

Bivariate relationships

Q-Step Summer Institute

29 June 2016

Andrew Eggers



[Home](#) > [About](#) > [Organisation](#) > [Finance and funding](#) > Financial Statements of the Oxford Colleges 2012-13

Financial Statements of the Oxford Colleges (2012-13)

HIS



The financial statements of the 36 colleges of Oxford University for the year ending 31 July 2013 are available as pdfs, together with an aggregated statement of financial activity (SOFA) and aggregated consolidated balance sheet.

The colleges are independent, self-governing and financially autonomous and their accounts are published under the accounting convention developed by the Charity Commissions for use by charities in the UK (the Charity SORP).

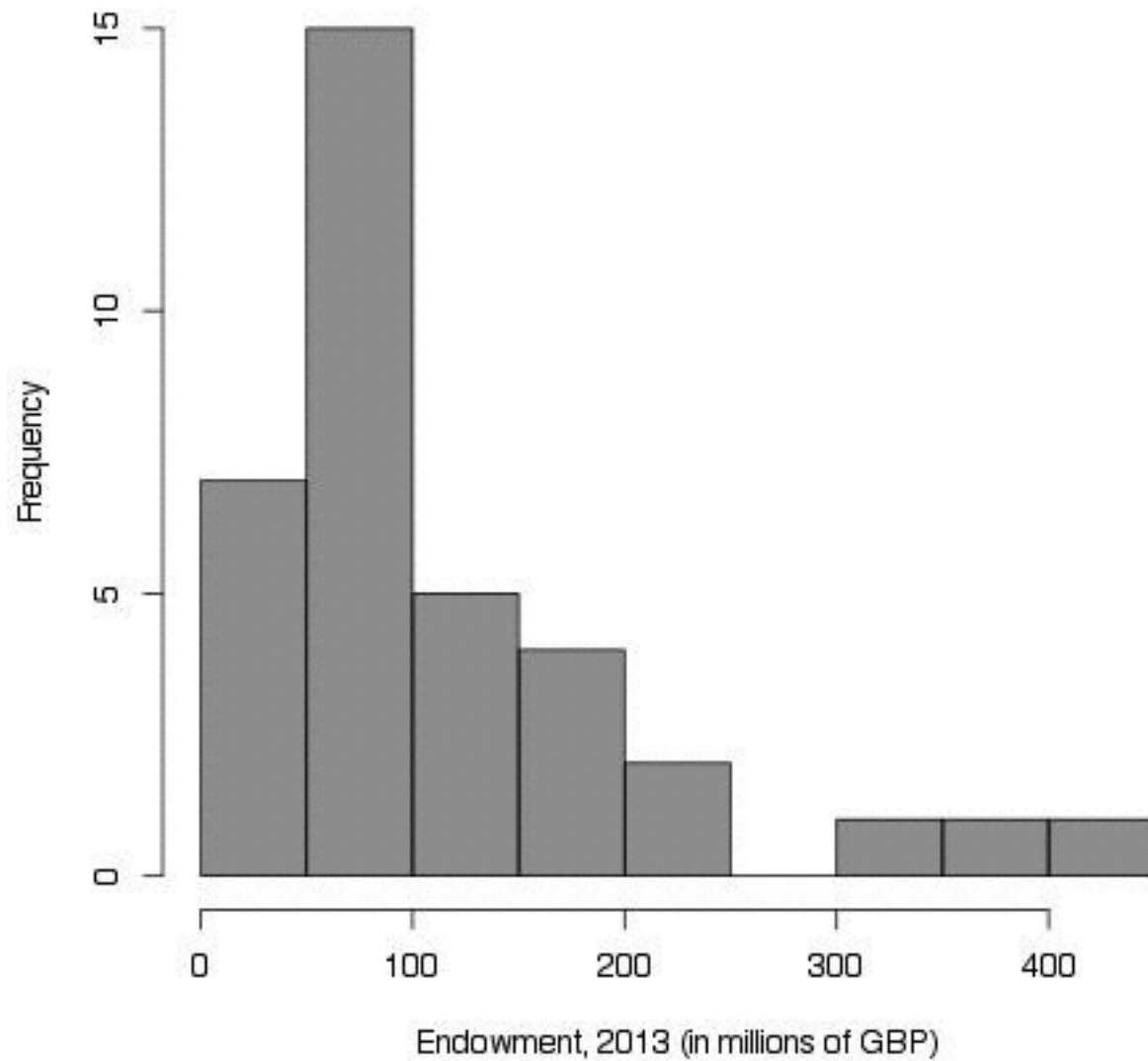
Kellogg College and St Cross College do not have Royal Charters and, for accounting purposes, are departments of the University. As such, their financial results are consolidated into the University's financial statements.

<http://goo.gl/1pJA2r>

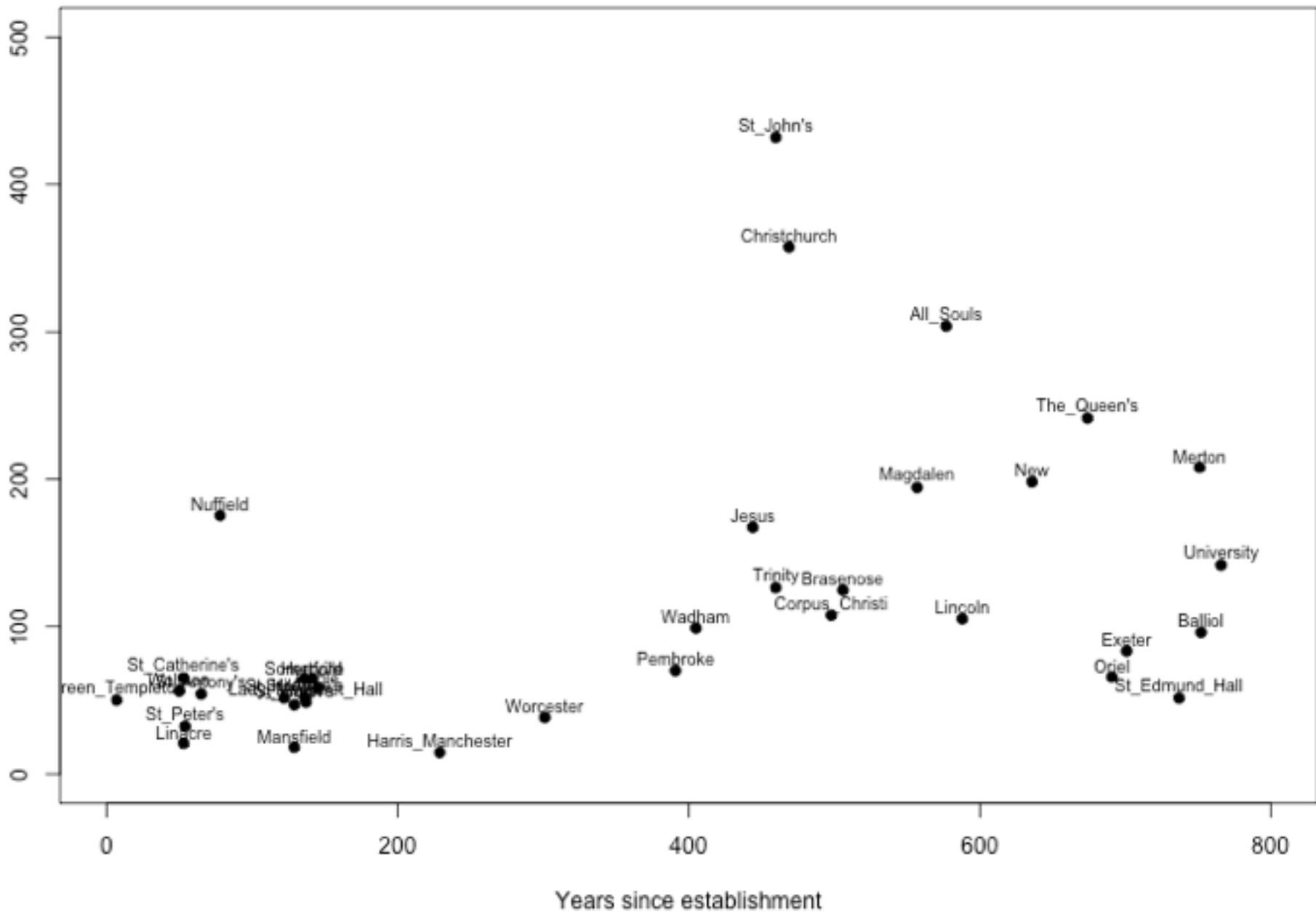
```
> ## data analysis for lecture
> d = read.csv("http://andy.egge.rs/data/college_stats_edited.csv")
> d$rank = 37 - rank(d$endowments, )
>
> # just a table of the data
> d[order(d$endowments, decreasing = T), c("rank", "college", "endowments")]
```

	rank	college	endowments
28	1	St_John's	432075
4	2	Christchurch	357667
1	3	All_Souls	303896
30	4	The_Queen's	241467
36	5	Merton	208054
17	6	New	198160
15	7	Magdalen	194344
18	8	Nuffield	175415
10	9	Jesus	167333
32	10	University	141872
31	11	Trinity	126350
3	12	Brasenose	124684
5	13	Corpus_Christi	107692
14	14	Lincoln	105245
33	15	Wadham	98935
2	16	Balliol	96044
6	17	Exeter	83383
20	18	Pembroke	70195

College endowments



Size of endowment, 2013 (millions of GBP)



Research questions you might have about Oxford colleges' endowments and age

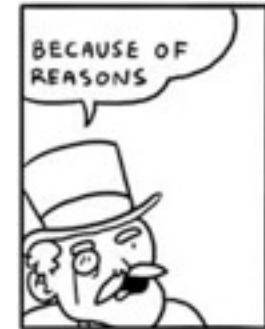
Descriptive/predictive questions:

- Do older colleges have more money?
- What is the average endowment of a college that is X years old?



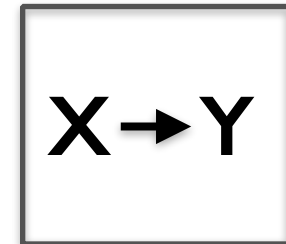
Explanatory questions (reverse causal):

- Why do some colleges have more money than others? (Maybe age is the/an answer.)



Forward causal questions:

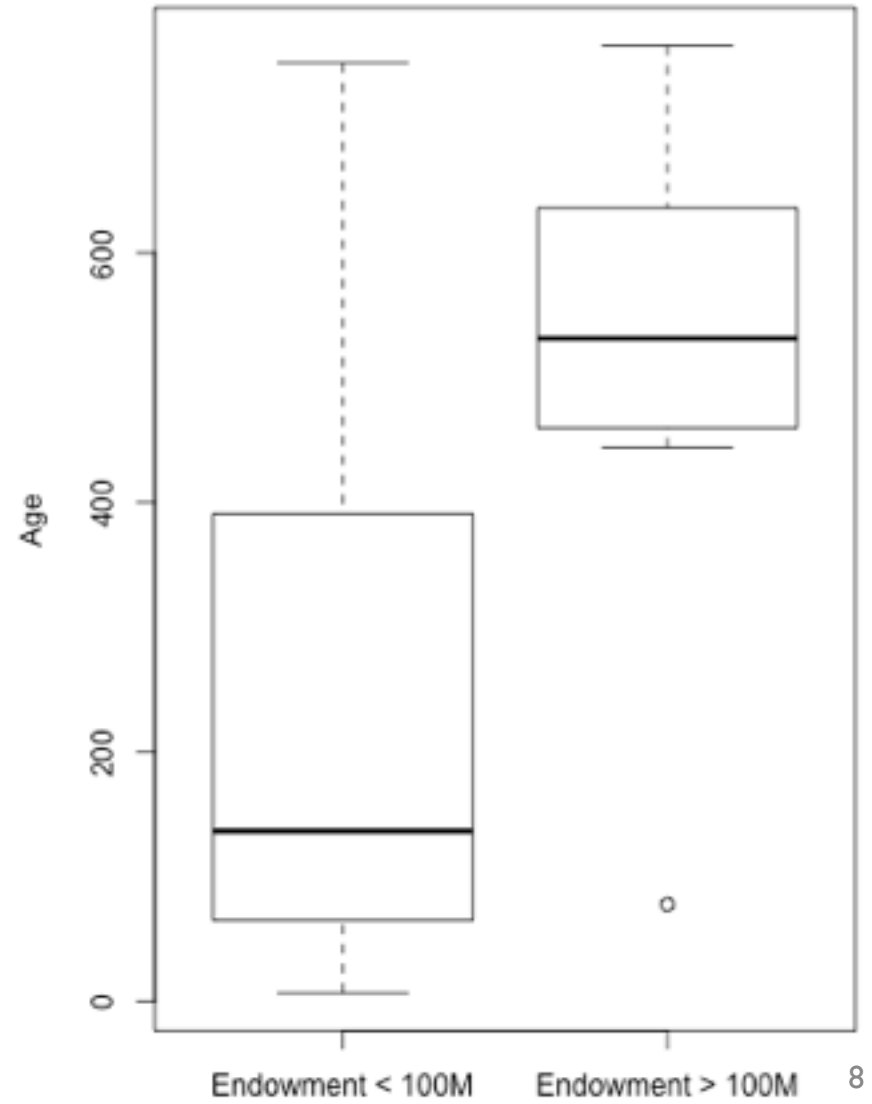
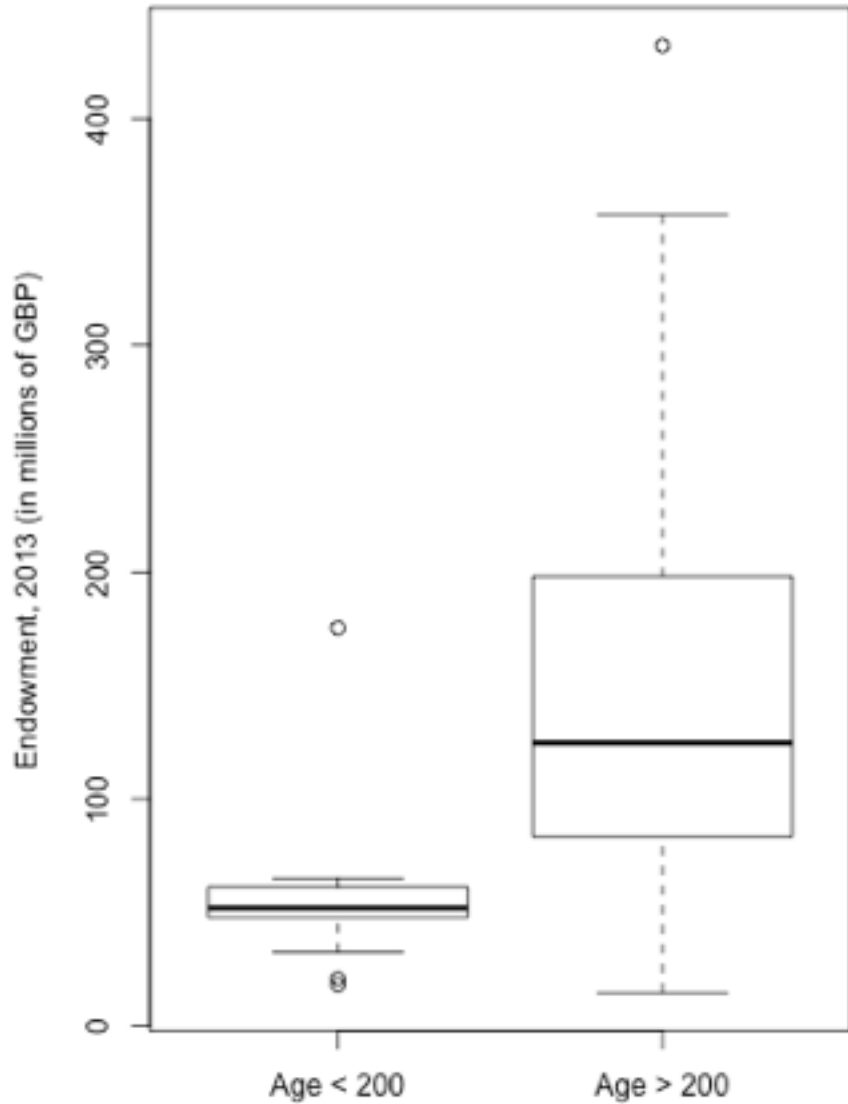
- (What is the effect of greater age on a college's endowment?)



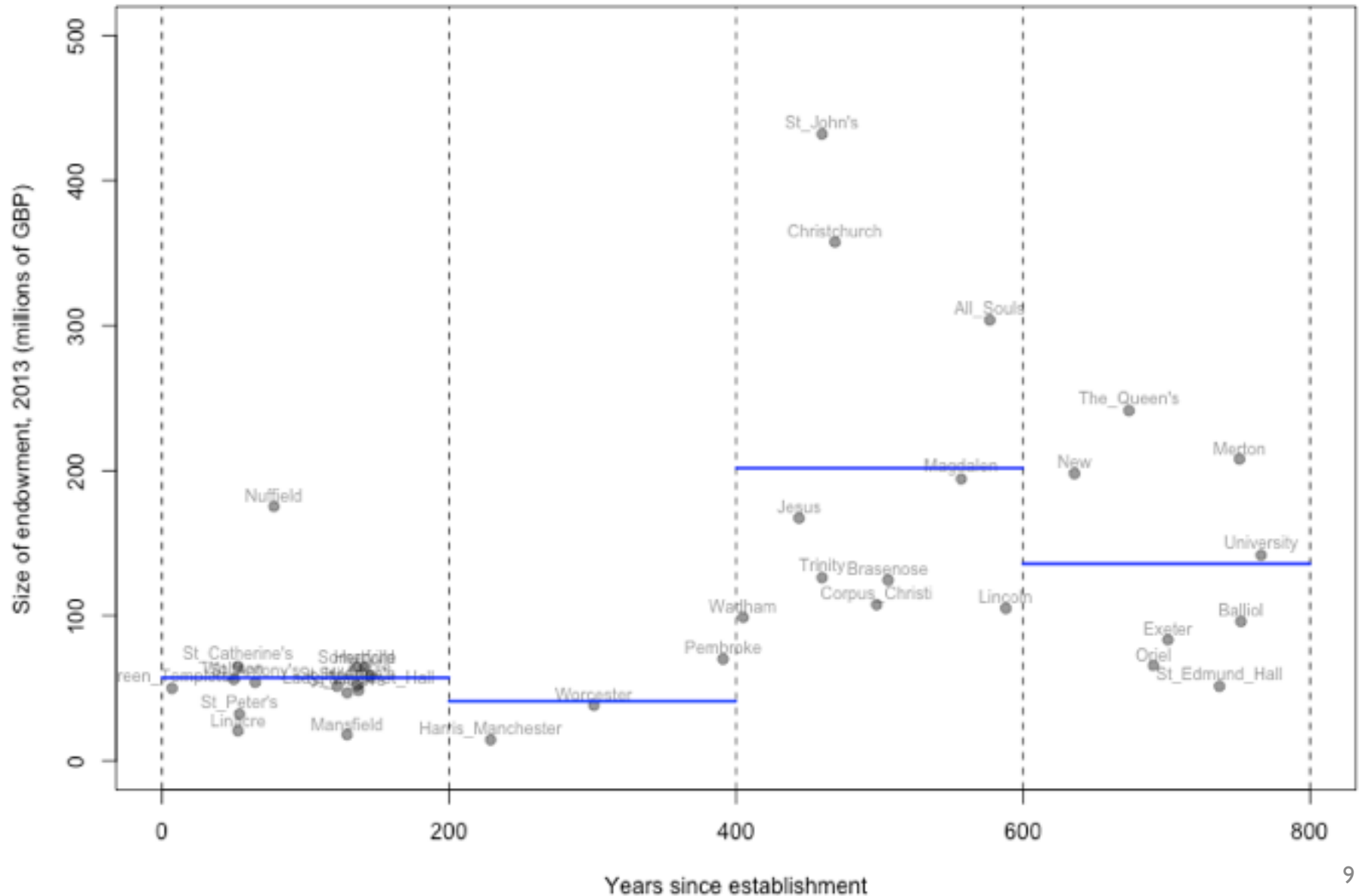
Various ways to summarize a relationship between two variables

- Show the scatterplot!
- Compare boxplots of one variable across categories of the other
- Show/report mean of one variable across categories of the other (binned means or kernel average smoother)
- Report covariance/correlation
- Report regression coefficient(s)
- Report predicted values of one variable based on the other from regression

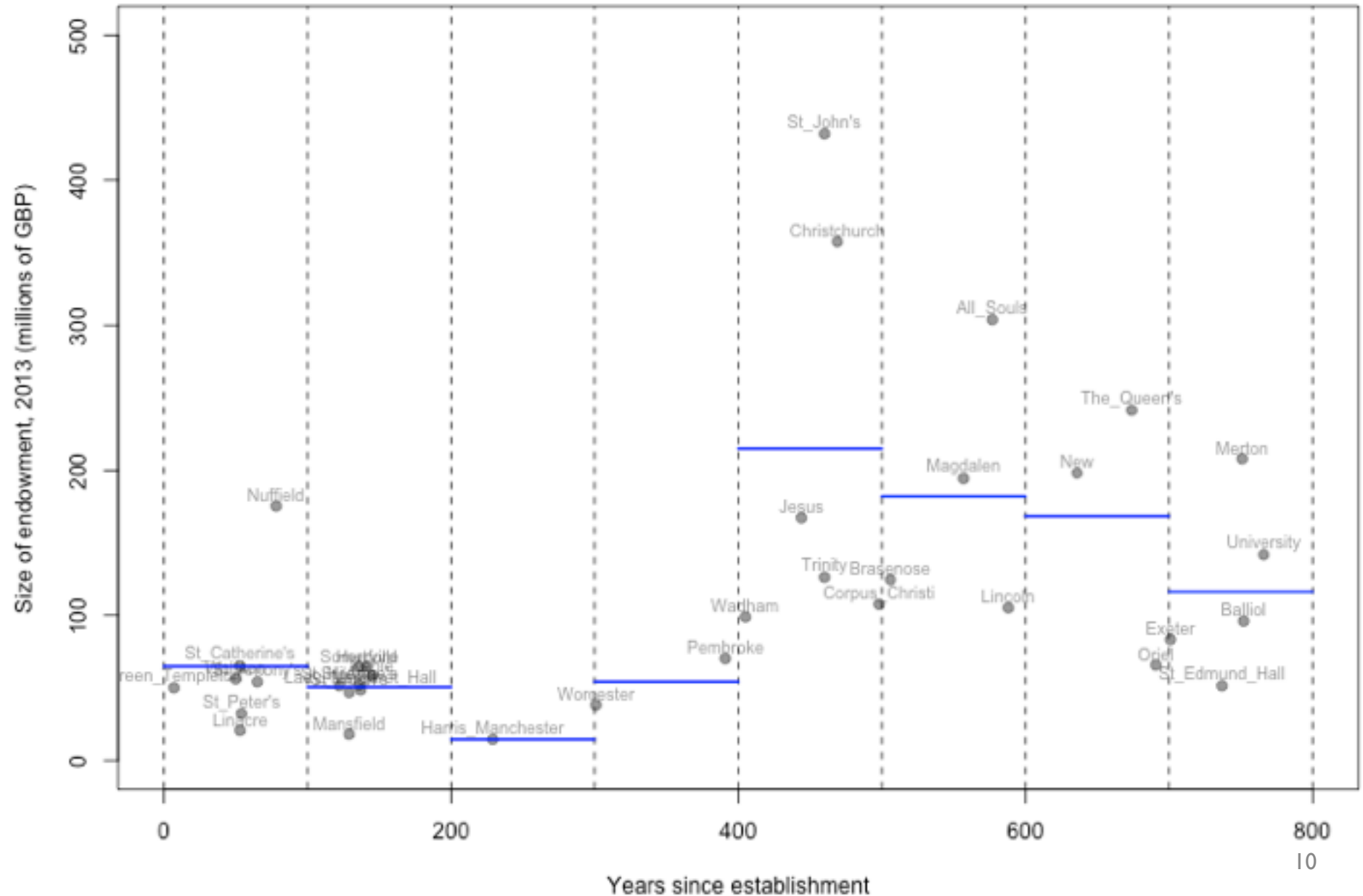
Boxplots: which do you prefer?



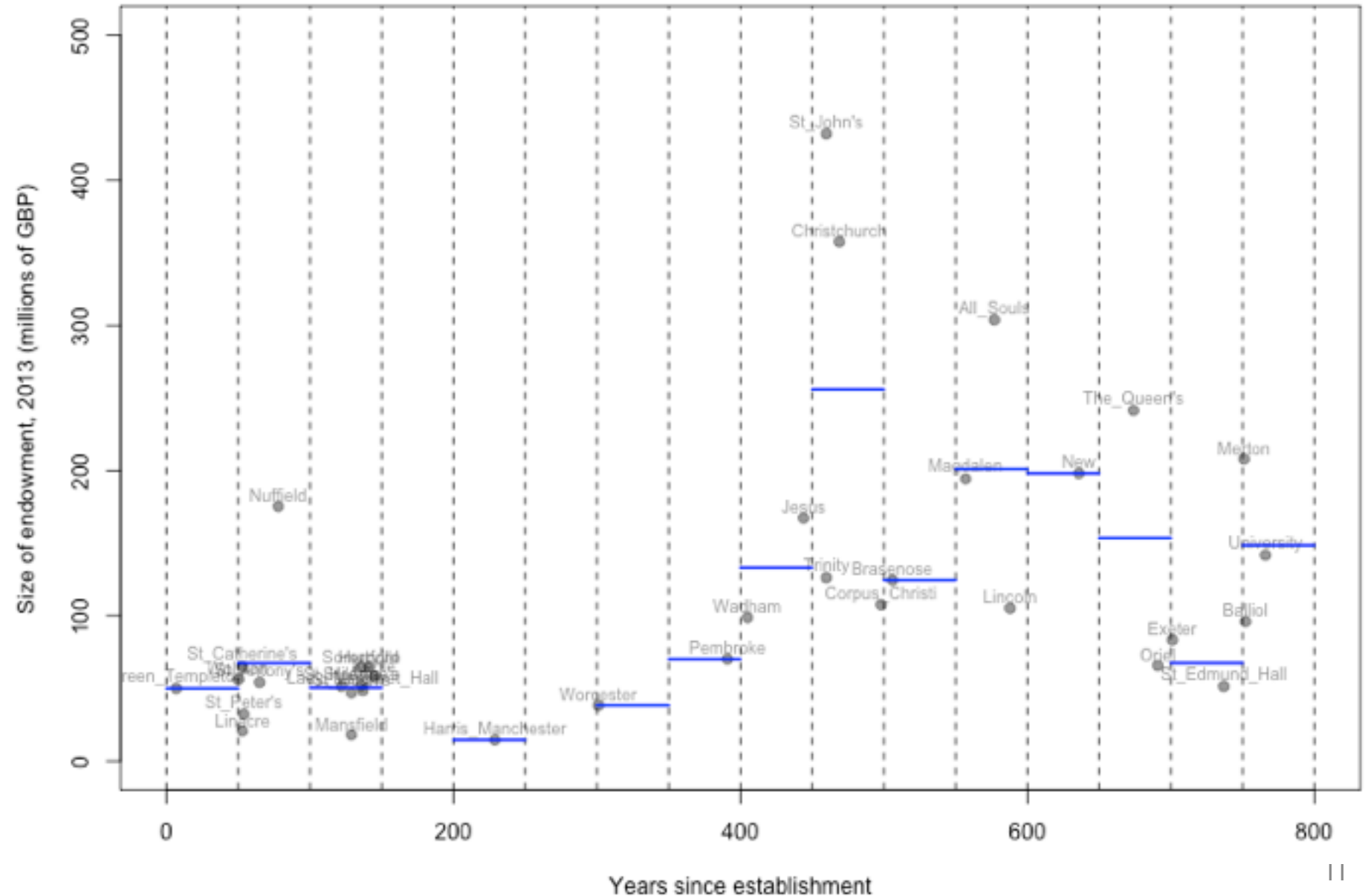
Mean endowment within 200-year intervals



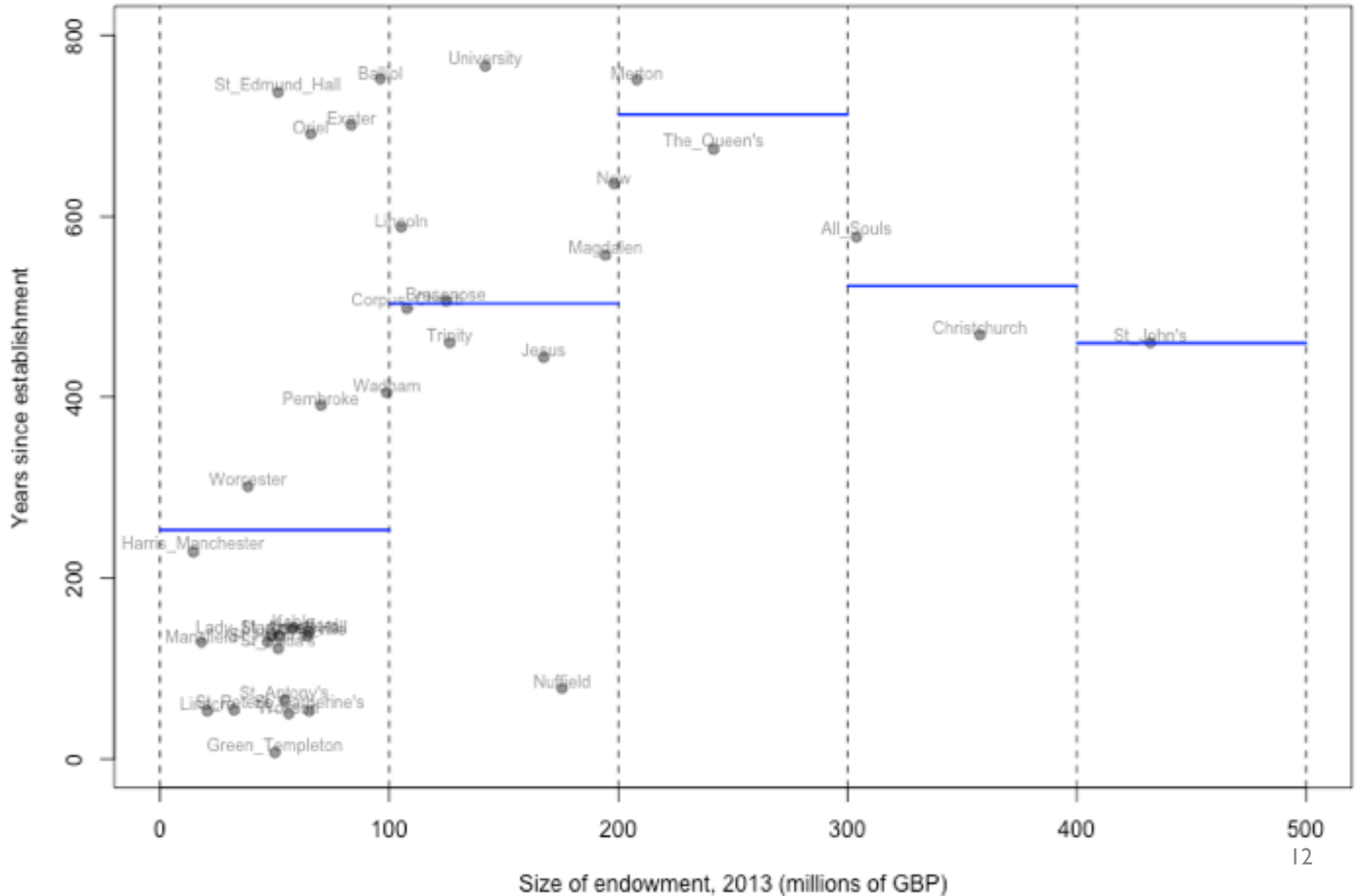
Mean endowment within 100-year intervals



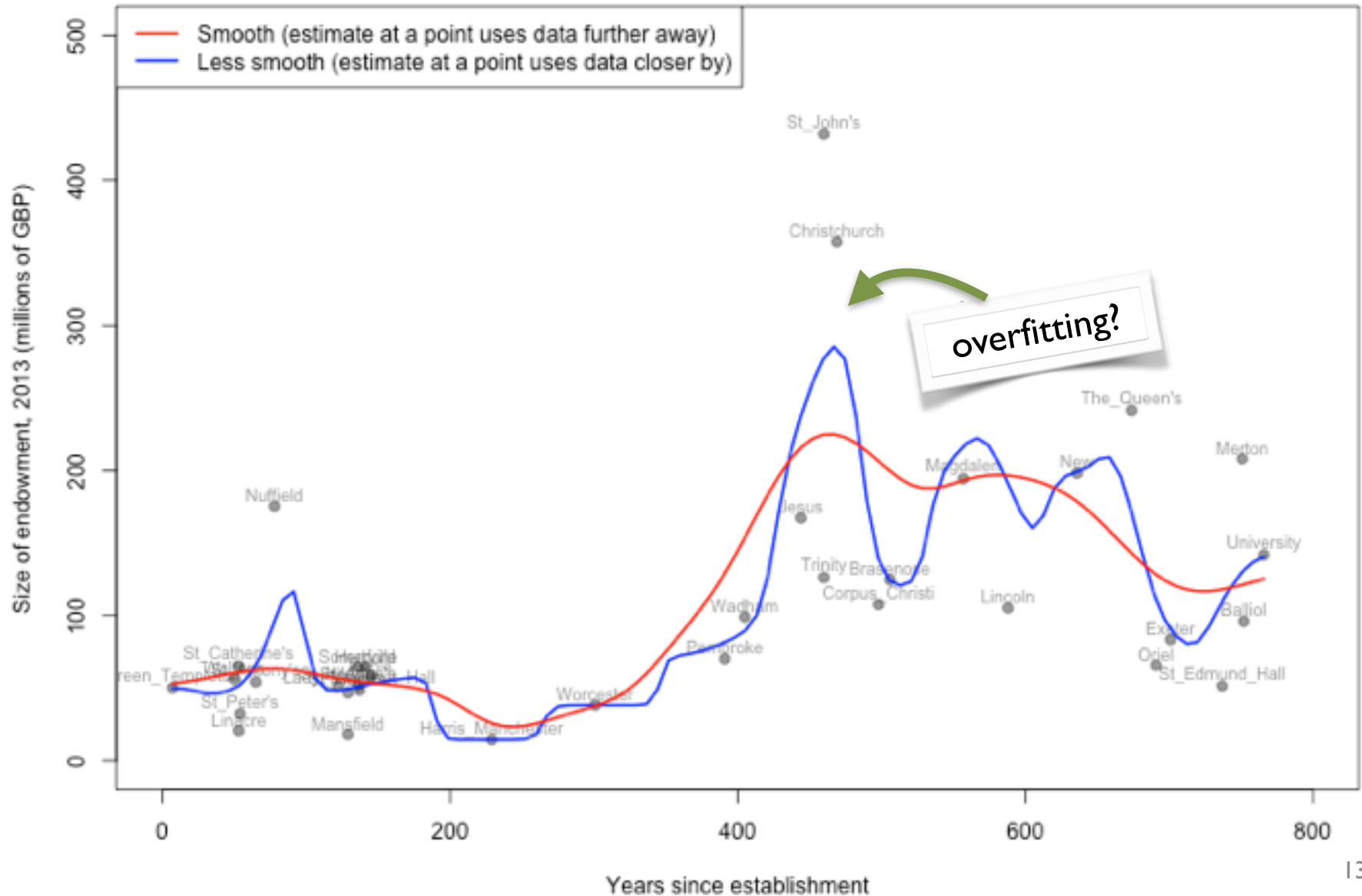
Mean endowment within 50-year intervals



Mean age within intervals of endowment



Kernel smoother: estimate at a point is (weighted) average of nearby points



Covariance, correlation, and regression

Covariance: a measure of linear association

How do x and y tend to move together, i.e. how do they **covary**?

When x is above its mean, is y also above its mean? By how much?

Covariance: a measure of linear association

How do x and y tend to move together, i.e. how do they **covary**?

When x is above its mean, is y also above its mean? By how much?

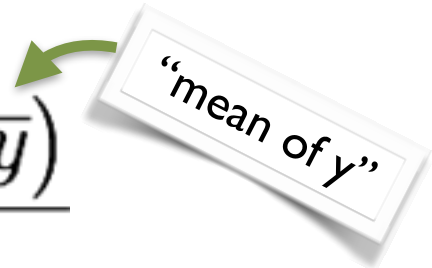
$$\text{Cov}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Covariance: a measure of linear association

How do x and y tend to move together, i.e. how do they **covary**?

When x is above its mean, is y also above its mean? By how much?

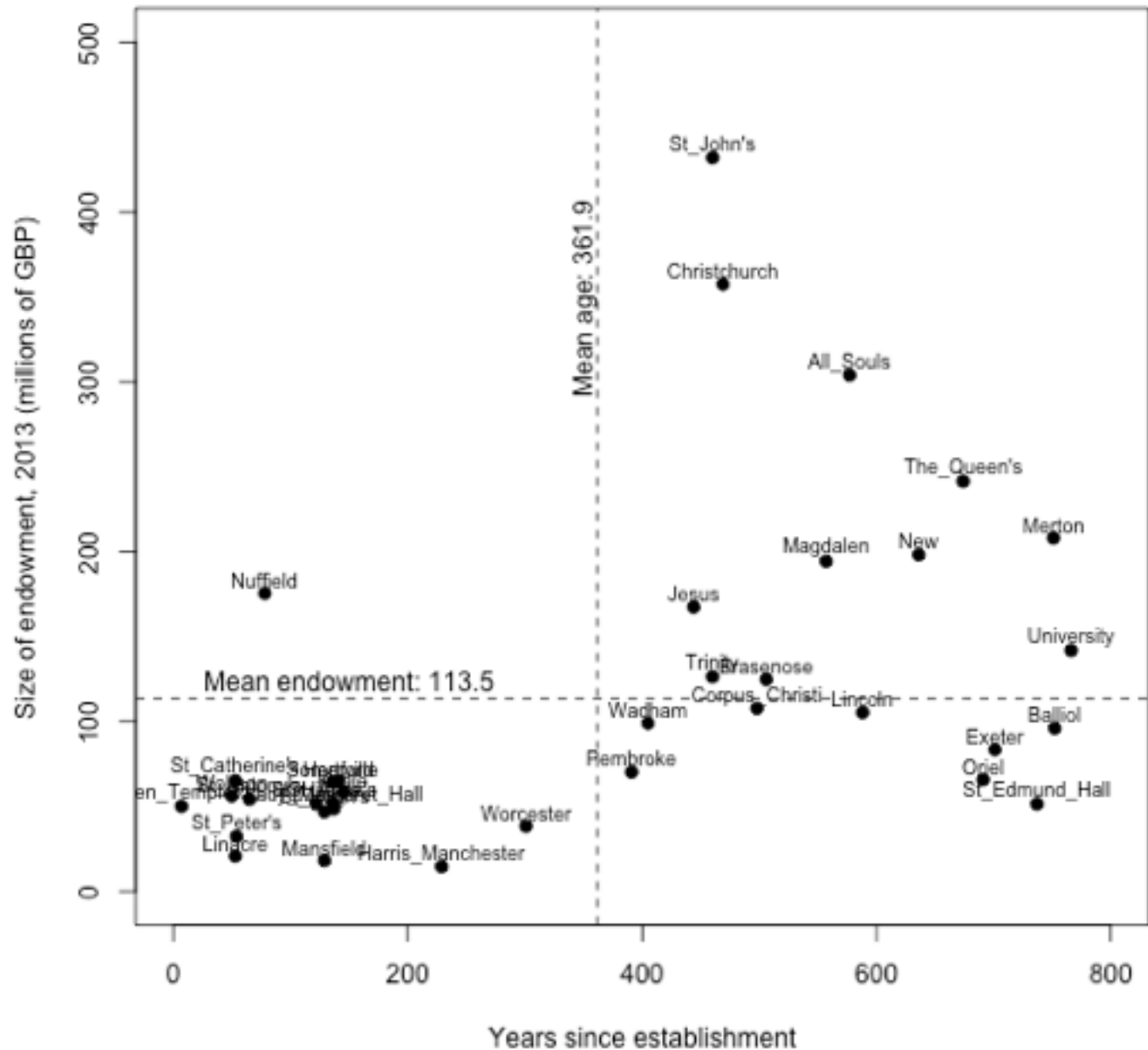
$$\text{Cov}(x, y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$



“mean of y ”

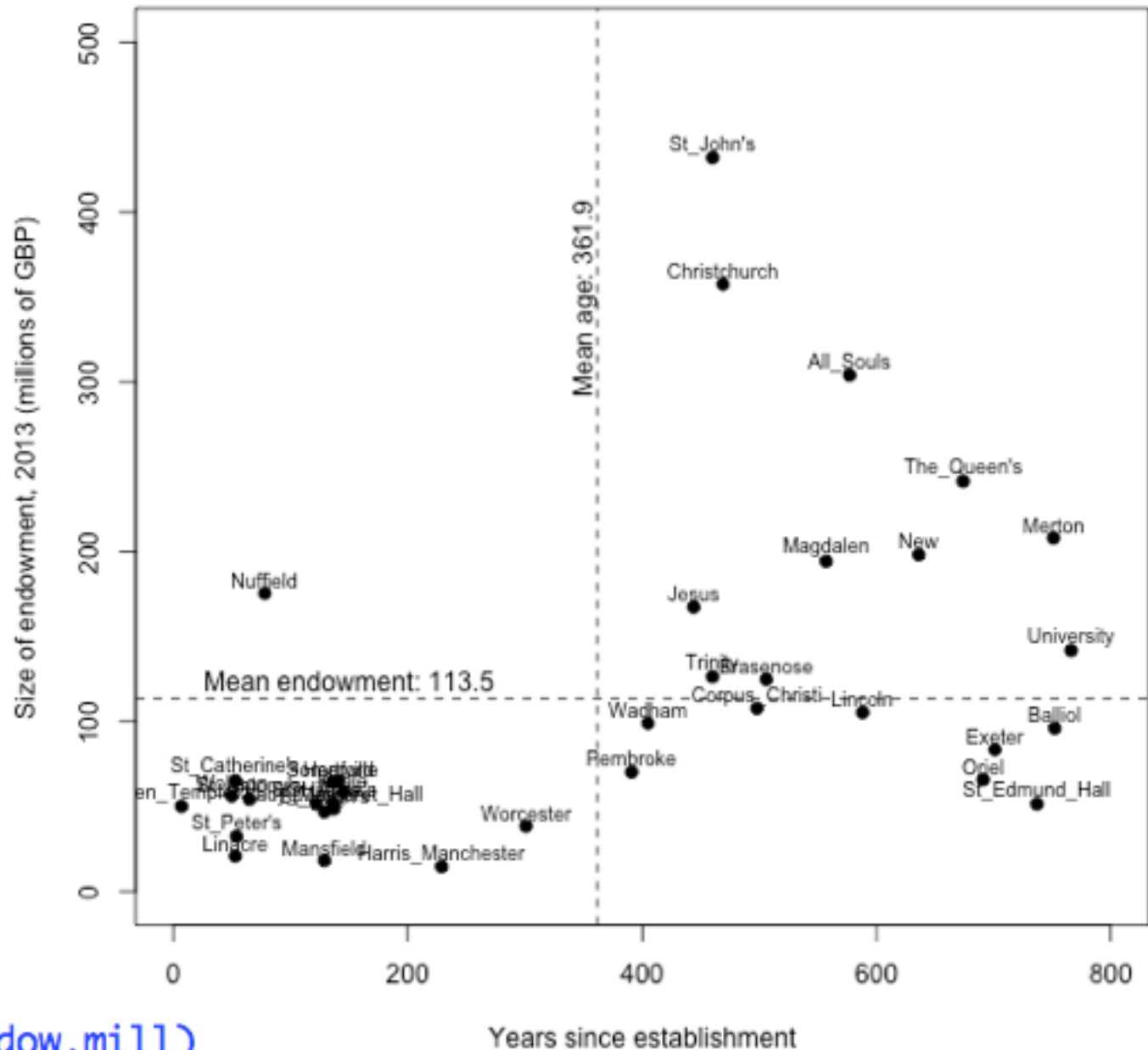
Is the covariance of endowment and age positive or negative?

When x is above its mean, is y also above its mean?



Is the covariance of endowment and age positive or negative?

When x is above its mean, is y also above its mean?



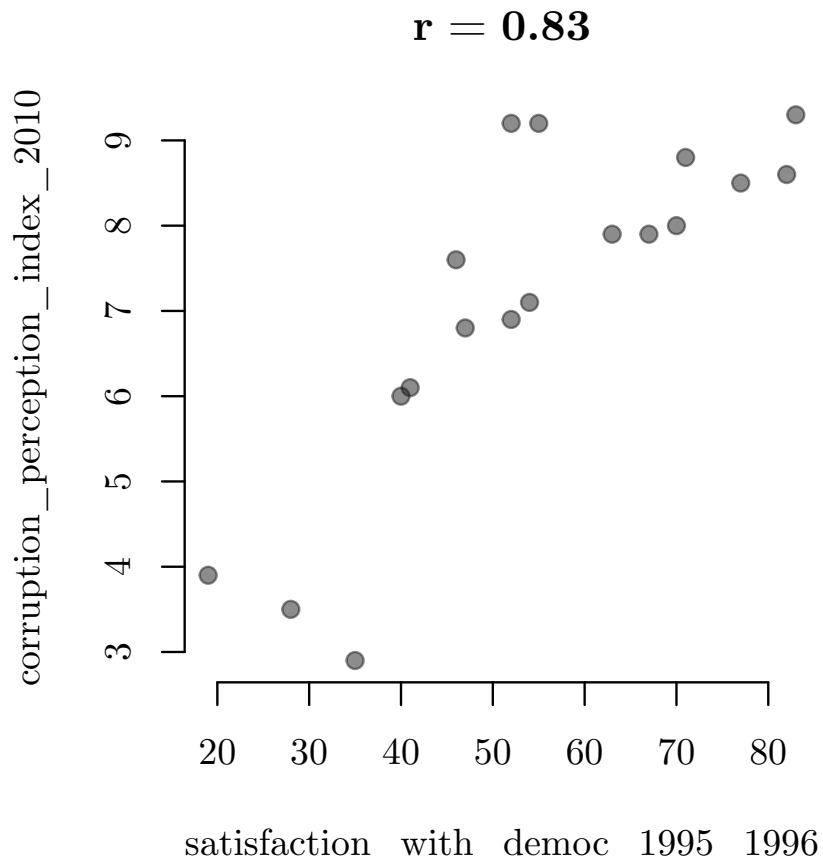
```
> cov(d$age, d$endow.mill)
[1] 11828.5
```

Correlation: a “scale-invariant” measure of linear association

If you plot x and y , how closely are the points arranged on a line (and is the slope of that line positive or negative)?

Correlation: a “scale-invariant” measure of linear association

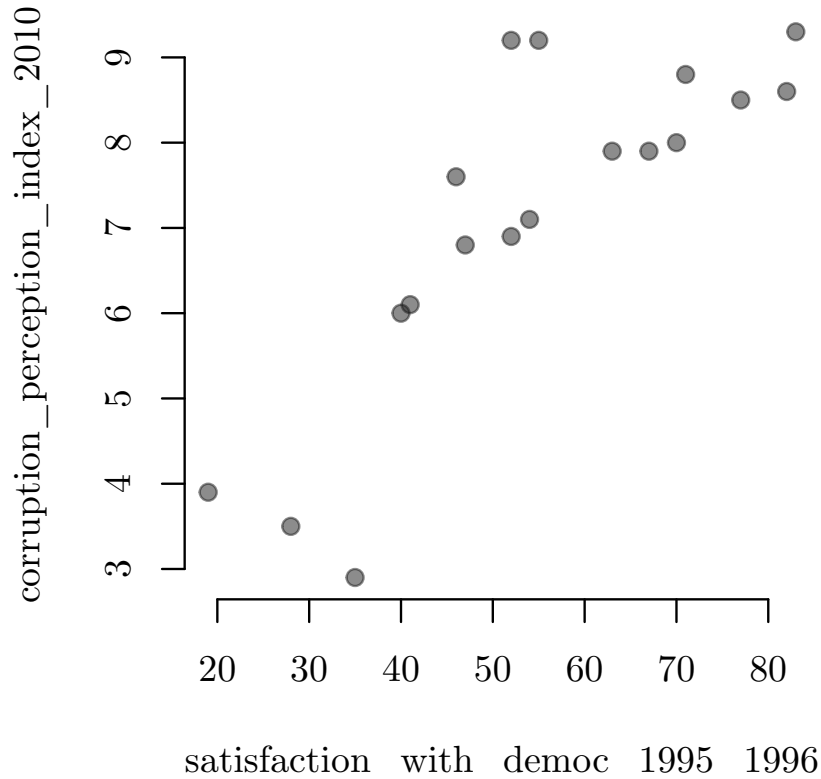
If you plot x and y , how closely are the points arranged on a line (and is the slope of that line positive or negative)?



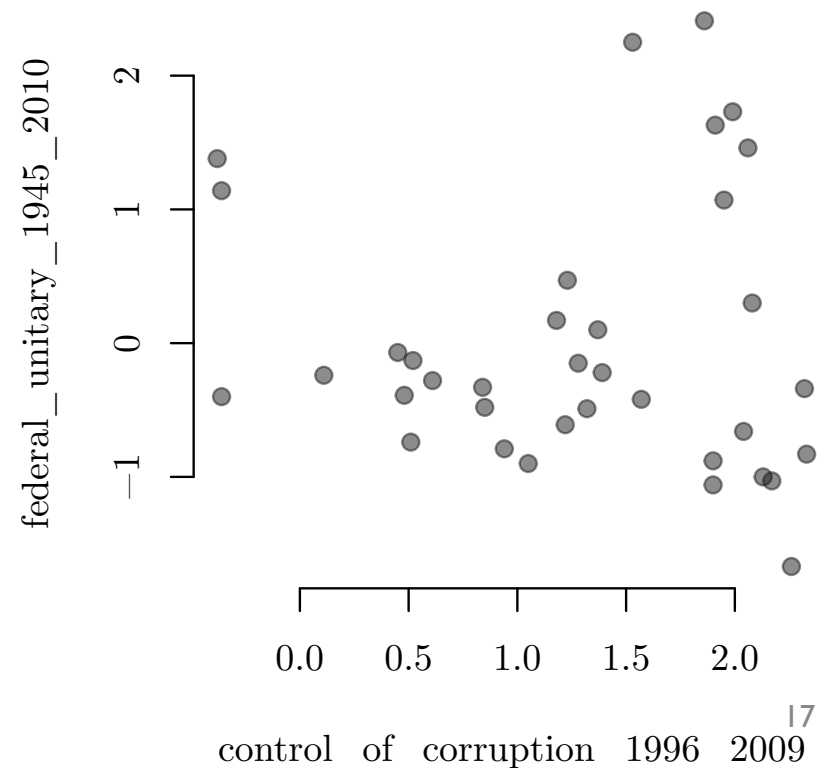
Correlation: a “scale-invariant” measure of linear association

If you plot x and y, how closely are the points arranged on a line (and is the slope of that line positive or negative)?

$r = 0.83$

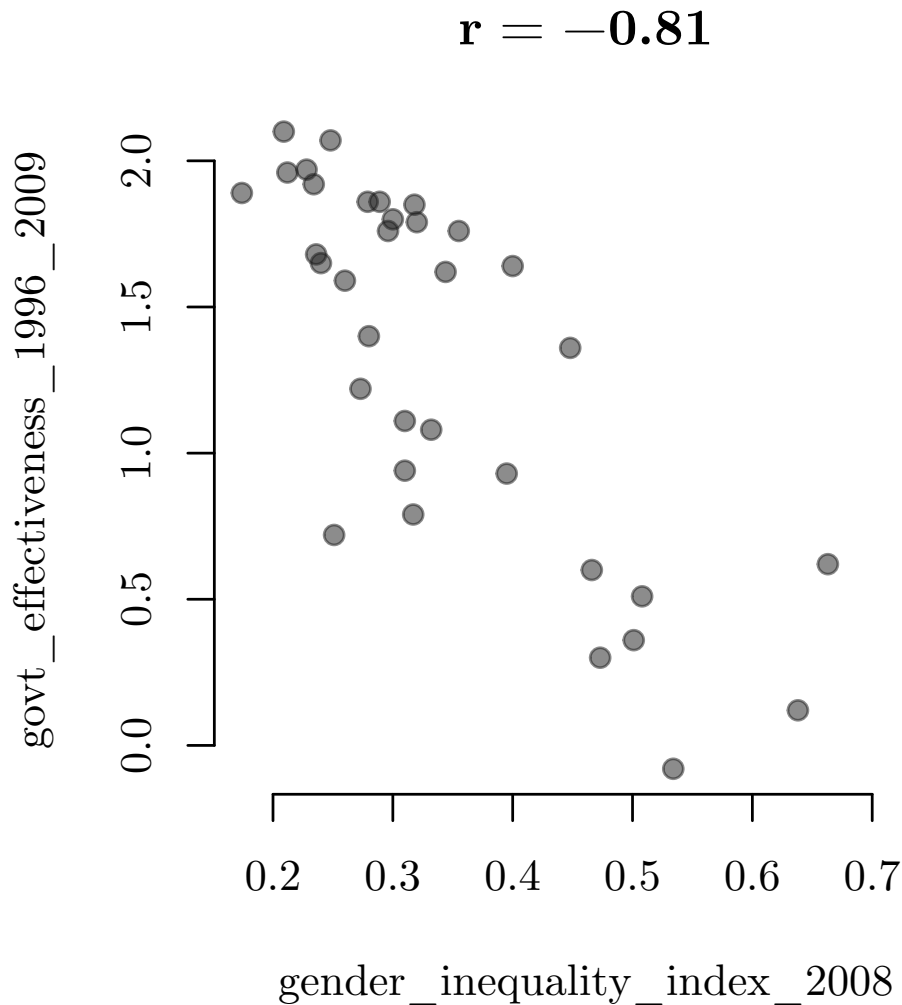


$r = -0.05$



Correlation: a scale-invariant measure of linear association

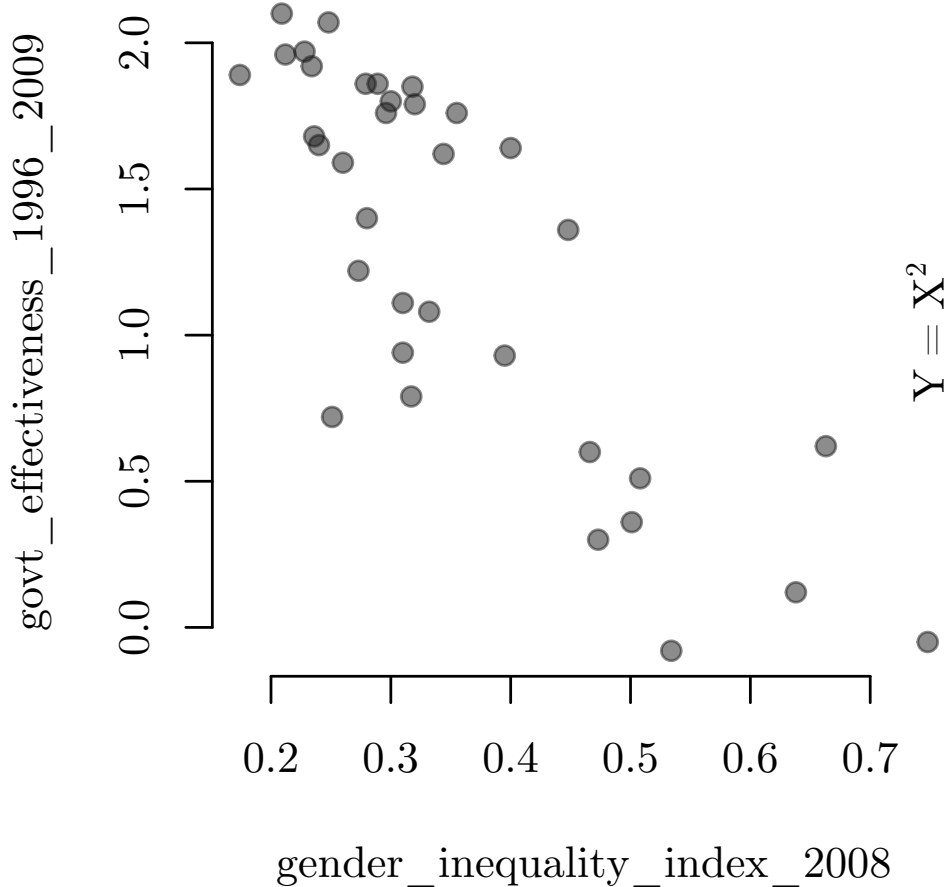
If you plot x and y, how closely are the points arranged on a line (and is the slope of that line positive or negative)?



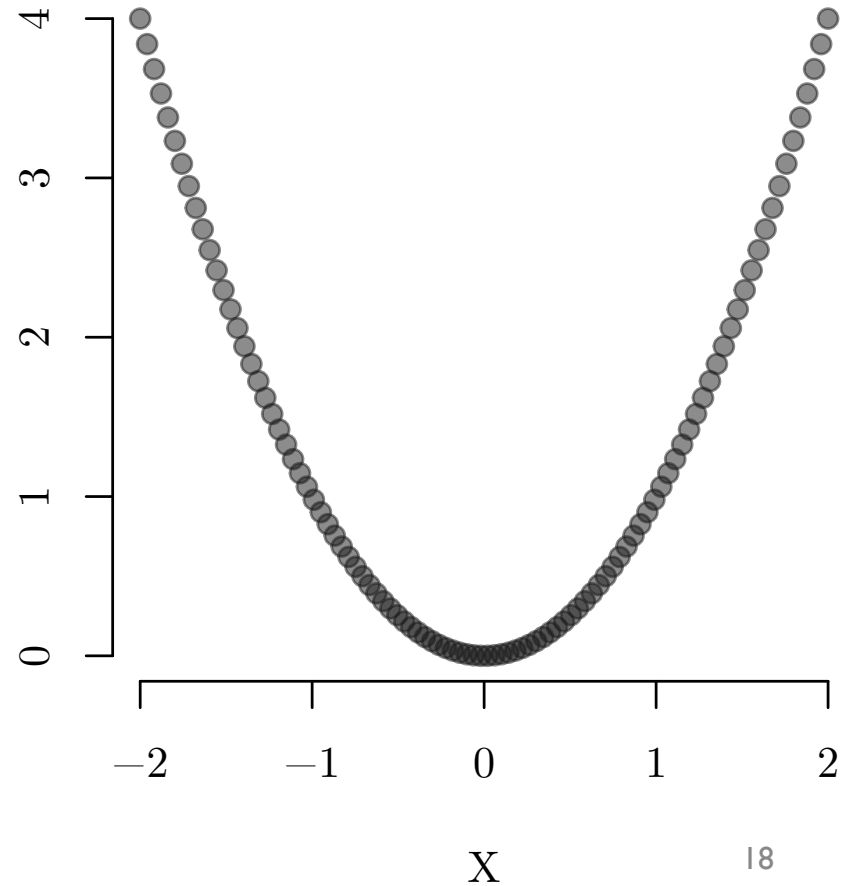
Correlation: a scale-invariant measure of linear association

If you plot x and y , how closely are the points arranged on a line (and is the slope of that line positive or negative)?

$$r = -0.81$$



$$r = 0$$



Correlation: a scale-invariant measure of linear association

Unlike covariance, correlation is **scale-invariant**: e.g. the correlation between x and y is the same as the correlation between $1000x$ and y

Correlation: a scale-invariant measure of linear association

Unlike covariance, correlation is **scale-invariant**: e.g. the correlation between x and y is the same as the correlation between $1000x$ and y

```
> cor(d$gender_inequality_index_2008,  
d$govt_effectiveness_1996_2009, use =  
"complete.obs")  
[1] -0.8116719
```

Correlation: a scale-invariant measure of linear association

Unlike covariance, correlation is **scale-invariant**: e.g. the correlation between x and y is the same as the correlation between $1000x$ and y

```
> cor(d$gender_inequality_index_2008,  
d$govt_effectiveness_1996_2009, use =  
"complete.obs")  
[1] -0.8116719
```

```
> cor(d$gender_inequality_index_2008,  
1000*d$govt_effectiveness_1996_2009,  
use = "complete.obs")  
[1] -0.8116719
```

Correlation: a scale-invariant measure of linear association

Unlike covariance, correlation is **scale-invariant**: e.g. the correlation between x and y is the same as the correlation between $1000x$ and y

Why? Because correlation is **normalized** by the standard deviations of x and y .

```
> cor(d$gender_inequality_index_2008,  
d$govt_effectiveness_1996_2009, use =  
"complete.obs")  
[1] -0.8116719
```

```
> cor(d$gender_inequality_index_2008,  
1000*d$govt_effectiveness_1996_2009,  
use = "complete.obs")  
[1] -0.8116719
```

Correlation: a scale-invariant measure of linear association

Unlike covariance, correlation is **scale-invariant**: e.g. the correlation between x and y is the same as the correlation between $1000x$ and y

```
> cor(d$gender_inequality_index_2008,  
d$govt_effectiveness_1996_2009, use =  
"complete.obs")  
[1] -0.8116719
```

```
> cor(d$gender_inequality_index_2008,  
1000*d$govt_effectiveness_1996_2009,  
use = "complete.obs")  
[1] -0.8116719
```

Why? Because correlation is **normalized** by the standard deviations of x and y .

$$\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{\text{sd}(x)\text{sd}(y)}$$

Correlation: a scale-invariant measure of linear association


Unlike covariance, correlation is **scale-invariant**: e.g. the correlation between x and y is the same as the correlation between $1000x$ and y

```
> cor(d$gender_inequality_index_2008,  
d$govt_effectiveness_1996_2009, use =  
"complete.obs")  
[1] -0.8116719
```

```
> cor(d$gender_inequality_index_2008,  
1000*d$govt_effectiveness_1996_2009,  
use = "complete.obs")  
[1] -0.8116719
```

Why? Because correlation is **normalized** by the standard deviations of x and y .

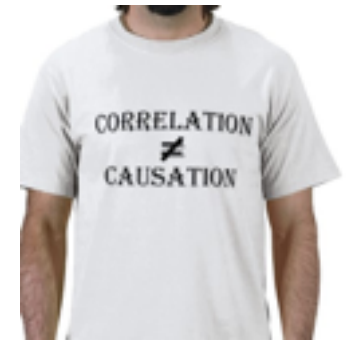
$$\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{\text{sd}(x)\text{sd}(y)}$$



“standard
deviation of y ”

Two things to remember about correlation

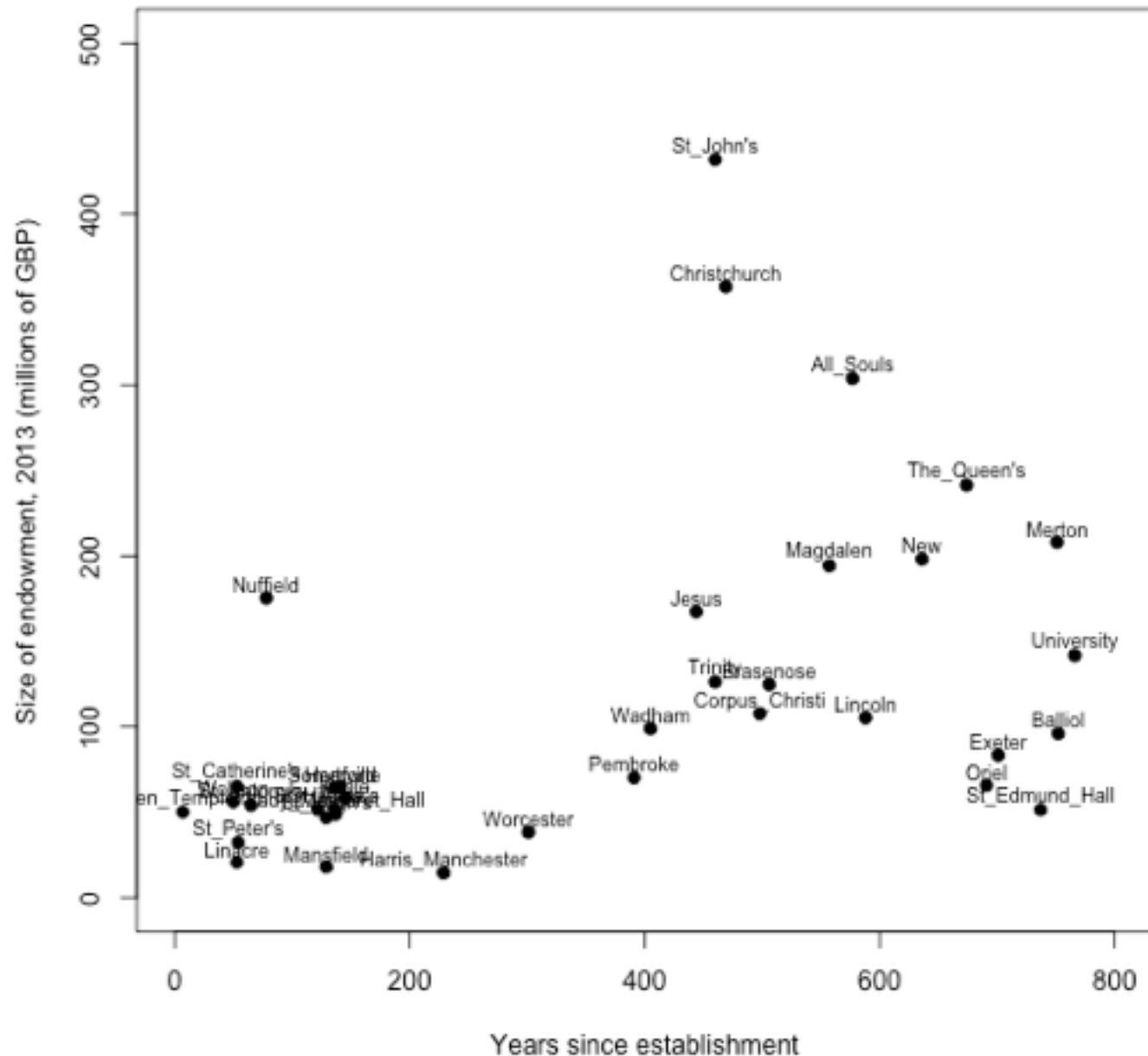
- It is always between -1 and 1.
- Correlation \neq causation.



Regression: linear prediction

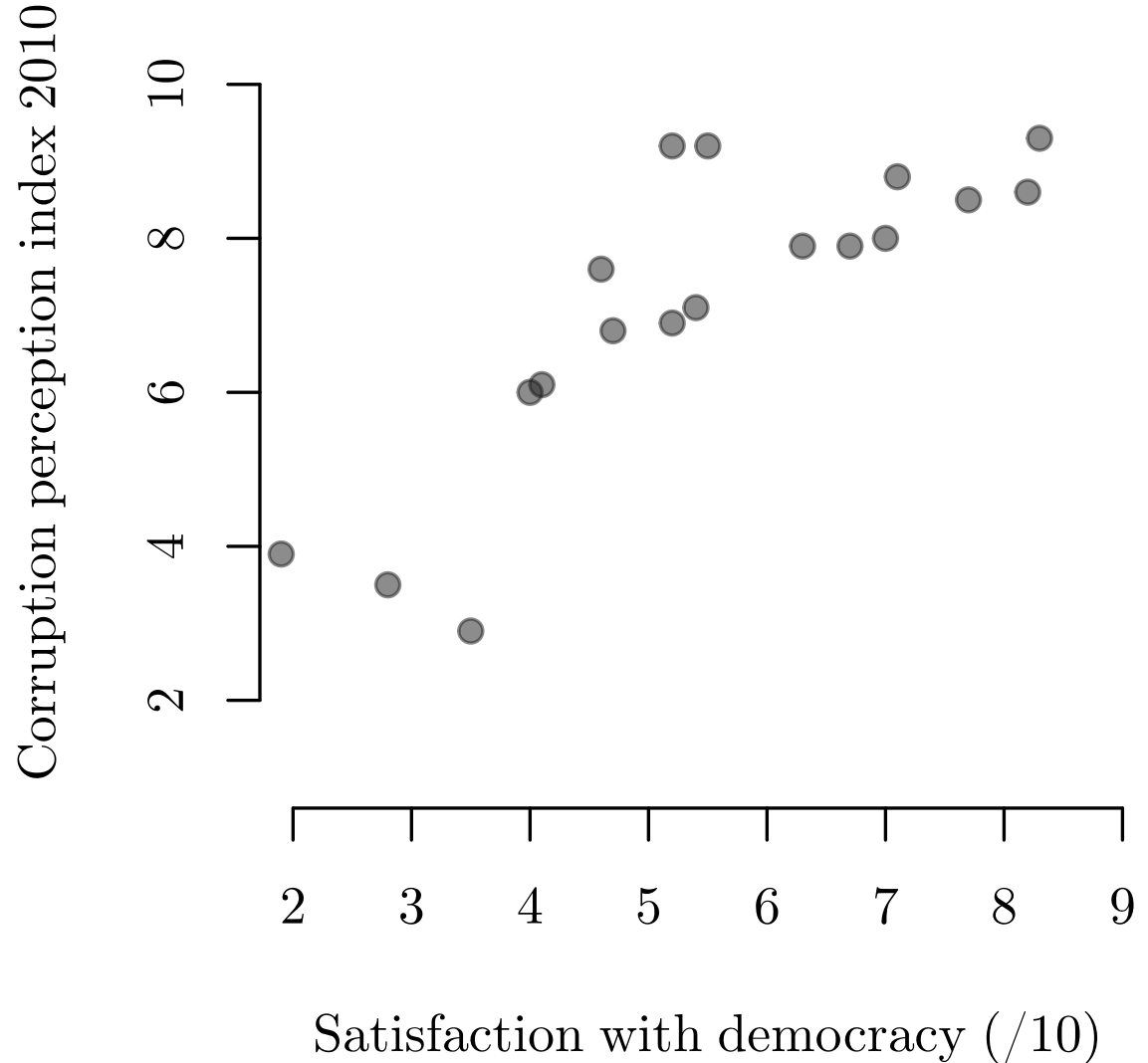
If you knew how old a college was, what would be your **best prediction** for the size of its endowment?

What about your **best linear prediction**?



Regression: linear prediction

If you knew the average satisfaction with democracy in a country in 1995-1996, what would be your **best linear prediction** for its corruption perception index in 2010?



A measure of predictive error: residuals

A measure of predictive error: residuals

Residual: The difference between the predicted value and the actual value.

A measure of predictive error: residuals

Residual: The difference between the predicted value and the actual value.

A measure of predictive error: residuals

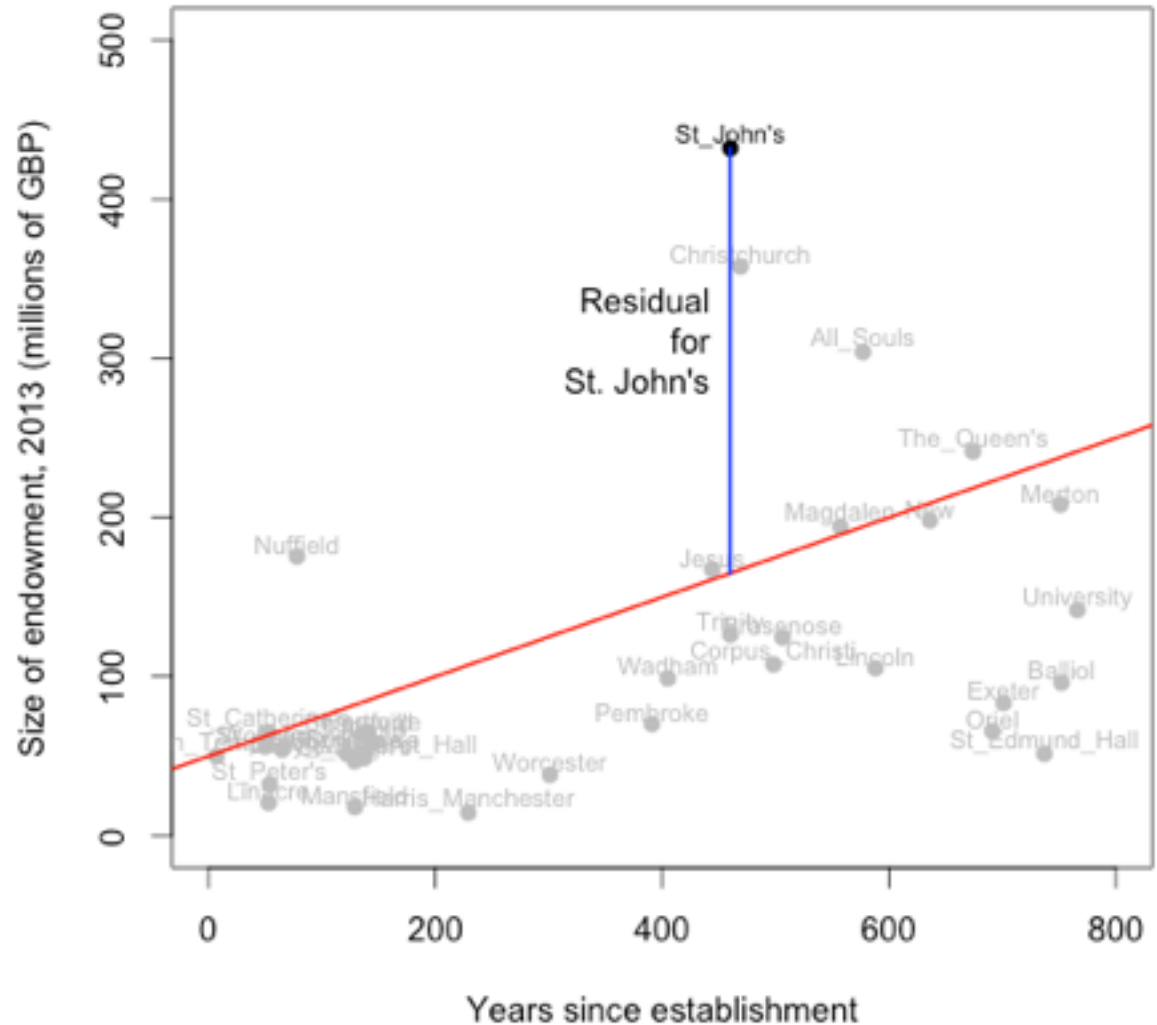
Residual: The difference between the predicted value and the actual value.

Given a linear prediction, the residual is the vertical distance from the point to the line.

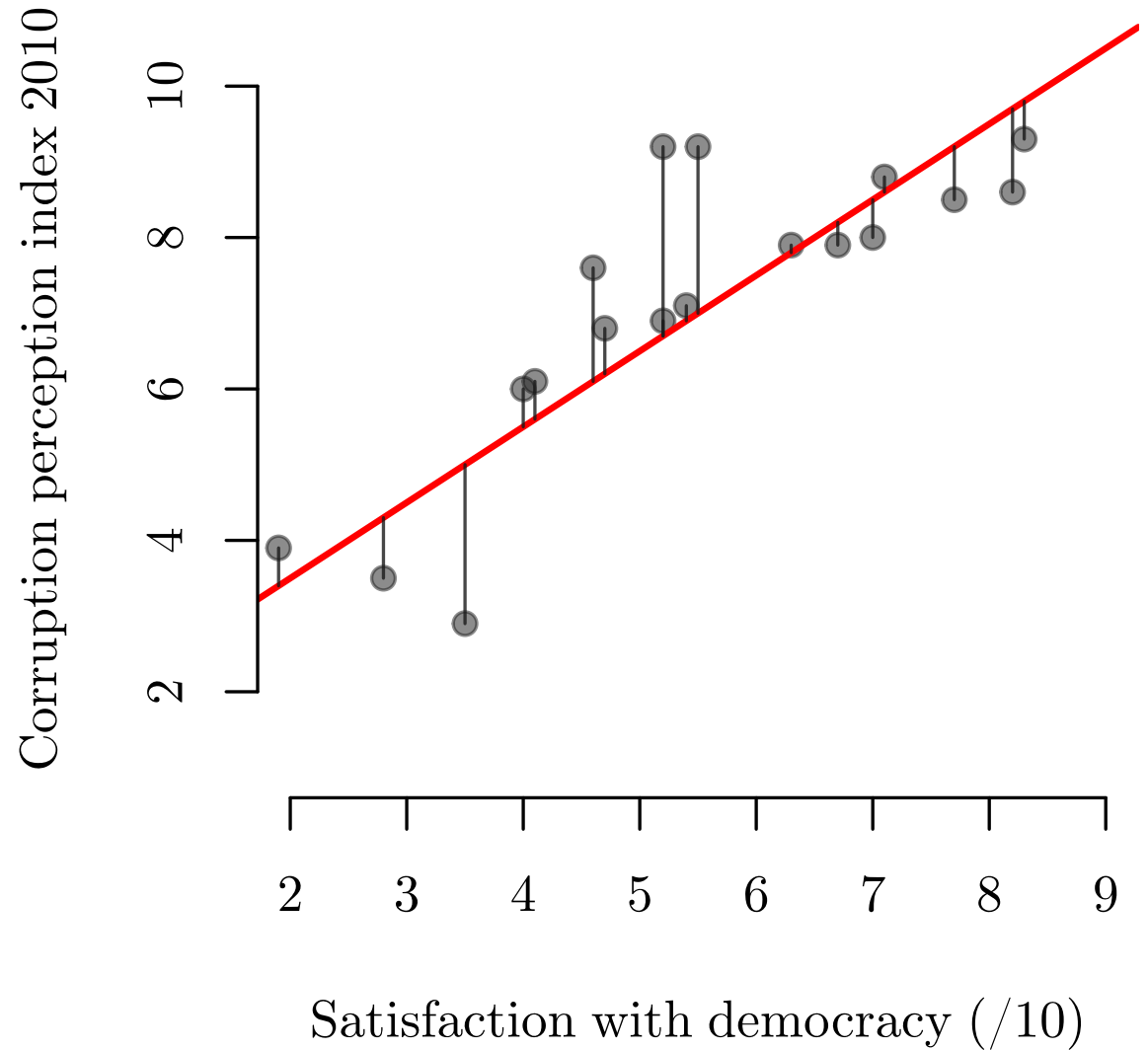
A measure of predictive error: residuals

Residual: The difference between the predicted value and the actual value.

Given a linear prediction, the residual is the vertical distance from the point to the line.

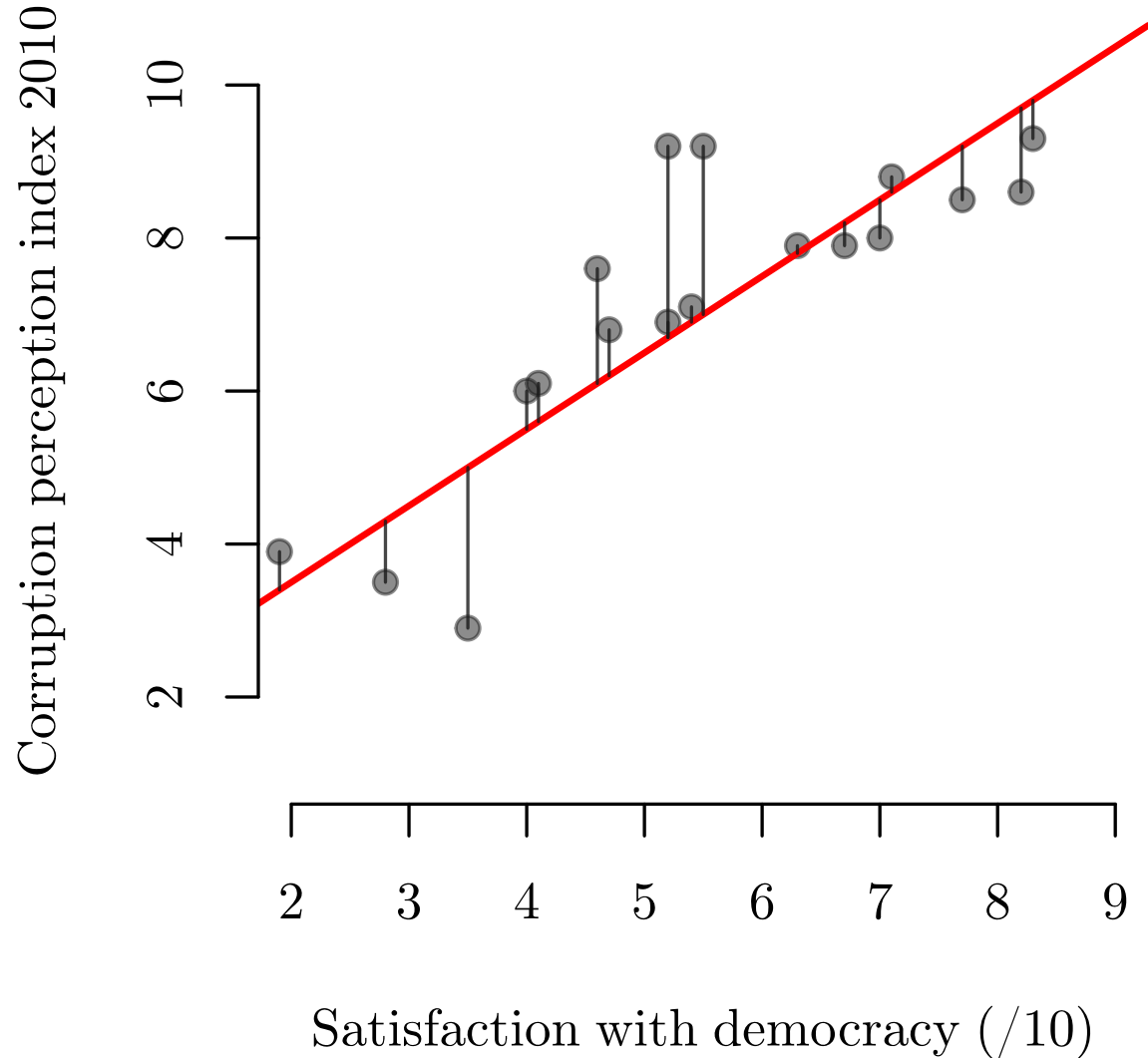


A measure of predictive error: residuals



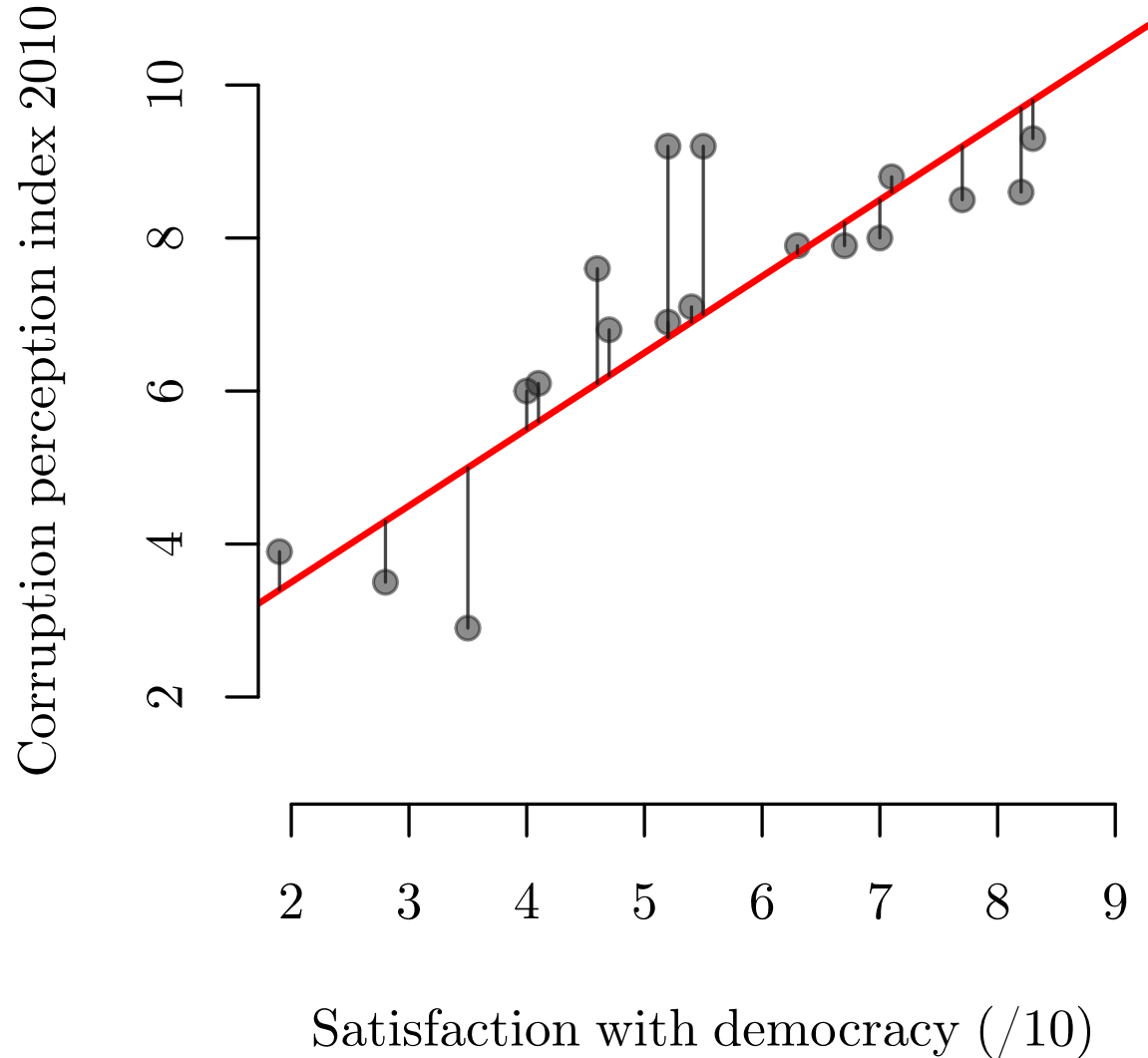
A measure of predictive error: residuals

Residual: The difference between the predicted value and the actual value.



A measure of predictive error: residuals

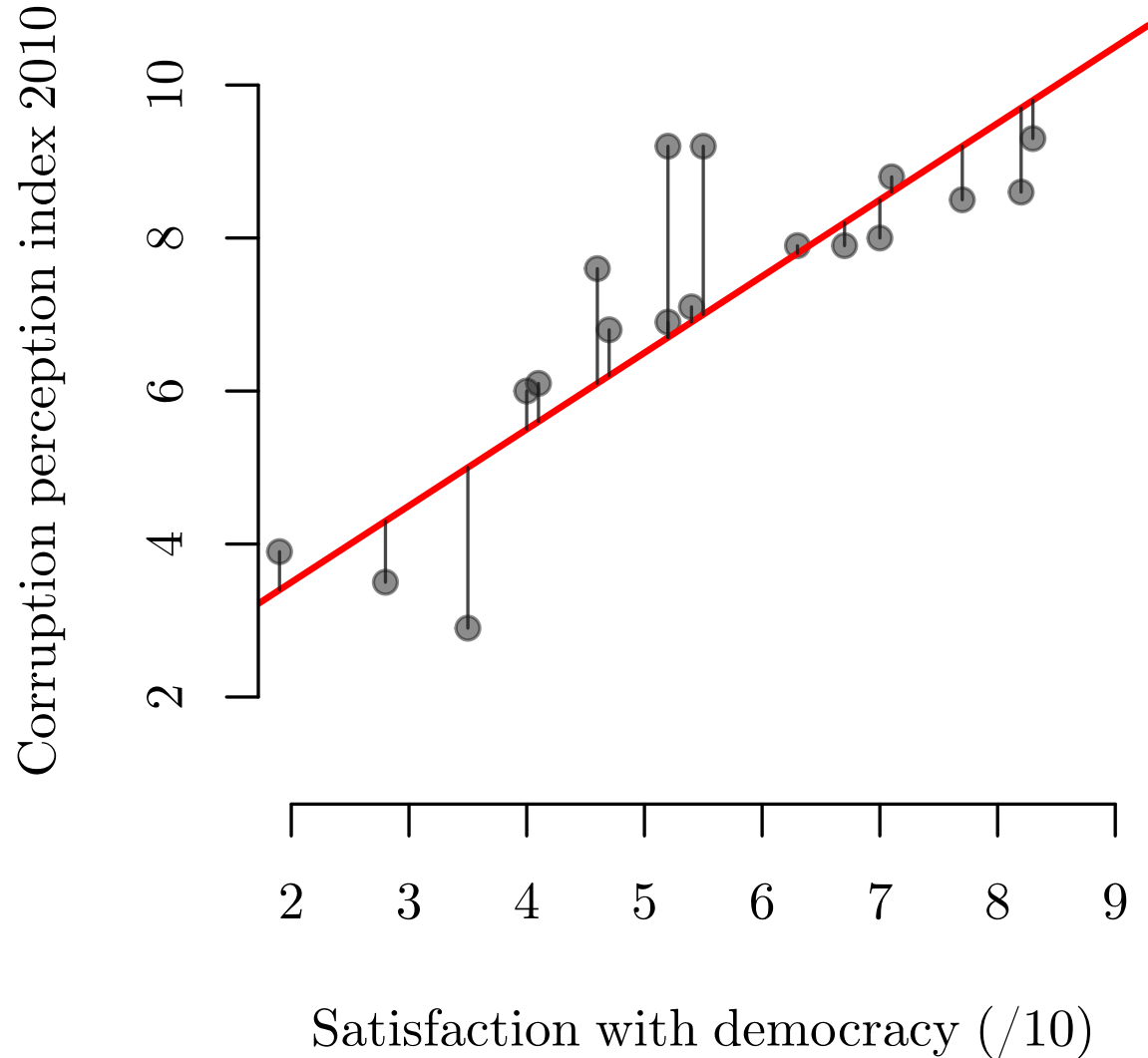
Residual: The difference between the predicted value and the actual value.



A measure of predictive error: residuals

Residual: The difference between the predicted value and the actual value.

Given a linear prediction, the residual is the vertical distance from the point to the line.



Ordinary least squares (OLS) chooses a predictive line by minimizing the sum of squared residuals

Ordinary least squares (OLS) chooses a predictive line by minimizing the sum of squared residuals

For each possible predictive line,

- calculate residuals
- square each residual
- add squared residuals

Choose the line that minimizes that sum.

OLS: Ordinary least squares

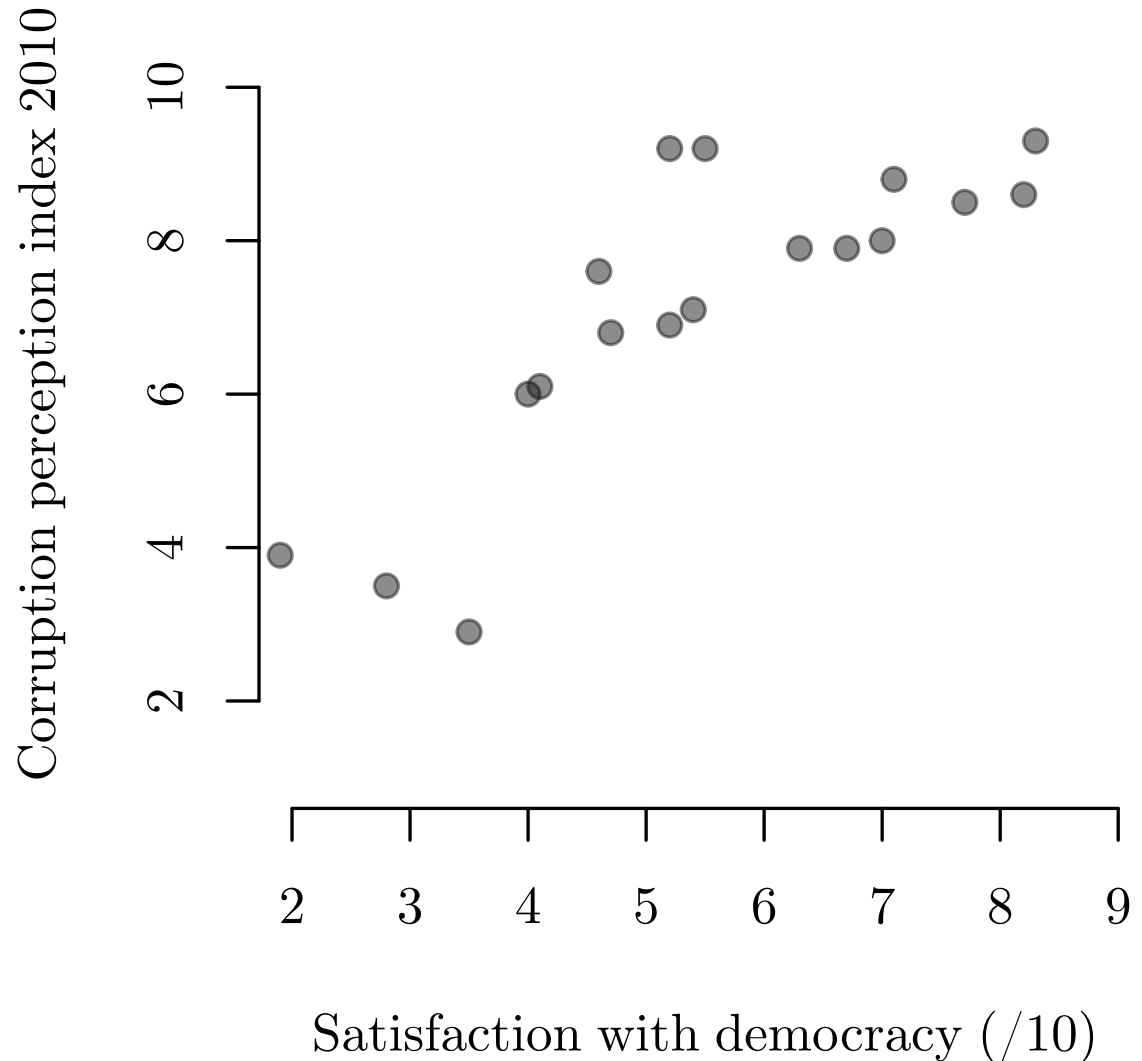
Ordinary least squares (OLS) chooses a predictive line by minimizing the sum of squared residuals

For each possible predictive line,

- calculate residuals
- square each residual
- add squared residuals

Choose the line that minimizes that sum.

OLS: Ordinary least squares



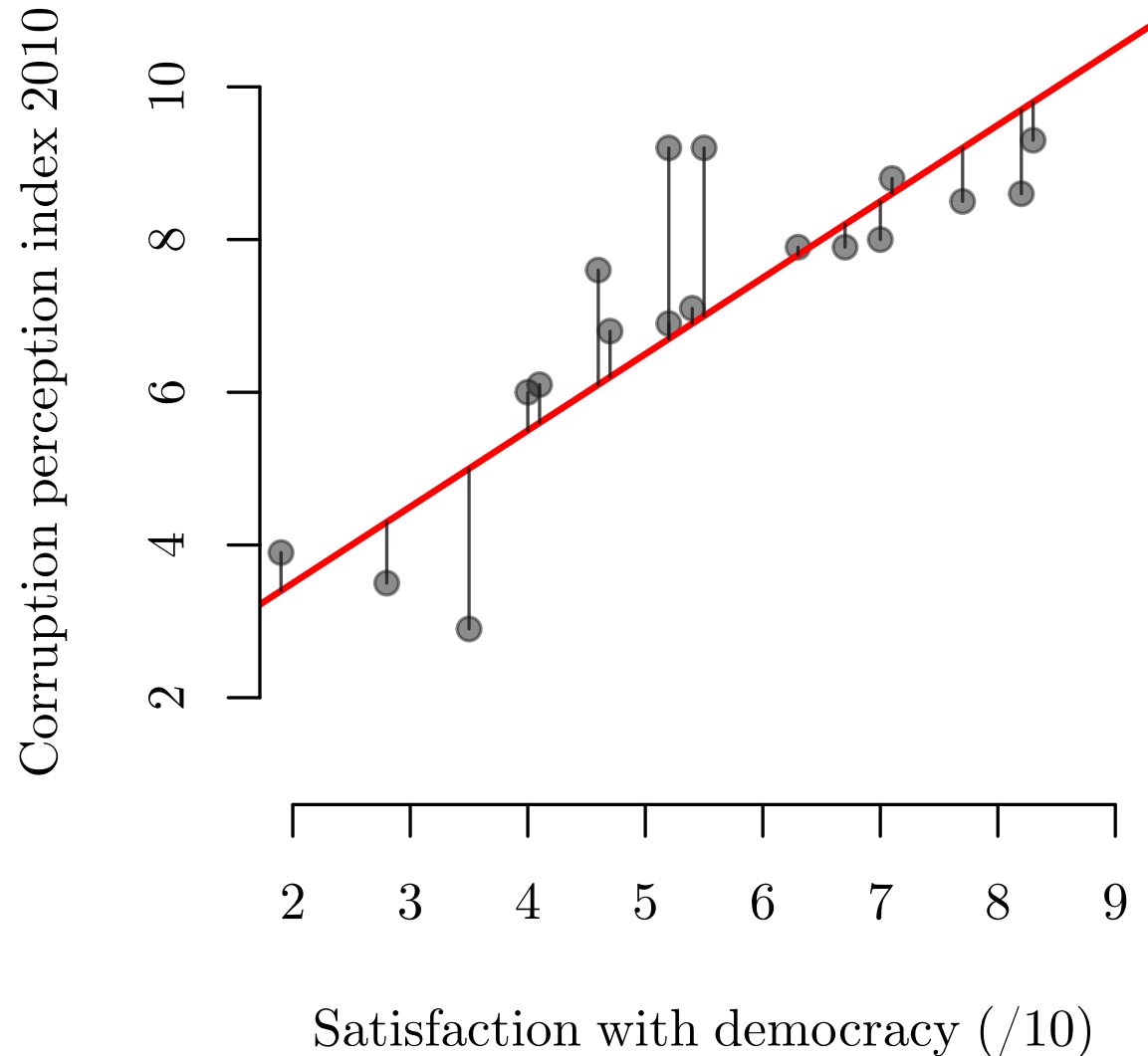
Ordinary least squares (OLS) chooses a predictive line by minimizing the sum of squared residuals

For each possible predictive line,

- calculate residuals
- square each residual
- add squared residuals

Choose the line that minimizes that sum.

OLS: Ordinary least squares



Ordinary least squares (OLS) chooses a predictive line by minimizing the sum of squared residuals

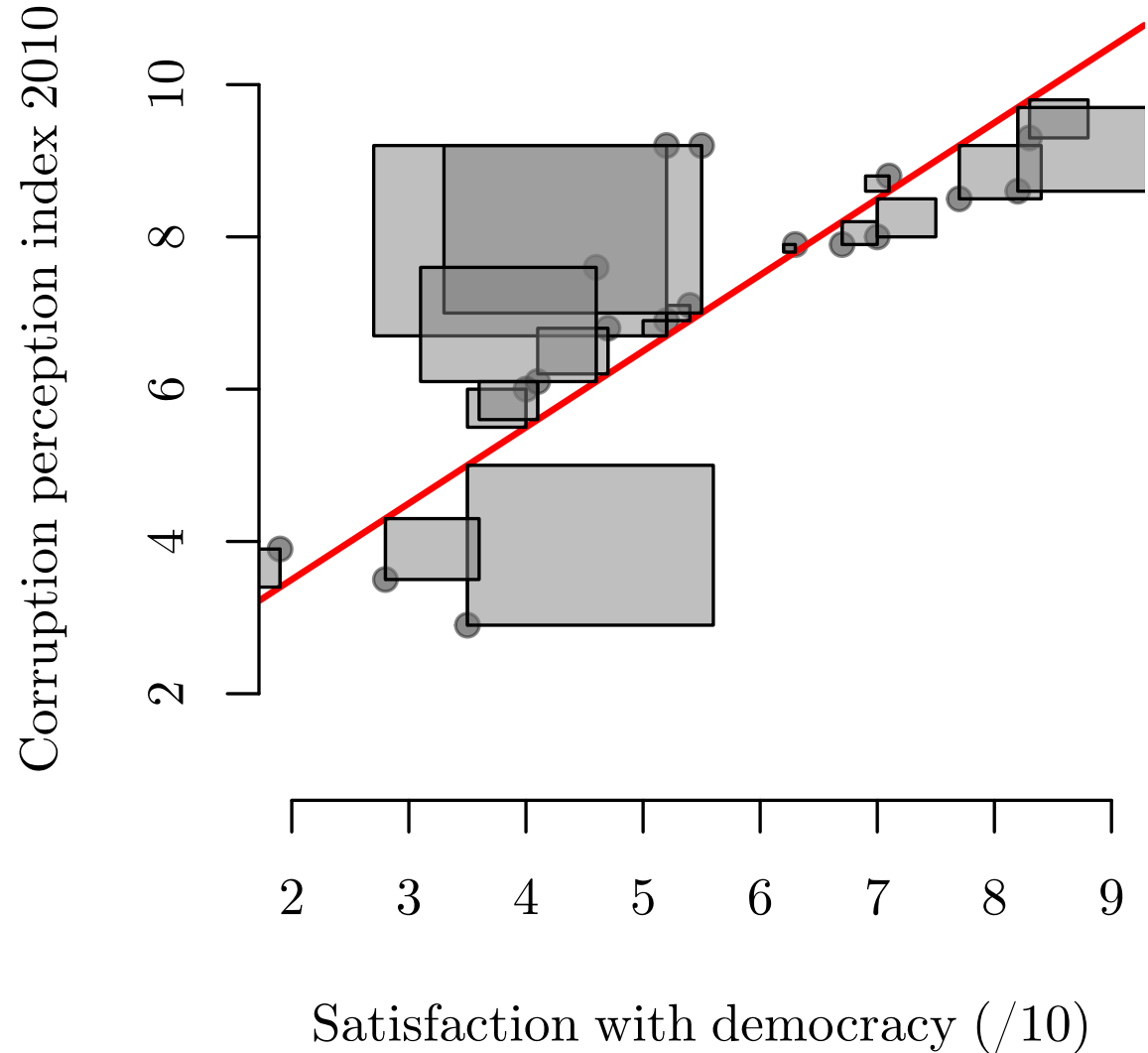
For each possible predictive line,

- calculate residuals
- square each residual
- add squared residuals

Choose the line that minimizes that sum.

OLS: Ordinary least squares

Sum of squared residuals:
21.92



Ordinary least squares (OLS) chooses a predictive line by minimizing the sum of squared residuals

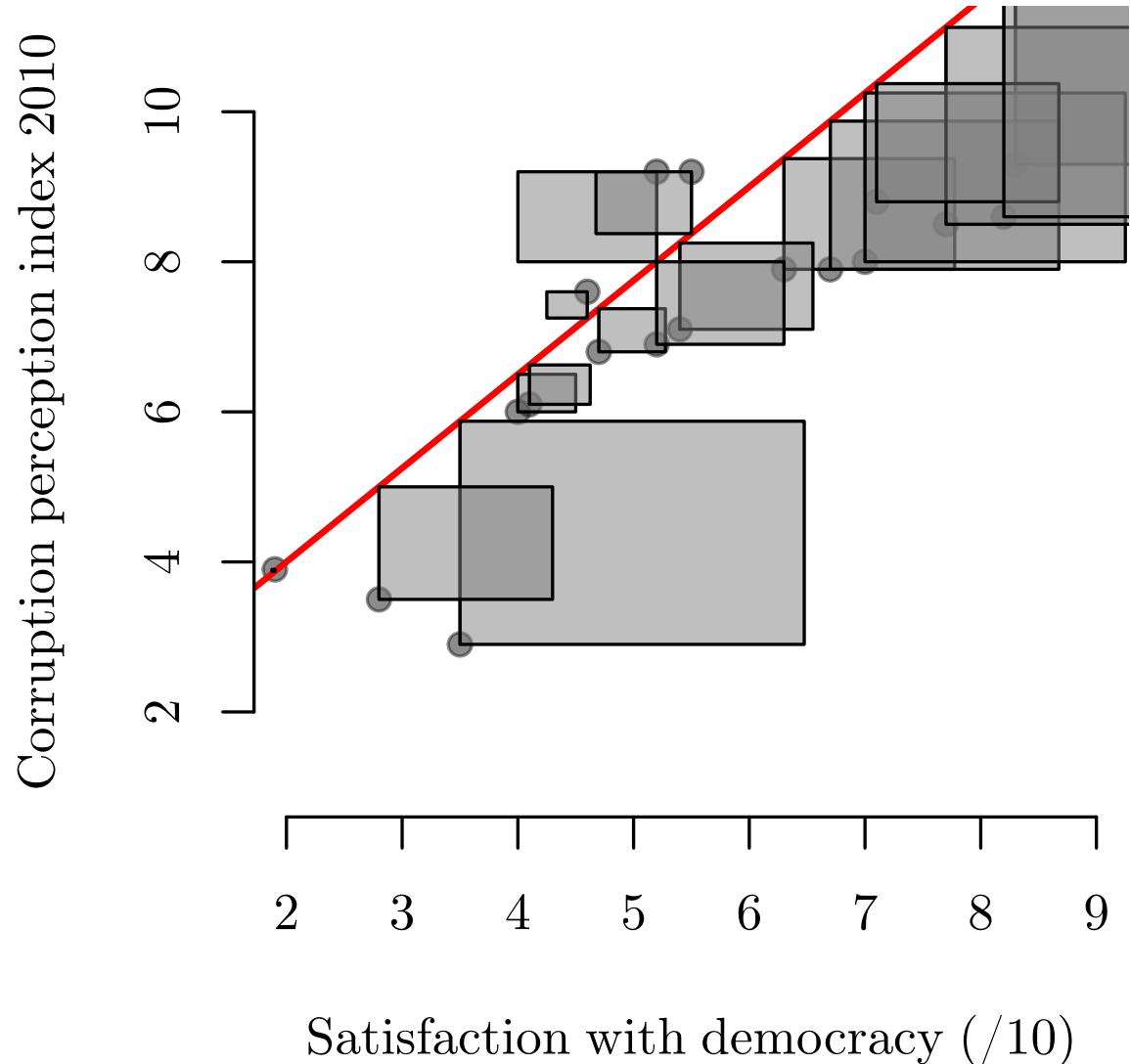
For each possible predictive line,

- calculate residuals
- square each residual
- add squared residuals

Choose the line that minimizes that sum.

OLS: Ordinary least squares

Sum of squared residuals:
53.796



Ordinary least squares (OLS) chooses a predictive line by minimizing the sum of squared residuals

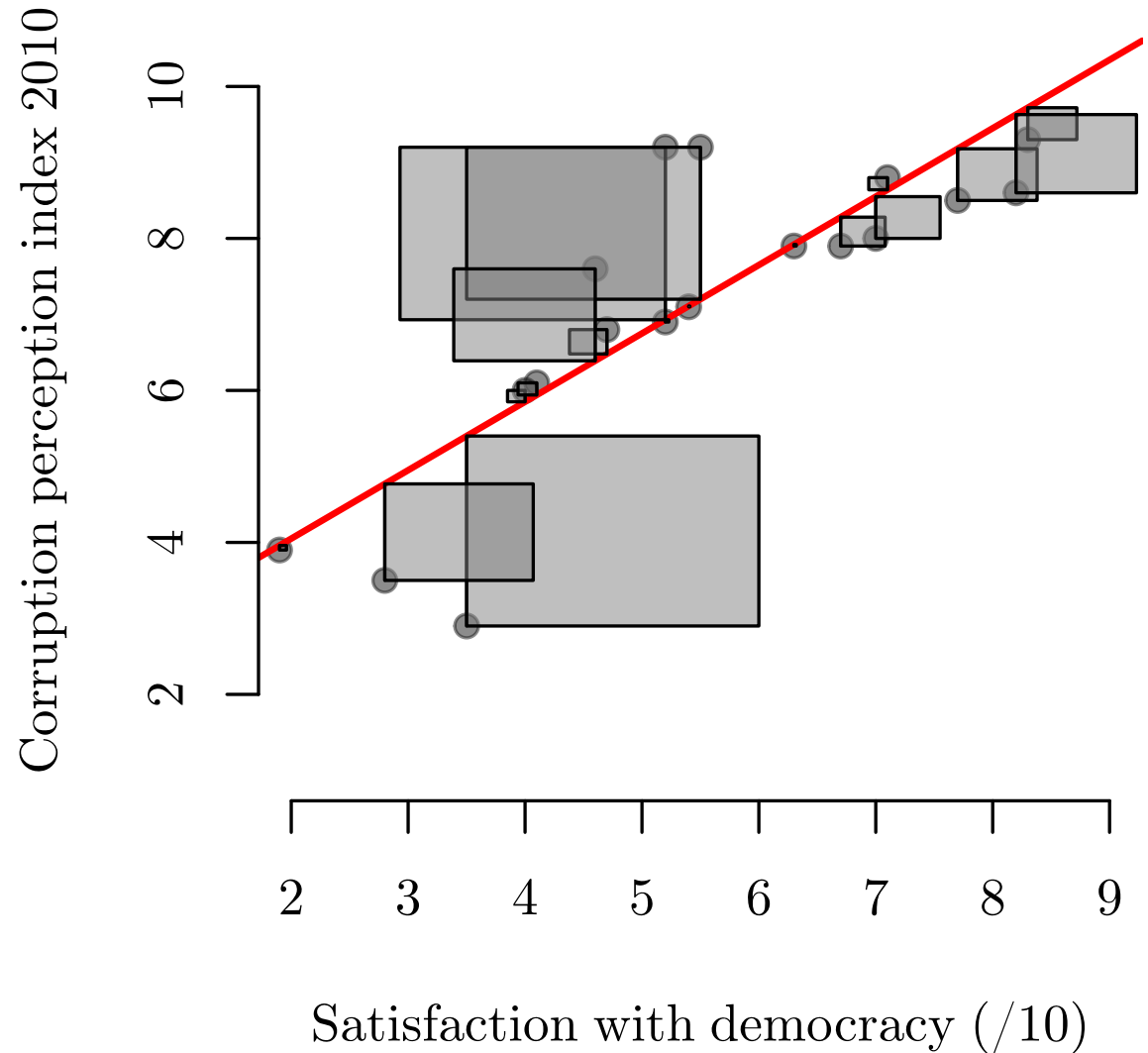
Sum of squared residuals:
20.808

For each possible predictive line,

- calculate residuals
- square each residual
- add squared residuals

Choose the line that minimizes that sum.

OLS: Ordinary least squares



A better way than trial and error!

A better way than trial and error!

OLS is a minimization problem, for which calculus is a great solution!

A better way than trial and error!

OLS is a minimization problem, for which calculus is a great solution!

A better way than trial and error!

OLS is a minimization problem, for which calculus is a great solution!

We let R do the calculus:

A better way than trial and error!

OLS is a minimization problem, for which calculus is a great solution!

We let R do the calculus:

```
> lm(D$corruption_perception_index_2010 ~ D$swd10)
```

Call:

```
lm(formula = D$corruption_perception_index_2010 ~ D$swd10)
```



Coefficients:

(Intercept)	D\$swd10
2.245	0.894

A better way than trial and error!

OLS is a minimization problem, for which calculus is a great solution!

We let R do the calculus:

Dependent variable (Y)  `> lm(D$corruption_perception_index_2010 ~ D$swd10)`  Independent variable (X)

Call:

```
lm(formula = D$corruption_perception_index_2010 ~ D$swd10)
```


Coefficients:

(Intercept)	D\$swd10
2.245	0.894

A better way than trial and error!

OLS is a minimization problem, for which calculus is a great solution!

We let R do the calculus:


Dependent variable (Y)  `> lm(D$corruption_perception_index_2010 ~ D$swd10)`  Independent variable (X)


Call:

```
lm(formula = D$corruption_perception_index_2010 ~ D$swd10)
```

Coefficients:

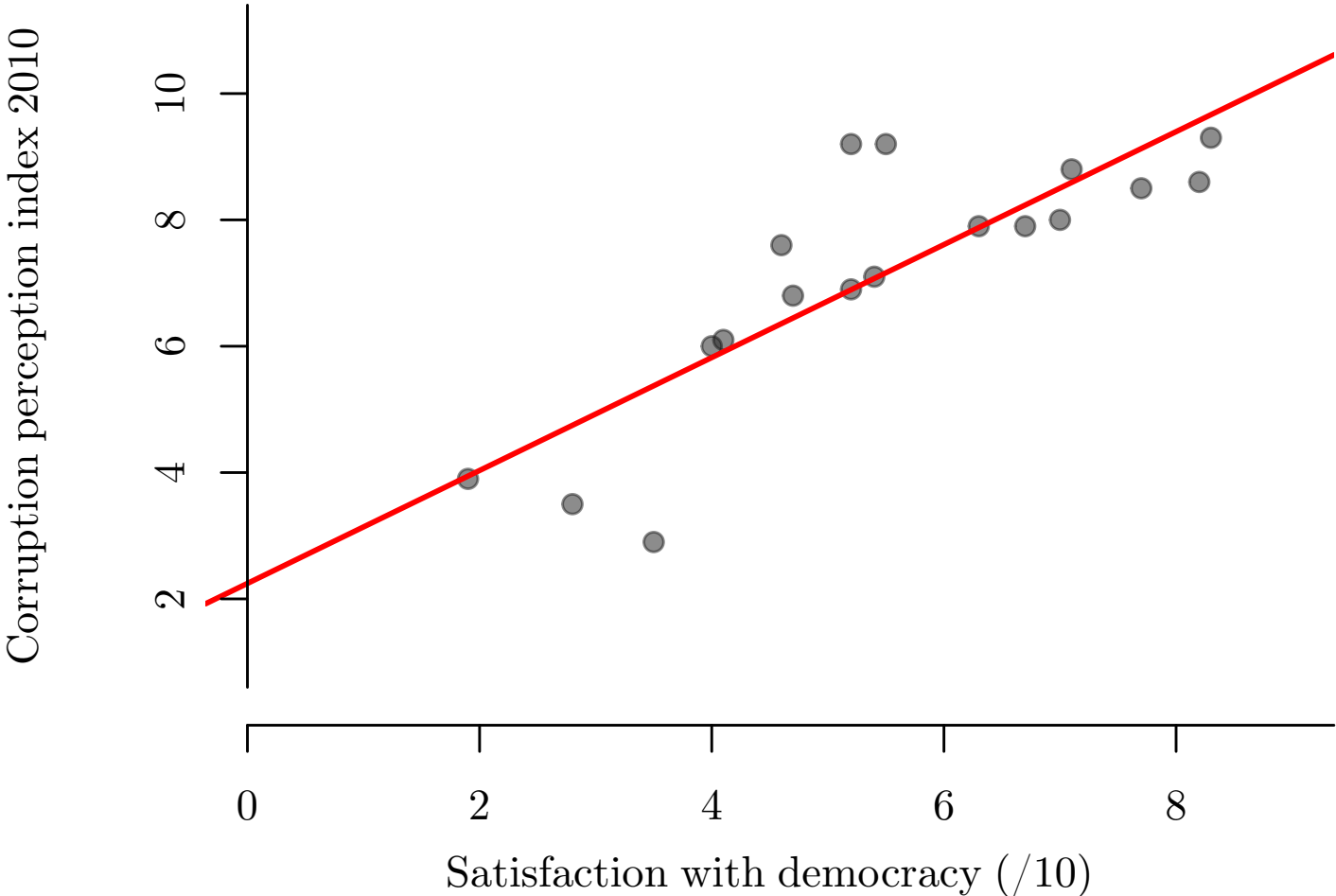
(Intercept)	D\$swd10
2.245	0.894

 Intercept of regression line

 Slope (gradient) of regression line

**Slope (gradient) and intercept:
remember $y = mx + c$**

The OLS regression line

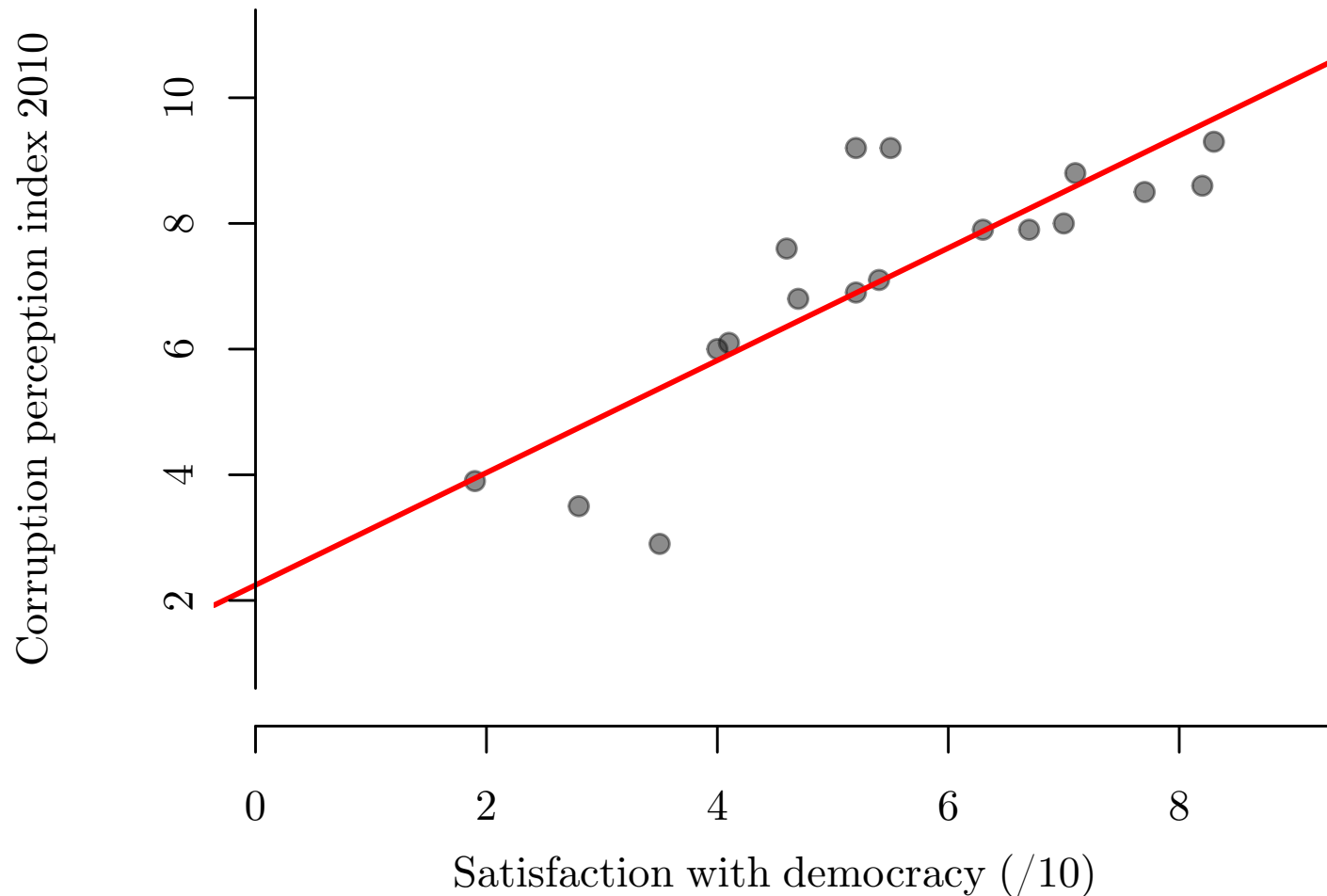


**Slope (gradient) and intercept:
remember $y = mx + c$?**

Coefficients:
(Intercept)
2.245

D\$swd10
0.894

The OLS regression line

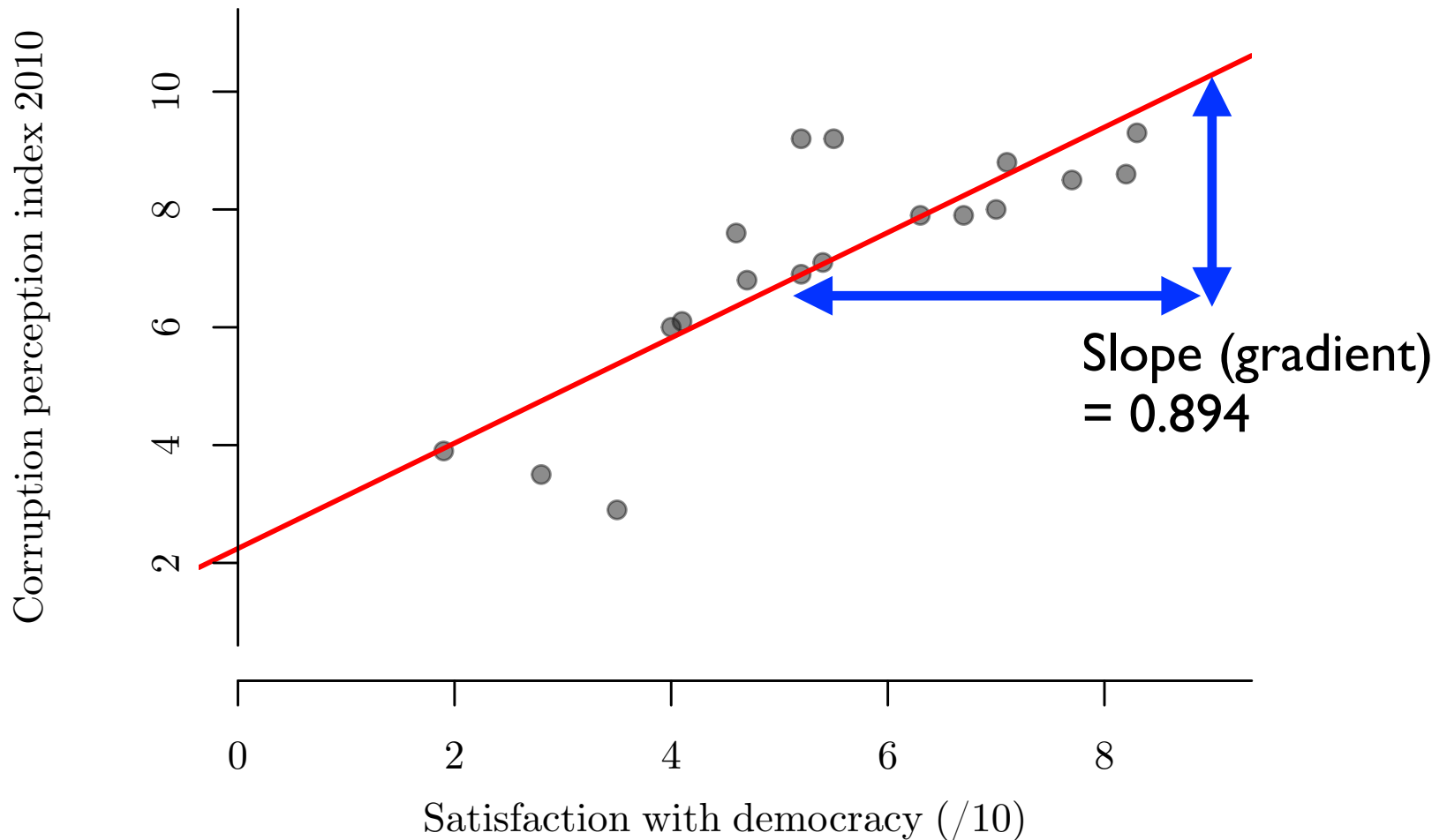


**Slope (gradient) and intercept:
remember $y = mx + c$**

Coefficients:
(Intercept)
2.245

D\$swd10
0.894

The OLS regression line

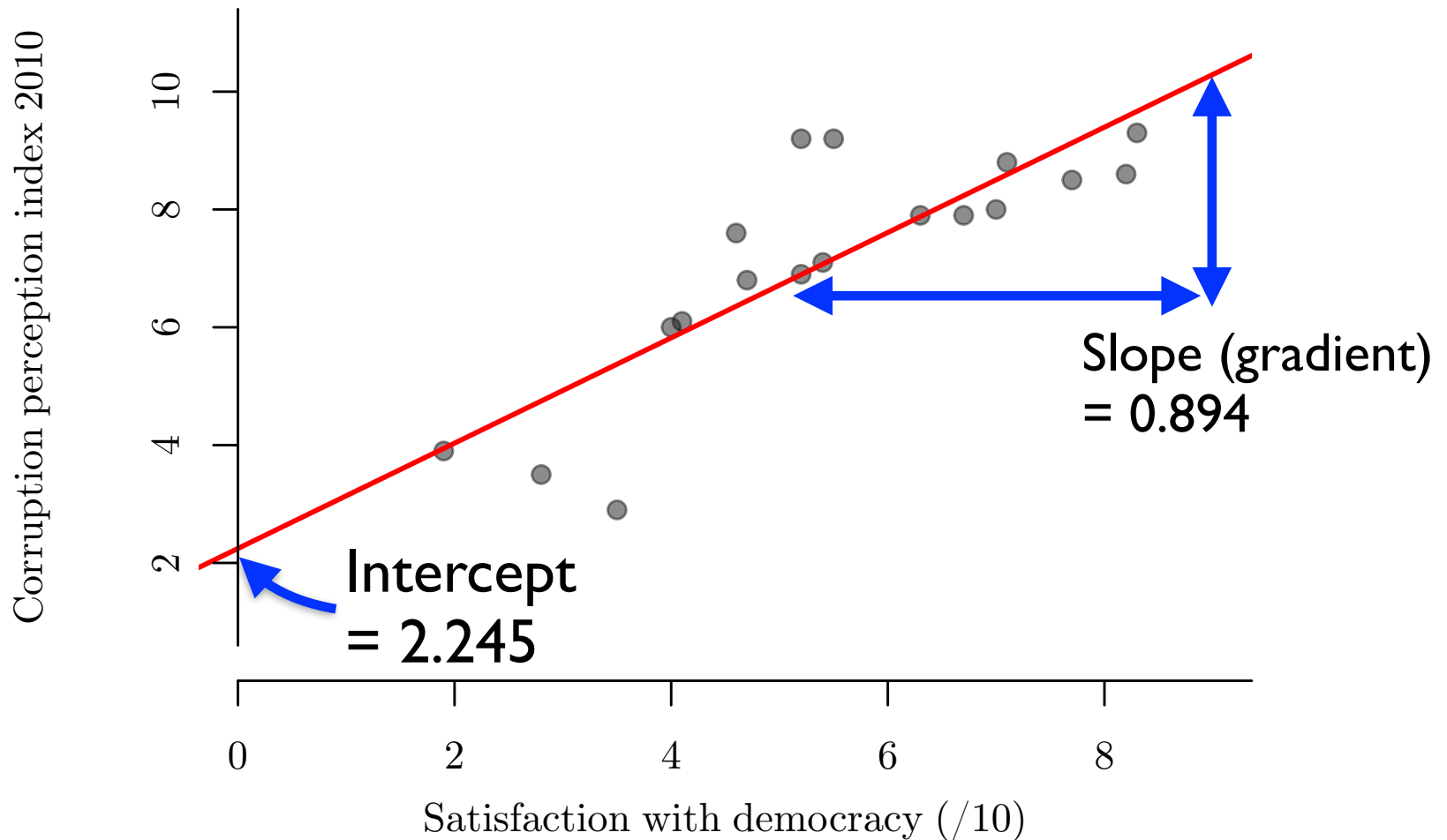


**Slope (gradient) and intercept:
remember $y = mx + c$**

Coefficients:
(Intercept)
2.245

D\$swd10
0.894

The OLS regression line



How well does our regression line predict the outcome? R^2

```
> summary(lm(D$corruption_perception_index_2010 ~ D$swd10))
```

```
Call:
```

```
lm(formula = D$corruption_perception_index_2010 ~ D$swd10)
```

```
Residuals:
```

```
      Min       1Q   Median       3Q      Max
-2.47405 -0.46842  0.01456  0.20319  2.30623
```

```
Coefficients:
```

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   2.2452      0.8667   2.590  0.0197 *
D$swd10        0.8940      0.1510   5.918 2.16e-05 ***
```

```
---
```

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

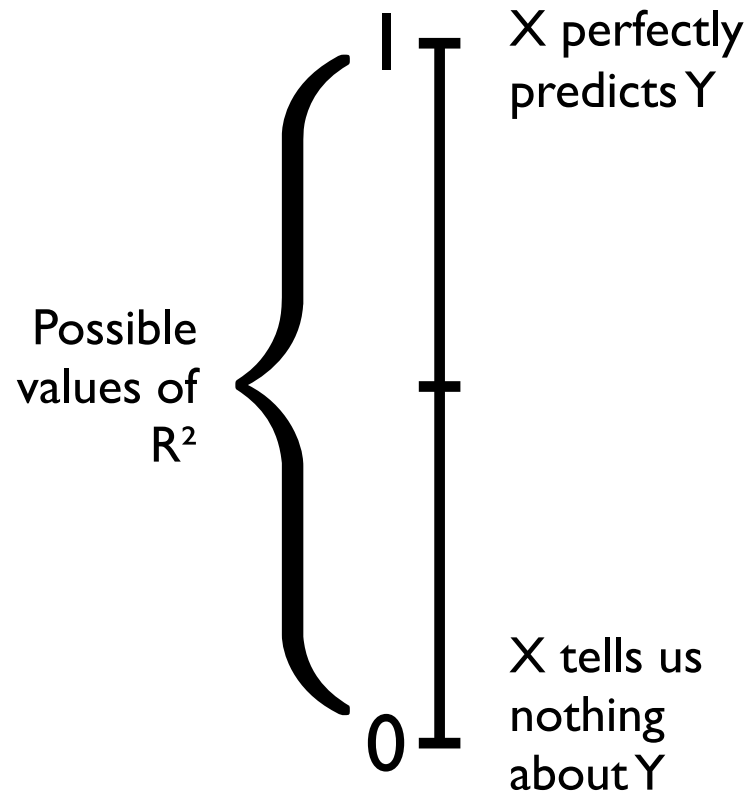
```
Residual standard error: 1.14 on 16 degrees of freedom
(18 observations deleted due to missingness)
```

```
Multiple R-squared: 0.6864, Adjusted R-squared: 0.6668
```

```
F-statistic: 35.03 on 1 and 16 DF, p-value: 2.162e-05
```

R^2 : intuition

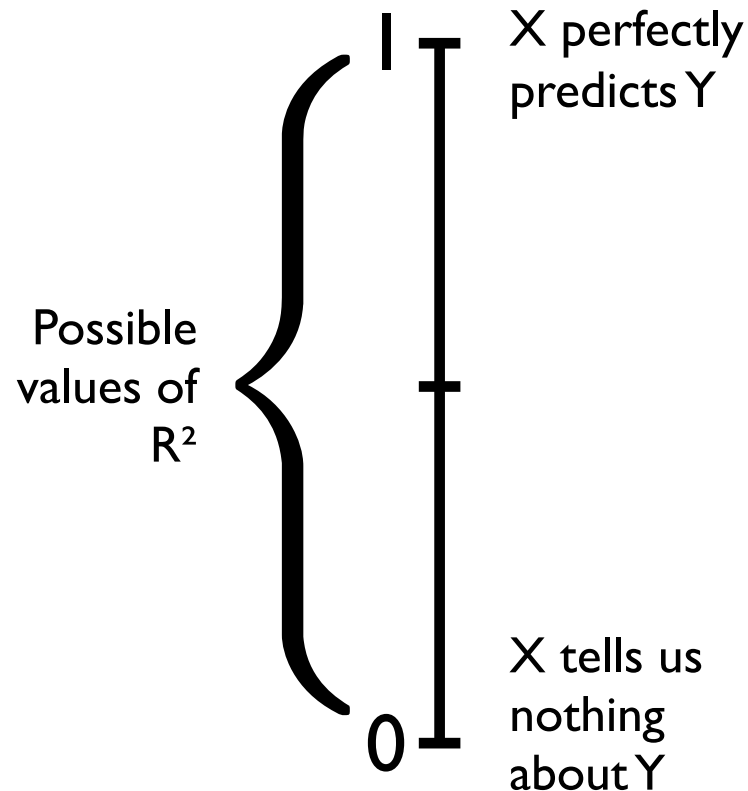
How much better are the predictions from our OLS regression line than the predictions from a flat line (i.e. not using X at all)?



R²: intuition

How much better are the predictions from our OLS regression line than the predictions from a flat line (i.e. not using X at all)?

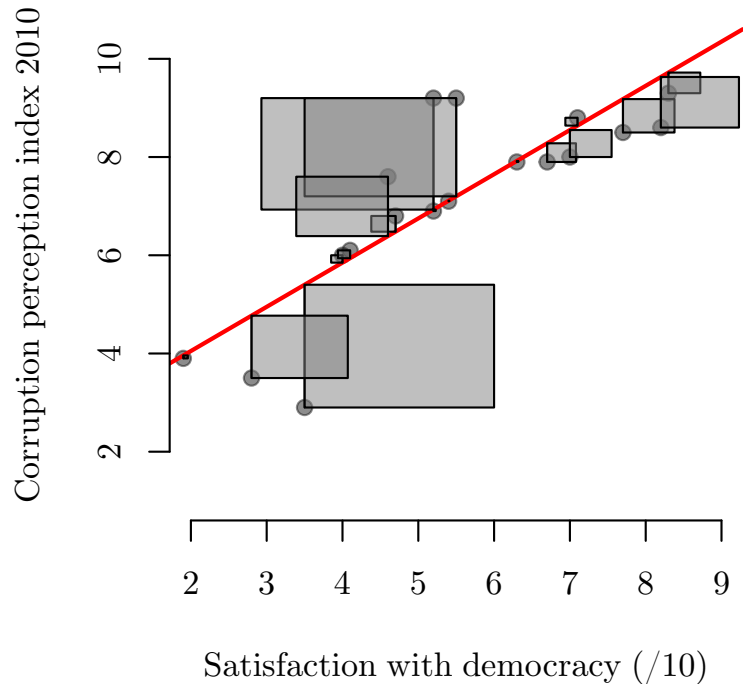
How much of the variation in Y is “explained” by the variation in X ?



R²: calculation

R²: calculation

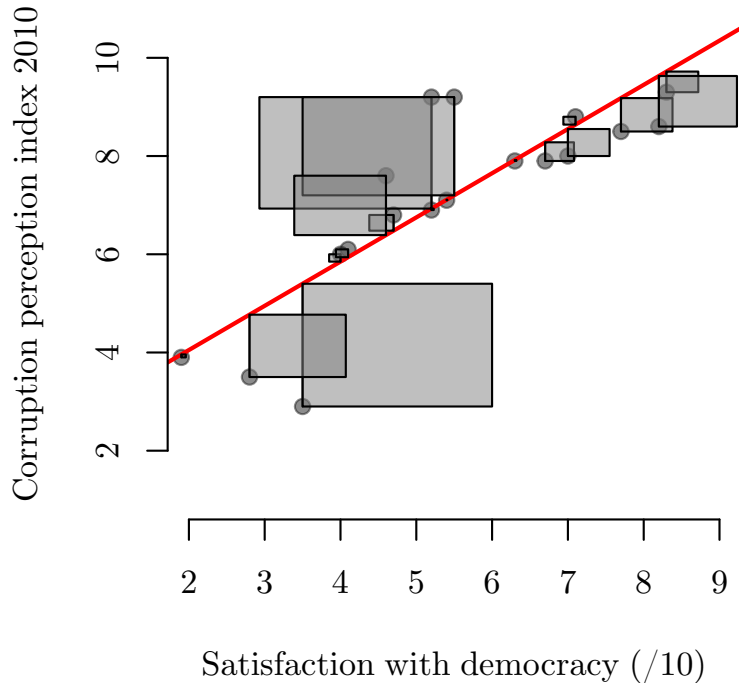
Sum of squared residuals:
20.808



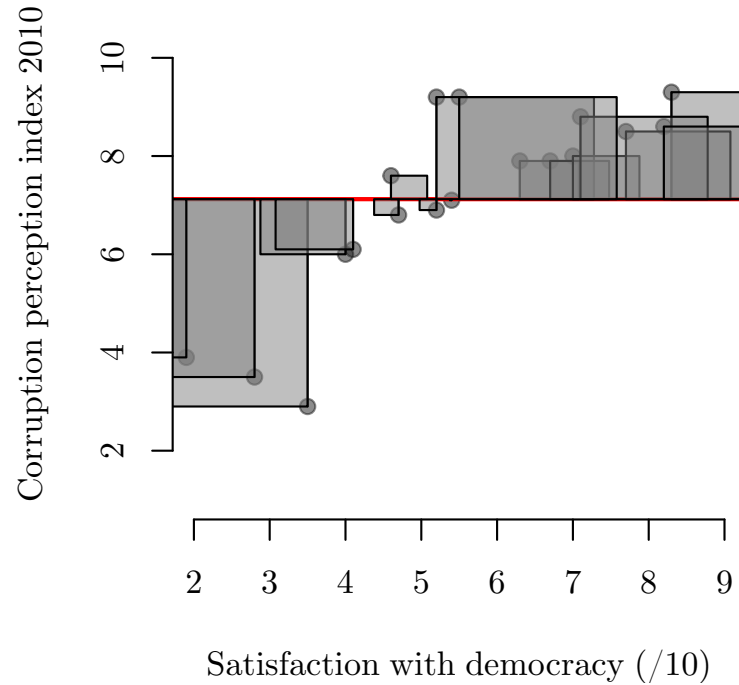
R²: calculation

“Total sum of squares”

Sum of squared residuals:
20.808



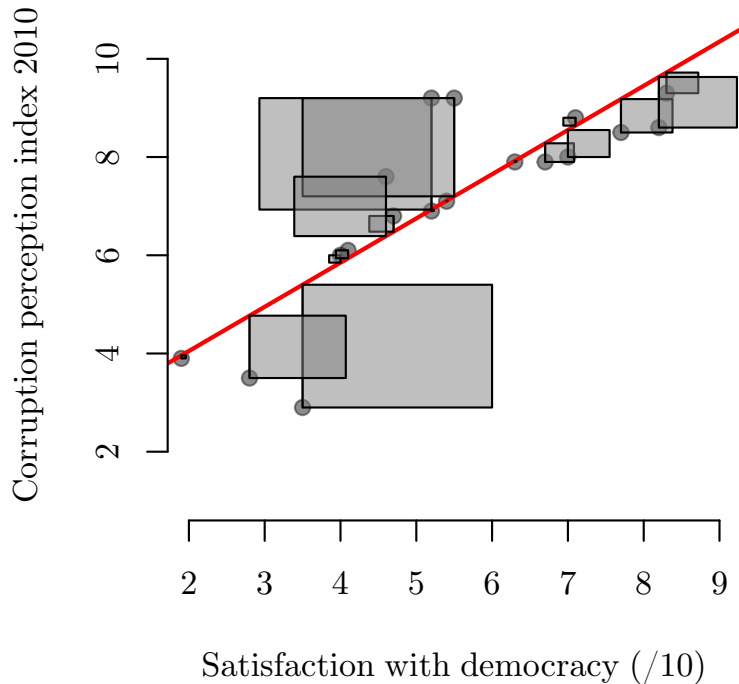
Sum of squared residuals:
66.271



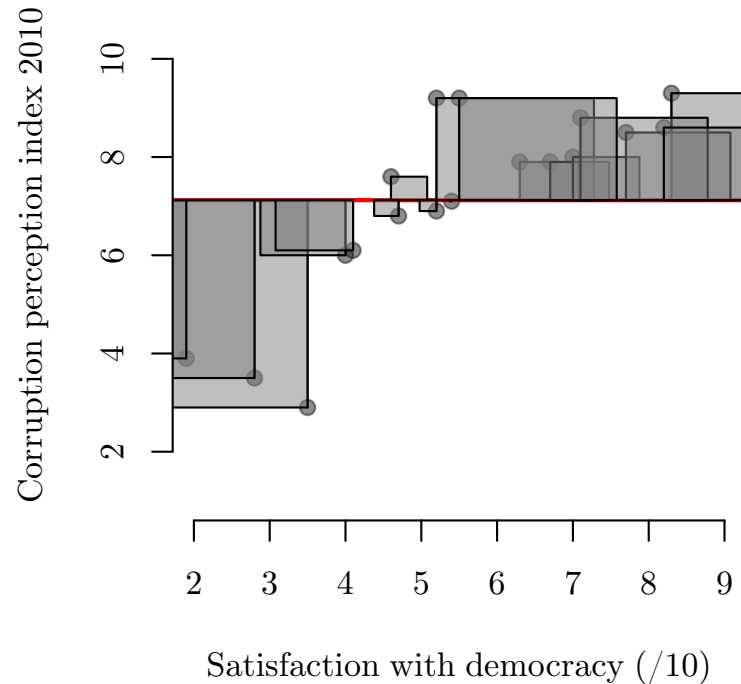
R²: calculation

“Total sum of squares”

Sum of squared residuals:
20.808



Sum of squared residuals:
66.271

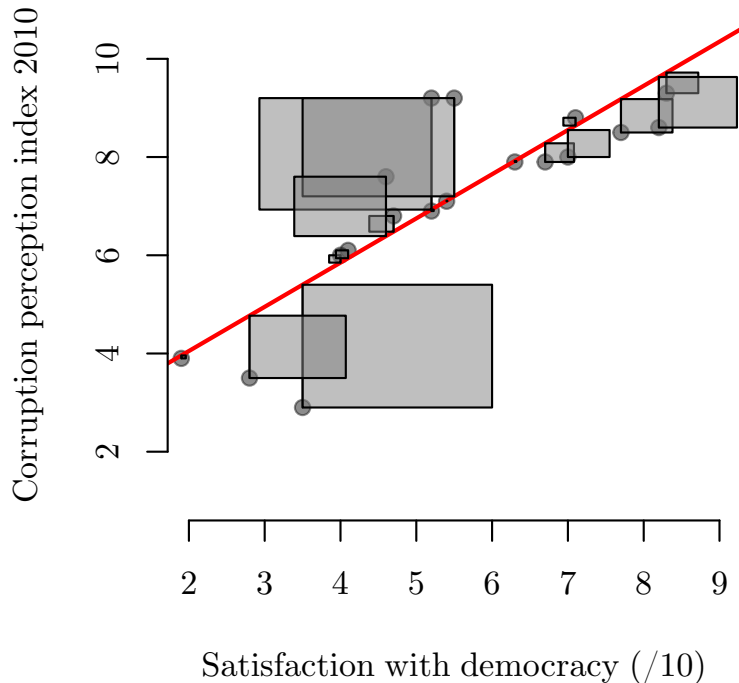


$$1 - \frac{20.808}{66.271} =$$

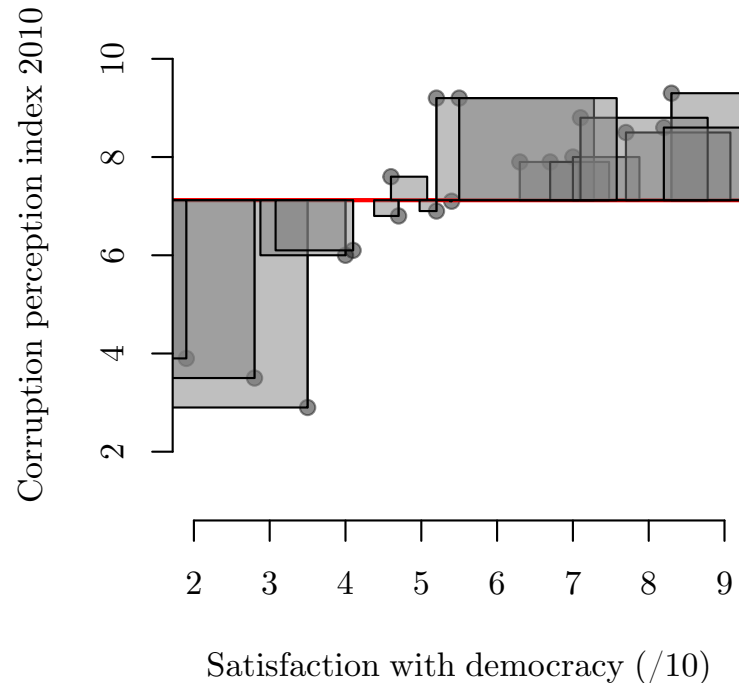
R²: calculation

“Total sum of squares”

Sum of squared residuals:
20.808



Sum of squared residuals:
66.271



$$1 - \frac{20.808}{66.271} = 0.6864$$

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression

output:

- intercept
- slope
- R^2

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression

output:

- intercept
- slope
- R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression output:
 - intercept
 - slope
 - R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

For bivariate relationships,
 $R^2 = \text{correlation}^2$

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression

output:

- intercept
- slope
- R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

For bivariate relationships, $R^2 = \text{correlation}^2$

Covariance and regression slope (but not correlation) depend on the units

Connections between measures of bivariate relationships

Key measures:

- covariance
- correlation
- OLS regression

output:

- intercept
- slope
- R^2

For any two variables, covariance, correlation, and regression slope will all have the same sign.

For bivariate relationships, $R^2 = \text{correlation}^2$

Regression slope (but not covariance or correlation) depends on which is Y and which is X

Covariance and regression slope (but not correlation) depend on the units

Extra slides

- Why do we call it “regression”?
- Regression to the mean in the heights of Oxford students and their parents

The roots of regression

ANTHROPOLOGICAL MISCELLANEA.

REGRESSION *towards* MEDIOCRITY *in* HEREDITARY STATURE.

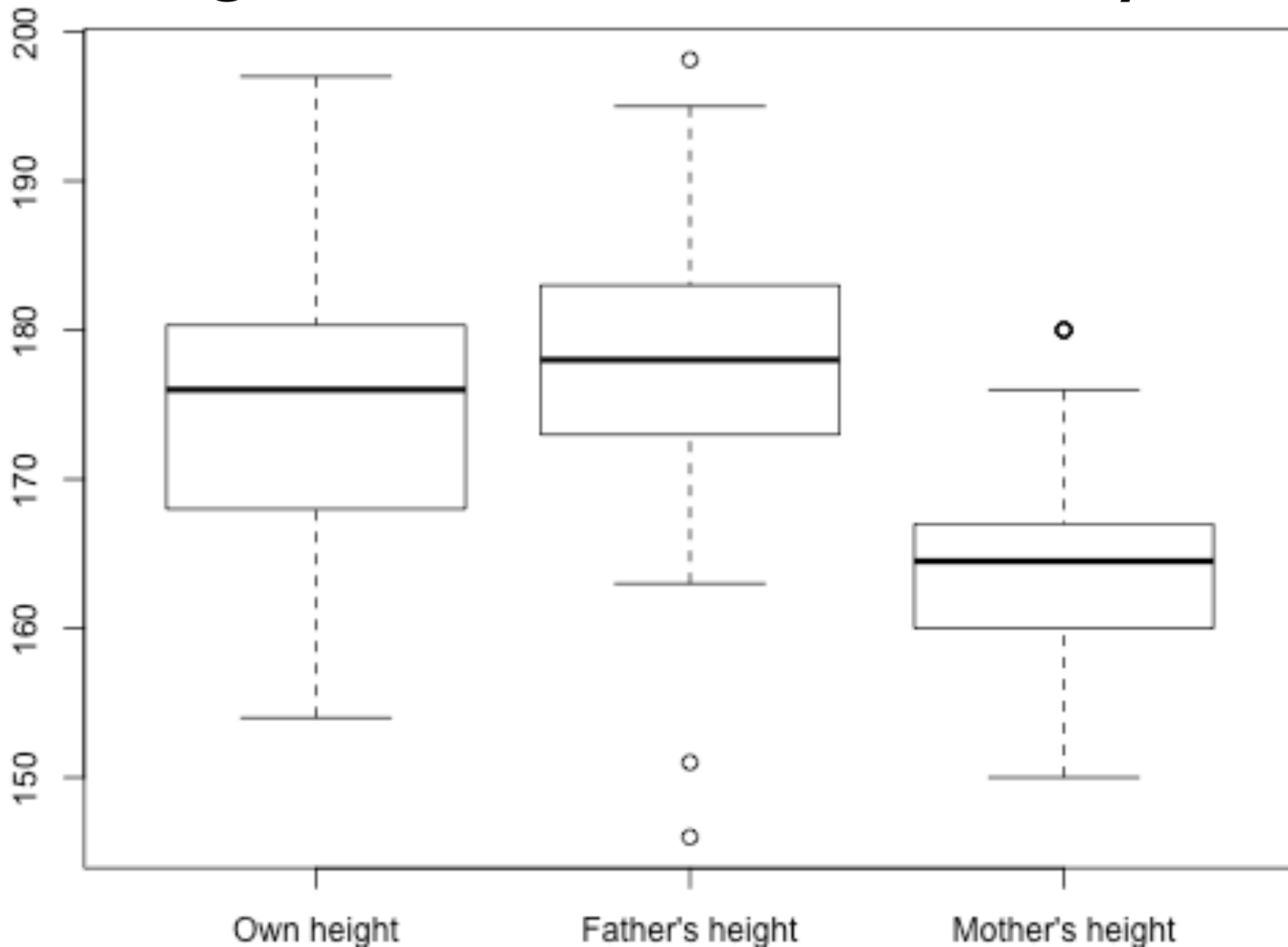
By FRANCIS GALTON, F.R.S., &c.

[WITH PLATES IX AND X.]

THIS memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. I reproduce here the portion of it which bears upon regression, together with some amplification where brevity had rendered it obscure, and I have added copies of the diagrams suspended at the meeting, without which the letterpress is necessarily difficult to follow. My object is to place beyond doubt the existence of a simple and far-reaching law that governs the hereditary transmission of, I believe, every one of those simple qualities which all possess, though in unequal degrees. I once before ventured to draw attention to this law on far more slender evidence than I now possess.

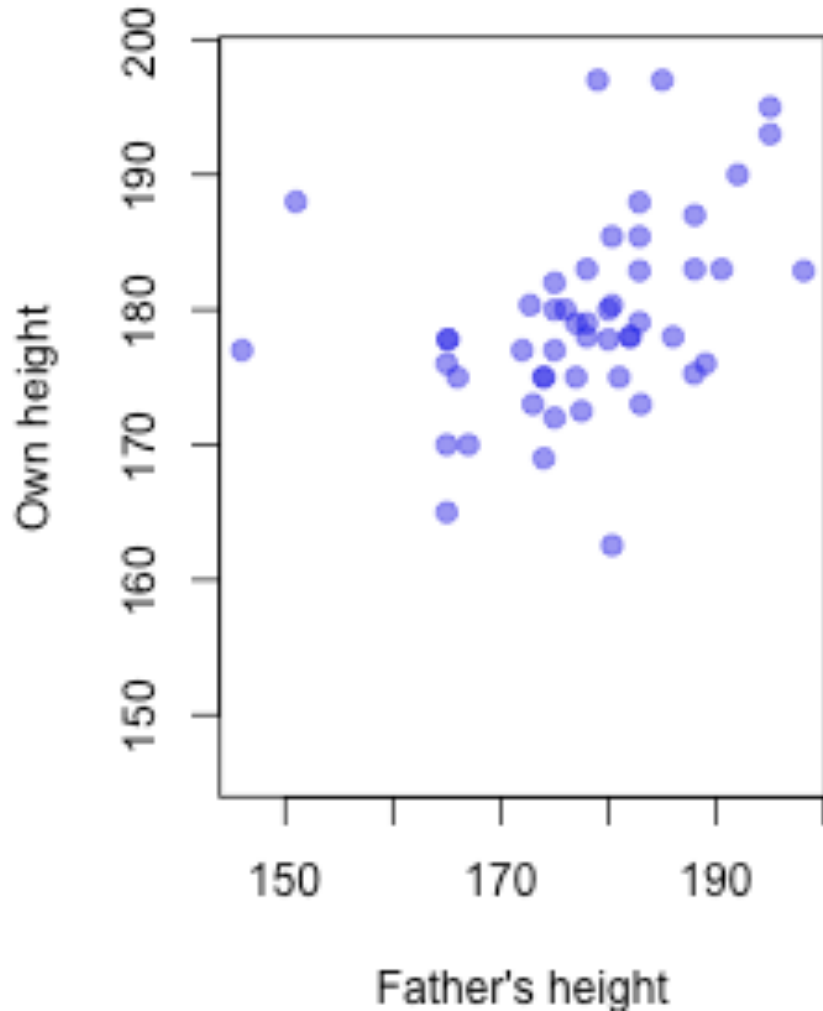


Height data from 2015 survey

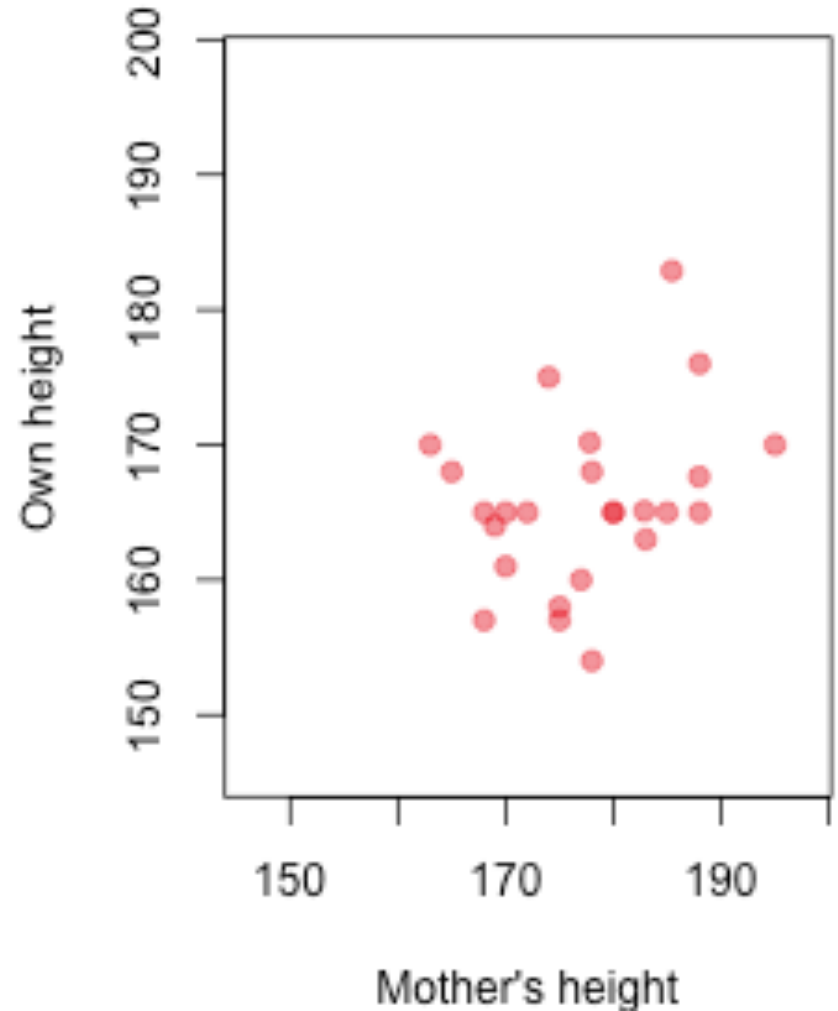


Height data, by gender

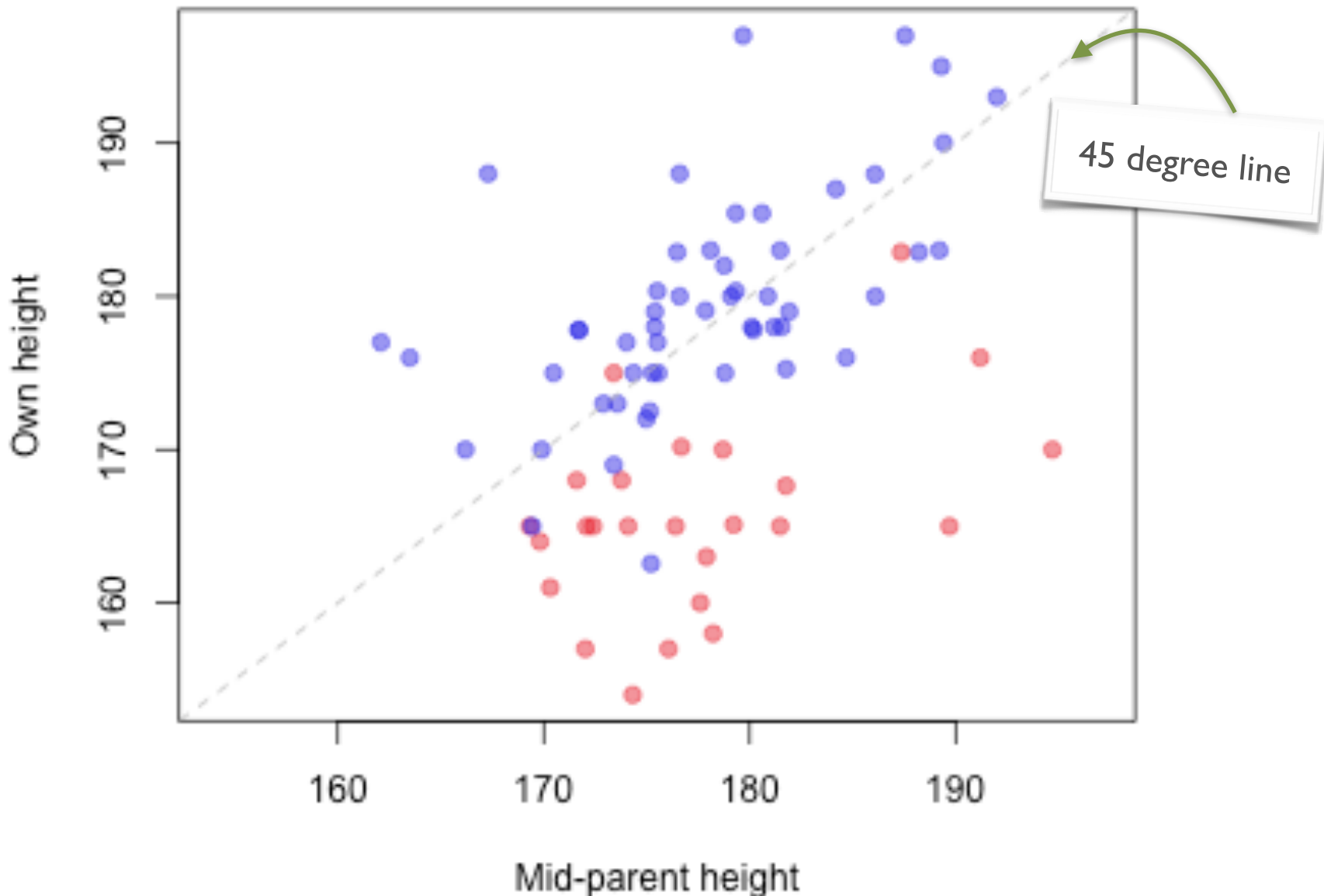
Male students



Female students



Own height and “mid-parent” height



Regression line shows “regression to the mean”!

