GOV 2000 Section 2

Konstantin Kashin[1]
Harvard University

September 12, 2012

[1]These notes and accompanying code draw on the notes from Molly Roberts, Maya Sen, Iain Osgood, Brandon Stewart, and TF's from previous years

Outline

Administrative Details

Regression

Goodness-of-Fit: R^2

Appendix

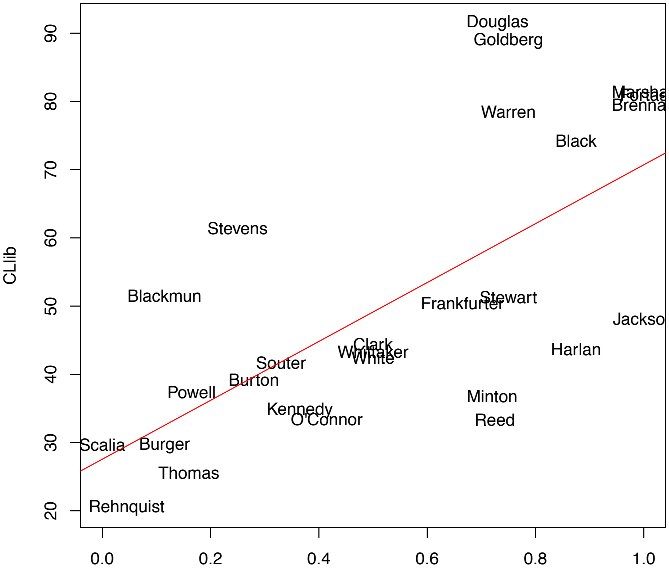
Problem Set Expectations

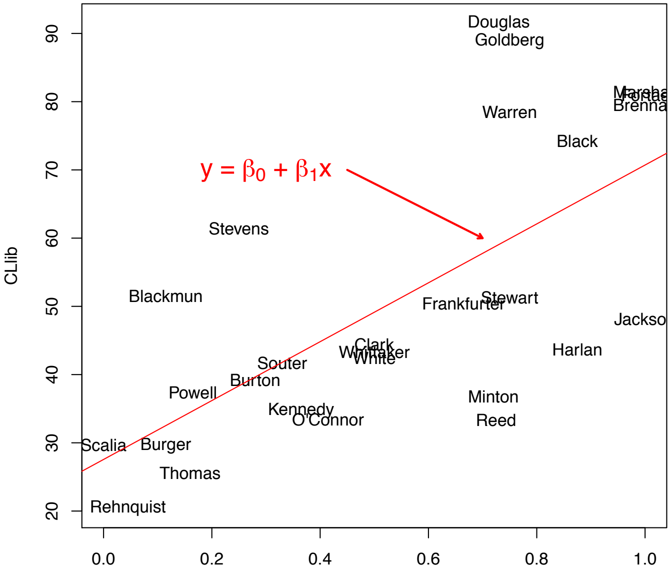
- First problem set distributed yesterday.
- Due next Tuesday.
- Must be typeset (LaTeX or Word) and submitted electronically
- Submit source-able, commented code with journal-quality graphics

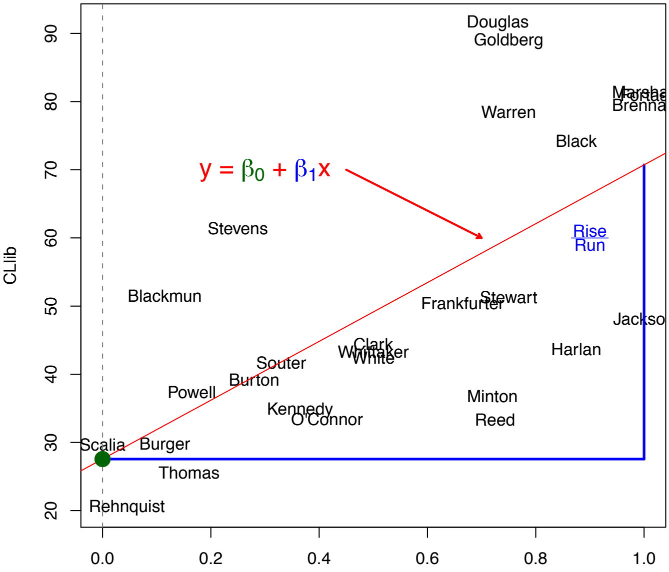
Getting Help

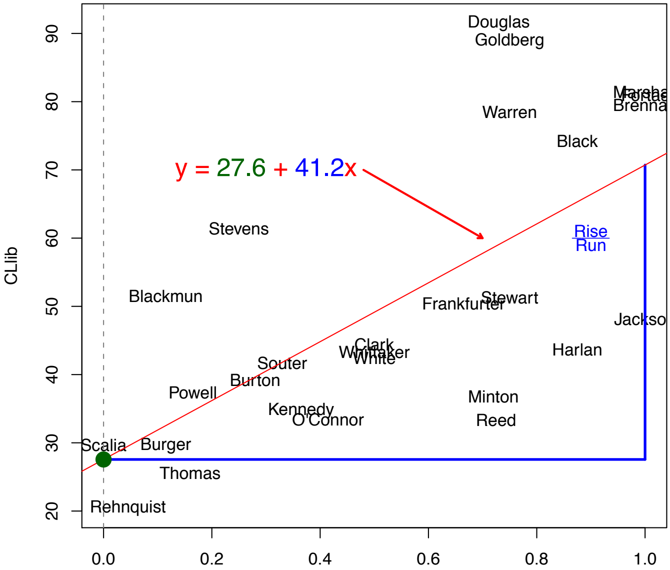
1. General questions should go to email list.
2. Response time: 24-response times during week, longer on weekends.
3. Office hours:
   - Adam: Tuesdays 4-6pm
   - Andy: Mondays 9-11am
   - Konstantin: Fridays 2-4pm (EXCEPT FOR THIS WEEK: 11am-1pm)
4. Formula Wiki

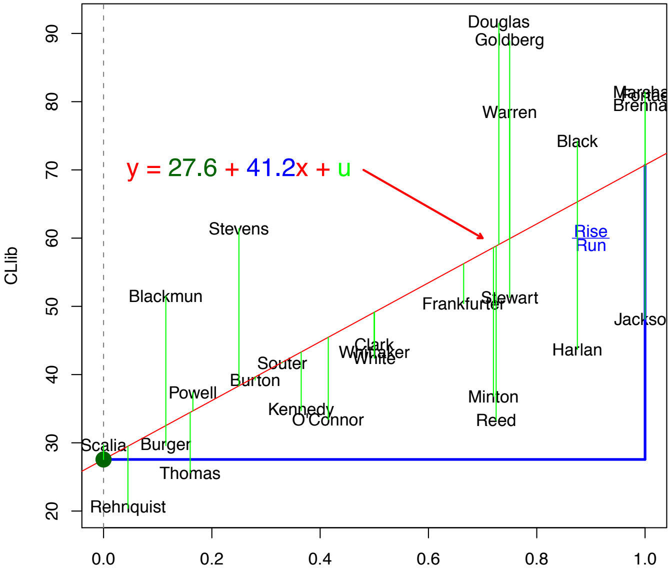
Outline


CLib
90
80
70
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
1.0
Douglas
Goldberg
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Harlan
Clark
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist


y = β0 + β1x
CLib
Douglas
Goldberg
Marshall
Fortas
Brennan
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackson
Clark
Whittaker
White
Harlan
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist
90
80
70
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
1.0


y = β0 + β1x
Rise
Run
CLib
Douglas
Goldberg
Marshall
Fortas
Brennan
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackson
Clark
Whittaker
White
Harlan
Souter
Burton
Powell
Minton
Kennedy
O'Connor
Reed
Scalia
Burger
Thomas
Rehnquist
20
30
40
50
60
70
80
90
0.0
0.2
0.4
0.6
0.8
1.0


y = 27.6 + 41.2x
Rise
Run
CLib
Douglas
Goldberg
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Harlan
Clark
Whittaker
White
Souter
Burton
Powell
Kennedy
O'Connor
Minton
Reed
Scalia
Burger
Thomas
Rehnquist
20
30
40
50
60
70
80
90
0.0
0.2
0.4
0.6
0.8
1.0


y = 27.6 + 41.2x + u
Rise
Run
CLib
Douglas
Goldberg
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Harlan
Clark
Whittaker
White
Souter
Burton
Powell
Kennedy
O'Connor
Minton
Reed
Scalia
Burger
Thomas
Rehnquist
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90

How to get B_0 and B_1

$$\hat{B}_0 = \bar{y} - \hat{B}_1 \bar{x}.$$

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

Outline

R^2

How much variance are we explaining?

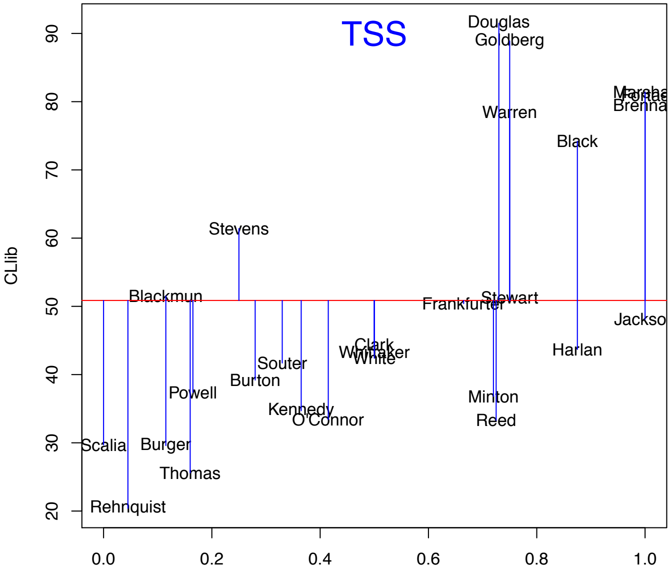
To calculate R^2, we need to think about the following two quantities:

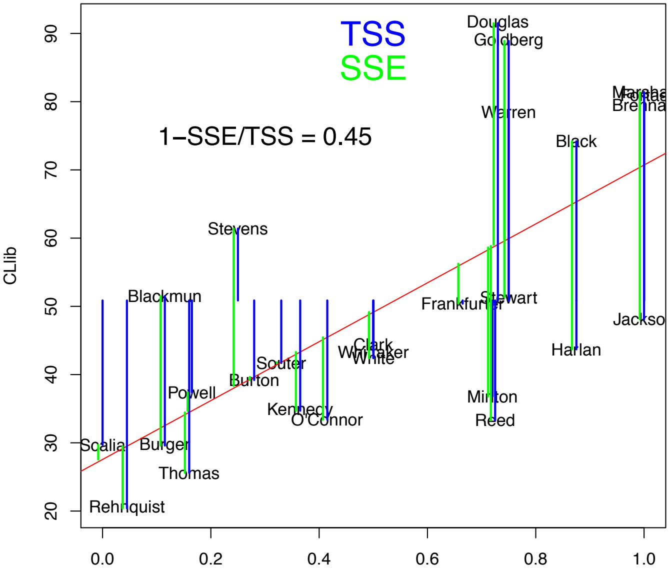
1. TSS: Total sum of squares
2. SSR: Sum of squared residuals

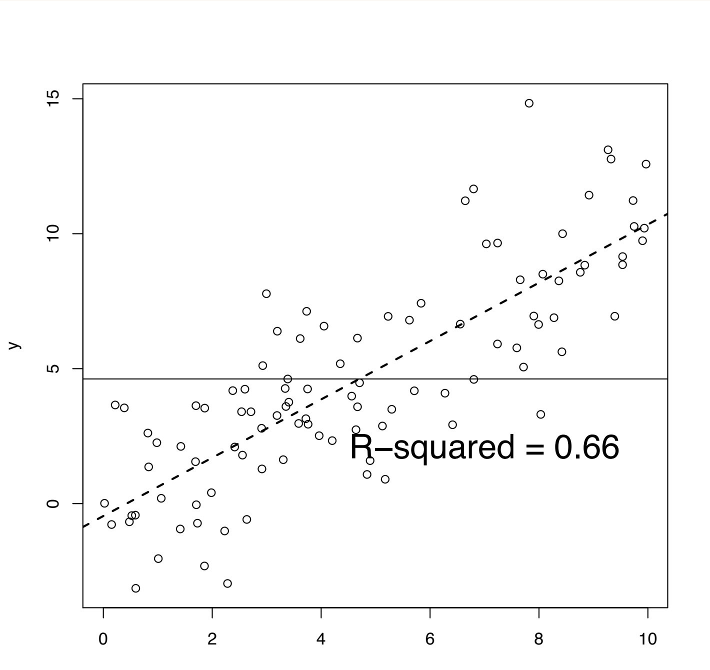
$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2.$$

$$SSR = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

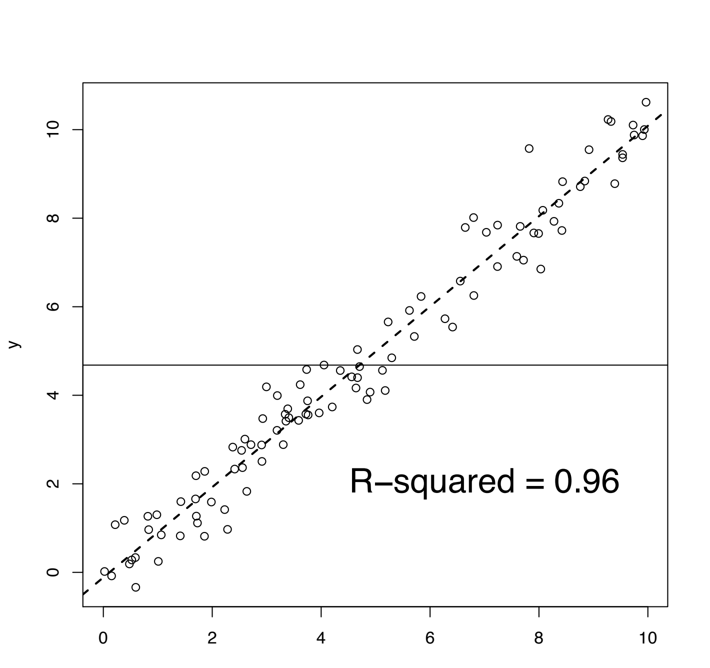
$$R^2 = 1 - \frac{SSR}{TSS}.$$

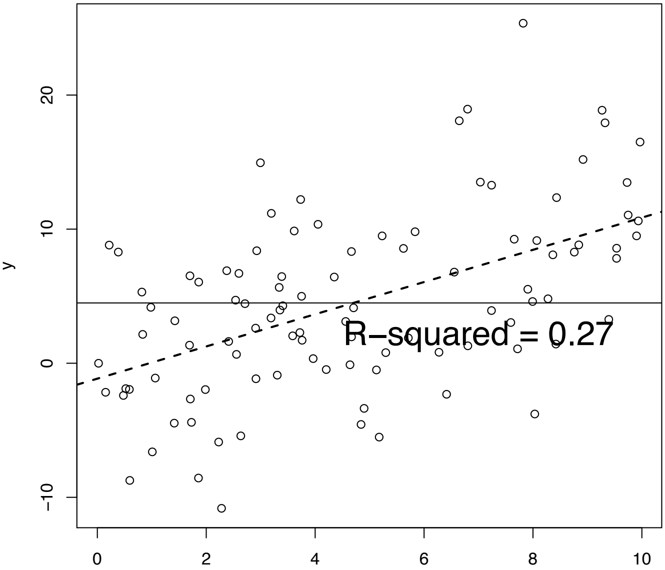

TSS
CLib
90
80
70
60
50
40
30
20
0.0
0.2
0.4
0.6
0.8
1.0
Douglas
Goldberg
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackso
Clark
Whittaker
White
Souter
Burton
Powell
Kennedy
O'Connor
Minton
Reed
Harlan
Scalia
Burger
Thomas
Rehnquist


TSS
SSE
1−SSE/TSS = 0.45
CLib
Douglas
Goldberg
Marshall
Fortas
Brennan
Warren
Black
Stevens
Blackmun
Frankfurter
Stewart
Jackson
Clark
Whittaker
White
Souter
Burton
Powell
Kennedy
O'Connor
Minton
Reed
Harlan
Scalia
Burger
Thomas
Rehnquist
0.0
0.2
0.4
0.6
0.8
1.0
20
30
40
50
60
70
80
90



15
10
5
0
y
0
2
4
6
8
10
R–squared = 0.66


R–squared = 0.96
y
0
2
4
6
8
10
0
2
4
6
8
10


R−squared = 0.27
y
20
10
0
−10
0
2
4
6
8
10

Questions?

Outline

Deriving the Linear Least Squares Estimator

Let $\tilde{\beta}_0$ and $\tilde{\beta}_1$ be possible values for β_0 and β_1 respectively, and

$$S(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - x_i \tilde{\beta}_1)^2.$$

1. Take partial derivatives of S with respect to $\tilde{\beta}_0$ and $\tilde{\beta}_1$.
2. Set each of the partial derivatives to 0
3. Substitute $\hat{\beta}_0$ and $\hat{\beta}_1$ for $\tilde{\beta}_0$ and $\tilde{\beta}_1$ and solve for $\hat{\beta}_0$ and $\hat{\beta}_1$

$$S(\tilde{\beta}_0, \tilde{\beta}_1) = \sum_{i=1}^{n} (y_i - \tilde{\beta}_0 - x_i \tilde{\beta}_1)^2$$
$$= \sum_{i=1}^{n} (y_i^2 - 2y_i\tilde{\beta}_0 - 2y_i\tilde{\beta}_1 x_i + \tilde{\beta}_0^2 + 2\tilde{\beta}_0\tilde{\beta}_1 x_i + \tilde{\beta}_1^2 x_i^2)$$

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_0} = \sum_{i=1}^{n} (-2y_i + 2\tilde{\beta}_0 + 2\tilde{\beta}_1 x_i)$$

and

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_1} = \sum_{i=1}^{n} (-2y_i x_i + 2\tilde{\beta}_0 x_i + 2\tilde{\beta}_1 x_i^2)$$

We set the partial derivatives to zero

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_0} = \sum_{i=1}^{n} (-2y_i + 2\tilde{\beta}_0 + 2\tilde{\beta}_1 x_i)$$

becomes

$$\hat{\beta}_0 n = \left(\sum_{i=1}^{n} y_i\right) - \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i\right)$$

and

$$\frac{\partial S(\tilde{\beta}_0, \tilde{\beta}_1)}{\partial \tilde{\beta}_1} = \sum_{i=1}^{n} (-2y_i x_i + 2\tilde{\beta}_0 x_i + 2\tilde{\beta}_1 x_i^2)$$

becomes

$$\hat{\beta}_1 \sum_{i=1}^{n} x_1^2 = \left(\sum_{i=1}^{n} x_i y_i\right) - \hat{\beta}_0 \left(\sum_{i=1}^{n} x_i\right)$$

Normal Equations: Two equations, two unkowns.

$$\hat{\beta}_0 n = \left(\sum_{i=1}^{n} y_i\right) - \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i\right)$$

$$\hat{\beta}_1 \sum_{i=1}^{n} x_1^2 = \left(\sum_{i=1}^{n} x_i y_i\right) - \hat{\beta}_0 \left(\sum_{i=1}^{n} x_i\right)$$

Solving for $\hat{\beta}_0$ is straight forward:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Let's solve for $\hat{\beta}_1$. We have two equations. We can manipulate them by multiplying the first by $\sum_{i=1}^{n} x_i$ and the second one by n.

$$\hat{\beta}_0 n \sum_{i=1}^{n} x_i + \hat{\beta}_1 \left(\sum_{i=1}^{n} x_i\right) \sum_{i=1}^{n} x_i = \left(\sum_{i=1}^{n} y_i\right) \sum_{i=1}^{n} x_i$$

$$\hat{\beta}_0 n \left(\sum_{i=1}^{n} x_i\right) + n\hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = n \left(\sum_{i=1}^{n} x_i y_i\right)$$

Putting them together

$$\hat{\beta}_1 \left(\left(\sum_{i=1}^{n} x_i\right) \sum_{i=1}^{n} x_i - n \sum_{i=1}^{n} x_i^2\right) = \left(\sum_{i=1}^{n} y_i\right) \sum_{i=1}^{n} x_i - n \left(\sum_{i=1}^{n} x_i y_i\right)$$

Rearranging we get

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

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

$$\hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$$

We also have from the assumptions the $E(\epsilon) = 0$ and $Cov(X_i \epsilon_j) = 0$ that

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

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

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