $R^2 = 0$ means no relationship
 - $R^2 = 1$ implies perfect linear fit
- Mechanically increases with additional covariates (better fit measures exist)

Linear model in matrix form

- Linear model is a system of n linear equations:

$$Y_1 = \mathbf{X}'_1 \boldsymbol{\beta} + e_1$$

$$Y_2 = \mathbf{X}'_2 \boldsymbol{\beta} + e_2$$

$$\vdots$$

$$Y_n = \mathbf{X}'_n \boldsymbol{\beta} + e_n$$

- We can write this more compactly using matrices and vectors:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbb{X} = \begin{pmatrix} \mathbf{X}'_1 \\ \mathbf{X}'_2 \\ \vdots \\ \mathbf{X}'_n \end{pmatrix} = \begin{pmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1k} \\ 1 & X_{21} & X_{22} & \cdots & X_{2k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{nk} \end{pmatrix}, \quad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

- Model is now just:

$$\mathbf{Y} = \mathbb{X} \boldsymbol{\beta} + \mathbf{e}$$

OLS estimator in matrix form

- Key relationship: sample sums can be written in matrix notation:

$$\sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' = \mathbb{X}'\mathbb{X}$$

$$\sum_{i=1}^n \mathbf{x}_i Y_i = \mathbb{X}'\mathbf{Y}$$

- Implies we can write the OLS estimator as

$$\hat{\boldsymbol{\beta}} = (\mathbb{X}'\mathbb{X})^{-1} \mathbb{X}'\mathbf{Y}$$

- Residuals:

$$\hat{\mathbf{e}} = \mathbf{Y} - \mathbb{X}\hat{\boldsymbol{\beta}} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} - \begin{bmatrix} 1\hat{\beta}_0 + X_{11}\hat{\beta}_1 + X_{12}\hat{\beta}_2 + \cdots + X_{1k}\hat{\beta}_k \\ 1\hat{\beta}_0 + X_{21}\hat{\beta}_1 + X_{22}\hat{\beta}_2 + \cdots + X_{2k}\hat{\beta}_k \\ \vdots \\ 1\hat{\beta}_0 + X_{n1}\hat{\beta}_1 + X_{n2}\hat{\beta}_2 + \cdots + X_{nk}\hat{\beta}_k \end{bmatrix}$$

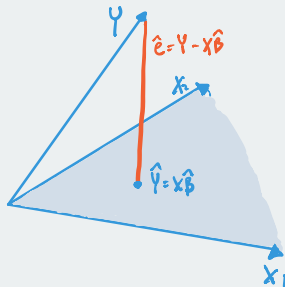
Least squares in matrix form

- OLS still minimizes sum of the squared residuals

$$\arg \min_{\mathbf{b} \in \mathbb{R}^{k+1}} \hat{\mathbf{e}}' \hat{\mathbf{e}} = \arg \min_{\mathbf{b} \in \mathbb{R}^{k+1}} (\mathbf{Y} - \mathbb{X}\mathbf{b})' (\mathbf{Y} - \mathbb{X}\mathbf{b})$$

- We can write the covariate-residual orthogonality as $\mathbb{X}'\hat{\mathbf{e}} = 0$.

Projection



- OLS can be seen as a projection of \mathbf{Y} onto the column space of \mathbf{X} , $\mathcal{S}(\mathbf{X})$.
 - Picture with $n = 3$ and $k = 2$: points in 3D space,
 - Column space of \mathbf{X} is a plane in this space.
- Intuition: $\hat{\boldsymbol{\beta}}$ defines the projection that gets is shortest distance between \mathbf{Y} and prediction.

Projection/hat matrix

- There are a couple of very important matrices in OLS algebra.
- **Projection matrix** $\mathbf{P} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$
- Also called the **hat matrix** it puts the “hat” on \mathbf{Y} :

$$\mathbf{PY} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \mathbf{X}\hat{\boldsymbol{\beta}} = \hat{\mathbf{Y}}$$

- Key properties:
 - \mathbf{P} is an $n \times n$ symmetric matrix
 - \mathbf{P} is **idempotent**: $\mathbf{PP} = \mathbf{P}$
 - Projecting \mathbf{X} onto itself returns itself: $\mathbf{PX} = \mathbf{X}$

Annihilator matrix

- **Annihilator matrix** projects onto the space spanned by the residual:

$$\mathbf{M} = \mathbf{I}_n - \mathbf{P} = \mathbf{I}_n - \mathbb{X}(\mathbb{X}'\mathbb{X})^{-1}\mathbb{X}'$$

- Also called the **residual maker**:

$$\mathbf{M}\mathbf{Y} = (\mathbf{I}_n - \mathbf{P})\mathbf{Y} = \mathbf{Y} - \mathbf{P}\mathbf{Y} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{e}$$

- “Annihilates” any function in the column space of \mathbb{X} , $\mathcal{S}(\mathbb{X})$:

$$\mathbf{M}\mathbb{X} = (\mathbf{I}_n - \mathbf{P})\mathbb{X} = \mathbb{X} - \mathbf{P}\mathbb{X} = \mathbb{X} - \mathbb{X} = \mathbf{0}$$

- Properties:

- \mathbf{M} is a symmetric $n \times n$ matrix.
- \mathbf{M} is idempotent so that $\mathbf{M}\mathbf{M} = \mathbf{M}$
- Admits a nice expression for the residual vector: $\hat{\mathbf{e}} = \mathbf{M}\mathbf{e}$

Partitioned regression

- Partition covariates and coefficients $\mathbb{X} = [\mathbb{X}_1 \ \mathbb{X}_2]$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)'$:

$$\mathbf{Y} = \mathbb{X}_1\boldsymbol{\beta}_1 + \mathbb{X}_2\boldsymbol{\beta}_2 + \mathbf{e}$$

- Can we find expressions for $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$?
- We can find $\hat{\boldsymbol{\beta}}_1$ by nested minimization:

$$\hat{\boldsymbol{\beta}}_1 = \arg \min_{\boldsymbol{\beta}_1} \left(\min_{\boldsymbol{\beta}_2} (\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2)' (\mathbf{Y} - \mathbb{X}_1\boldsymbol{\beta}_1 - \mathbb{X}_2\boldsymbol{\beta}_2) \right)$$

- First find the minimum of the sum of the squared residuals over $\boldsymbol{\beta}_2$ fixing $\boldsymbol{\beta}_1$
- Then find $\boldsymbol{\beta}_1$ that minimizes the resulting SSR.

Partitioned regression results

- The projection and annihilator matrices are defined only by covariates.
 - $\mathbf{M}_1 = \mathbf{I}_n - \mathbb{X}_1(\mathbb{X}_1'\mathbb{X}_1)^{-1}\mathbb{X}_1'$
 - $\mathbf{M}_2 = \mathbf{I}_n - \mathbb{X}_2(\mathbb{X}_2'\mathbb{X}_2)^{-1}\mathbb{X}_2'$
 - Creates residuals from a regression on \mathbb{X}_1 or \mathbb{X}_2
- Solving the nested minimization gives:

$$\hat{\beta}_1 = (\mathbb{X}_1'\mathbf{M}_2\mathbb{X}_1)^{-1} (\mathbb{X}_1'\mathbf{M}_2\mathbf{Y})$$

$$\hat{\beta}_2 = (\mathbb{X}_2'\mathbf{M}_1\mathbb{X}_2)^{-1} (\mathbb{X}_2'\mathbf{M}_1\mathbf{Y})$$

- When will $\hat{\beta}_1$ will be the same regardless of whether \mathbb{X}_2 is included?
 - If \mathbb{X}_1 and \mathbb{X}_2 are orthogonal so $\mathbb{X}_2'\mathbb{X}_1 = 0$ so $\mathbf{M}_2\mathbb{X}_1 = \mathbb{X}_1$

Residual regression

- Define two sets of residuals:
 - $\tilde{\mathbb{X}}_2 = \mathbf{M}_1 \mathbb{X}_2$ = residuals from regression of \mathbb{X}_2 on \mathbb{X}_1
 - $\tilde{\mathbf{e}}_1 = \mathbf{M}_1 \mathbf{Y}$ = residuals from regression of \mathbf{Y} on \mathbb{X}_1 .
- Then remembering that \mathbf{M}_1 is symmetric and idempotent:

$$\begin{aligned}\hat{\beta}_2 &= (\mathbb{X}_2' \mathbf{M}_1 \mathbb{X}_2)^{-1} (\mathbb{X}_2' \mathbf{M}_1 \mathbf{Y}) \\ &= (\mathbb{X}_2' \mathbf{M}_1 \mathbf{M}_1 \mathbb{X}_2)^{-1} (\mathbb{X}_2' \mathbf{M}_1 \mathbf{M}_1 \mathbf{Y}) \\ &= (\tilde{\mathbb{X}}_2' \tilde{\mathbb{X}}_2)^{-1} (\tilde{\mathbb{X}}_2' \tilde{\mathbf{e}}_1)\end{aligned}$$

- $\hat{\beta}_2$ can be obtained from a regression of $\tilde{\mathbf{e}}_1$ on $\tilde{\mathbb{X}}_2$.
 - Same result applies when using \mathbf{Y} in place of $\tilde{\mathbf{e}}_1$.
 - Intuition: residuals are orthogonal
 - Called the **Frisch-Waugh-Lovell Theorem**
 - Sample version of the results we saw for the linear projection.

Outliers, leverage points, and influential observations

- Least square heavily penalizes large residuals.
- Implies a just a few unusual observations can be extremely influential.
 - Dropping them leads to large changes in the estimated $\hat{\beta}$.
 - Not all “unusual” observations have the same effect, though.
- Useful to categorize:
 1. **Leverage point:** extreme in one X direction
 2. **Outlier:** extreme in the Y direction
 3. **Influence point:** extreme in both directions

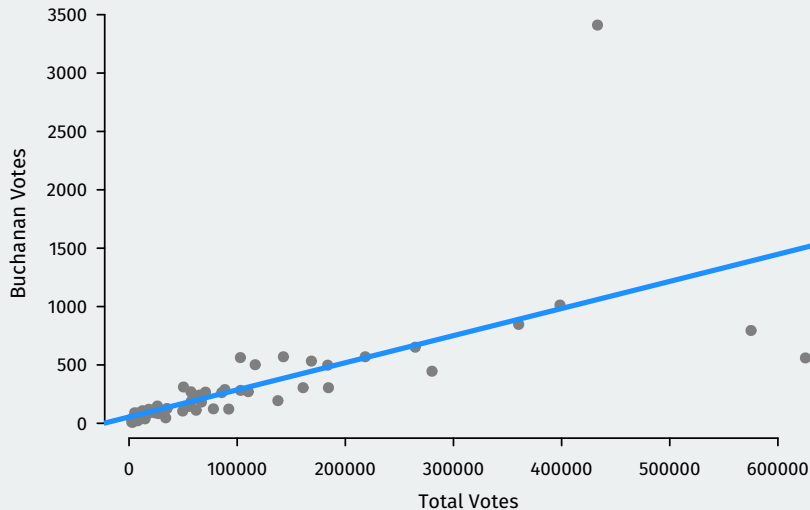
Example: Buchanan votes in Florida, 2000

- 2000 Presidential election in FL (Wand et al., 2001, APSR)

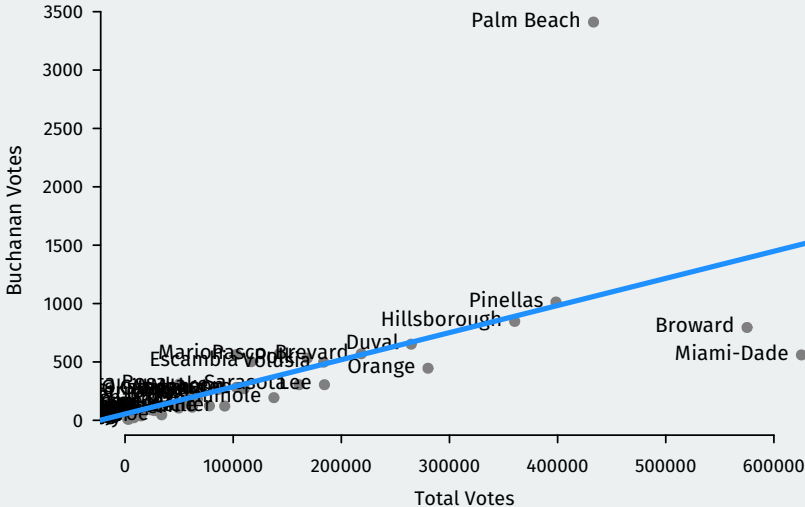
OFFICIAL BALLOT, GENERAL ELECTION PALM BEACH COUNTY, FLORIDA NOVEMBER 7, 2000		
es will electors.)	(REPUBLICAN)	
	GEORGE W. BUSH - PRESIDENT	3 ➡
	DICK CHENEY - VICE PRESIDENT	
	(DEMOCRATIC)	
	AL GORE - PRESIDENT	5 ➡
	JOE LIEBERMAN - VICE PRESIDENT	
	(LIBERTARIAN)	
	HARRY BROWNE - PRESIDENT	7 ➡
	ART OLIVIER - VICE PRESIDENT	
	(GREEN)	
	RALPH NADER - PRESIDENT	9 ➡
	WINONA LaDUKE - VICE PRESIDENT	
	(SOCIALIST WORKERS)	
JAMES HARRIS - PRESIDENT	11 ➡	
MARGARET TROWE - VICE PRESIDENT		
(NATURAL LAW)		
JOHN HAGELIN - PRESIDENT	13 ➡	
NAT GOLDBABER - VICE PRESIDENT		

OFFICIAL BALLOT, GENERAL ELECTION PALM BEACH COUNTY, FLORIDA NOVEMBER 7, 2000	
4 ⬅	(REFORM) PAT BUCHANAN - PRESIDENT EZOLA FOSTER - VICE PRESIDENT
6 ⬅	(SOCIALIST) DAVID McREYNOLDS - PRESIDENT MARY CAL HOLLIS - VICE PRESIDENT
8 ⬅	(CONSTITUTION) HOWARD PHILLIPS - PRESIDENT J. CURTIS FRAZIER - VICE PRESIDENT
10 ⬅	(WORKERS WORLD) MONICA MOOREHEAD - PRESIDENT GLORIA La RIVA - VICE PRESIDENT
WRITE-IN CANDIDATE To vote for a write-in candidate, follow the directions on the long stub of your ballot card.	

Example: Buchanan votes in Florida, 2000



Example: Buchanan votes in Florida, 2000

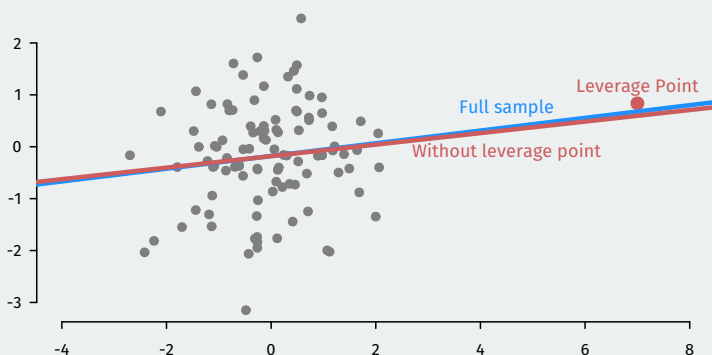


Example: Buchanan votes

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

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  54.22945    49.14146    1.10    0.27
## edaytotal     0.00232     0.00031    7.48 2.4e-10 ***
## ---
## Signif. codes:
## 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 333 on 65 degrees of freedom
## Multiple R-squared:  0.463, Adjusted R-squared:  0.455
## F-statistic: 56 on 1 and 65 DF, p-value: 2.42e-10
```

Leverage point definition



- Values that are extreme in the X dimension
- That is, values far from the center of the covariate distribution

Leverage values

- Let h_{ij} be the (i, j) entry of \mathbf{P} . Then:

$$\hat{\mathbf{Y}} = \mathbf{PY} \quad \Rightarrow \quad \hat{Y}_i = \sum_{j=1}^n h_{ij} Y_j$$

- h_{ij} = importance of observation j is for the fitted value \hat{Y}_i
- Leverage/hat values:** h_{ii} diagonal entries of the hat matrix
- With a simple linear regression, we have

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{\sum_{j=1}^n (X_j - \bar{X})^2}$$

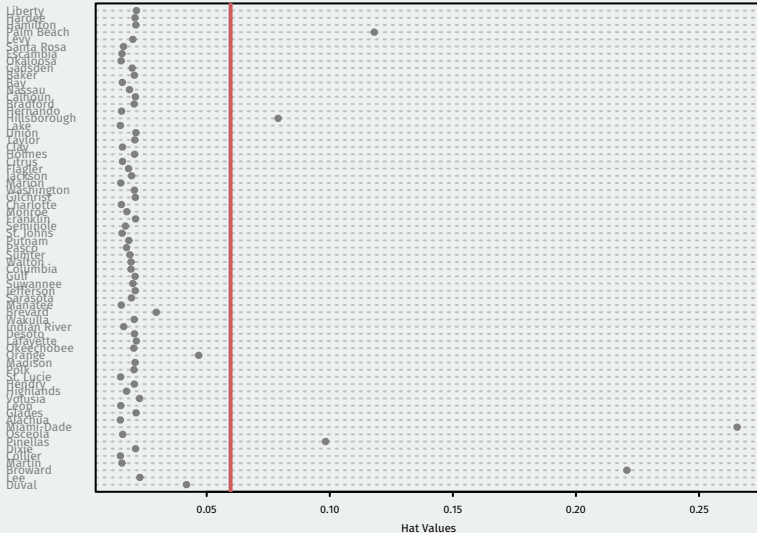
- \rightsquigarrow how far i is from the center of the X distribution
- Rule of thumb:** examine hat values greater than $2(k+1)/n$

Buchanan hats

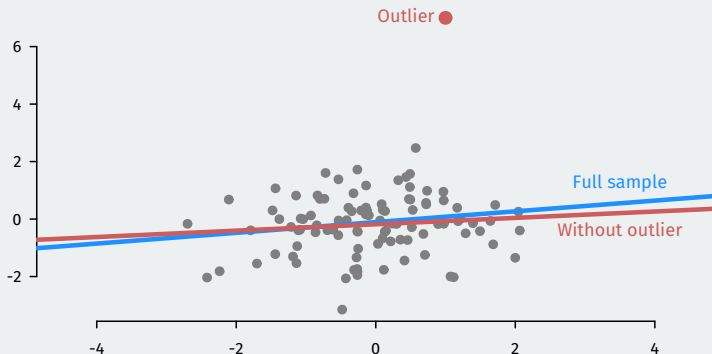
```
head(hatvalues(mod), 5)
```

```
##           1           2           3           4           5  
## 0.0418 0.0228 0.2207 0.0156 0.0149
```


Buchanan hats



Outlier definition



- An **outlier** is far away from the center of the Y distribution.
- Intuitively: a point that would be poorly predicted by the regression.

Detecting outliers

- Want values poorly predicted? Look for big residuals, right?
 - Problem: we use i to estimate $\hat{\beta}$ so \hat{Y} aren't valid predictions.
 - unit might pull the regression line toward itself \rightsquigarrow small residual
- Better: **leave-one-out prediction errors**,

1. Regress $\mathbf{Y}_{(-i)}$ on $\mathbb{X}_{(-i)}$, where these omit unit i :

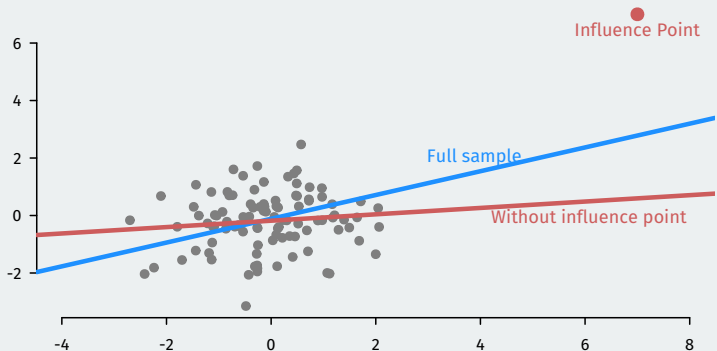
$$\hat{\beta}_{(-i)} = (\mathbb{X}'_{(-i)} \mathbb{X}_{(-i)})^{-1} \mathbb{X}'_{(-i)} \mathbf{Y}_{(-i)}$$

2. Calculate predicted value of Y_i using that regression: $\tilde{Y}_i = \mathbf{x}'_i \hat{\beta}_{(-i)}$
3. Calculate prediction error: $\tilde{e}_i = Y_i - \tilde{Y}_i$

- Simple closed-form expressions:

$$\hat{\beta}_{(-i)} = \hat{\beta} - (\mathbb{X}'\mathbb{X})^{-1} \mathbf{x}_i \tilde{e}_i \quad \tilde{e}_i = \frac{\hat{e}_i}{1 - h_{ii}}$$

Influence points



- An **influence point** is one that is both an outlier and a leverage point.
- Extreme in both the X and Y dimensions

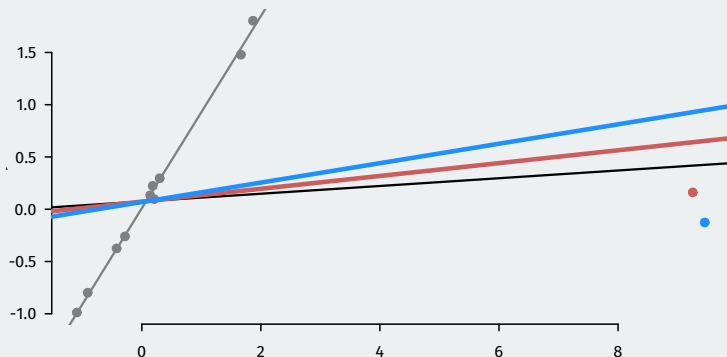
Overall measures of influence

- Influence of i can be measured by change in predictions:

$$\widehat{Y}_i - \widetilde{Y}_i = h_{ii}\tilde{e}_i$$

- How much does excluding i from the regression change its predicted value?
- Equal to “leverage \times outlier-ness”
- Lots of diagnostics exist, but are mostly heuristic.
 - Does removing the point change a coefficient by a lot?

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

What to do about outliers and influential units?

- 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, least absolute deviations)