# Bivariate relationships

Week 6

23 February, 2015

Prof. Andrew Eggers



# Financial Statements of the Oxford Colleges (2012-13)

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.

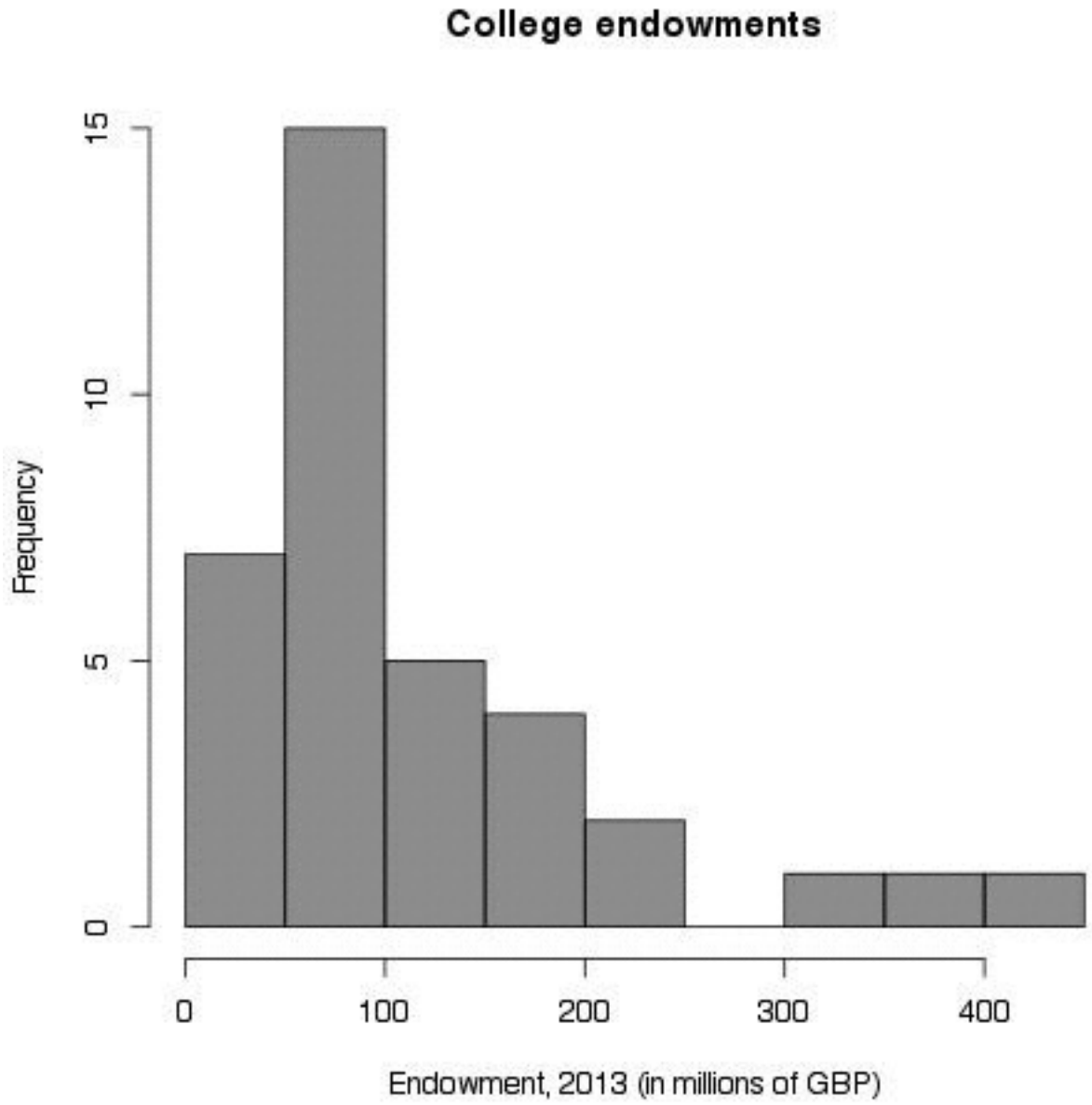
The colleges are independent, self-governing and financially autonomous and their accounts are published under the accounting convention developed by the Charity Commissions for use by charities in the UK (the Charity SORP).

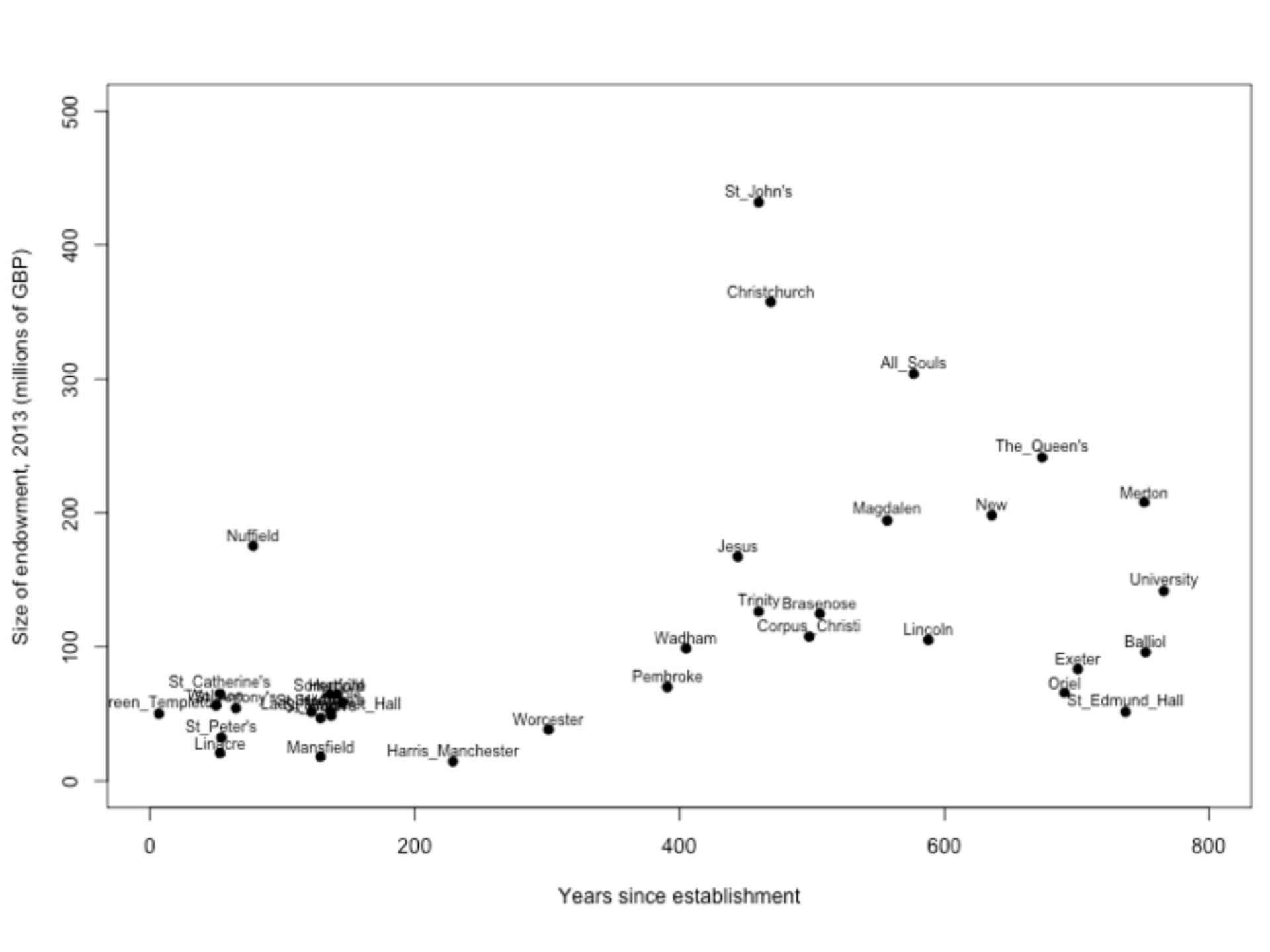
Kellogg College and St Cross College do not have Royal Charters and, for accounting purposes, are departments of the University. As such, their financial results are consolidated into the University's financial statements.

http://goo.gl/1pJA2r

Console ~/

```
> ## 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
```

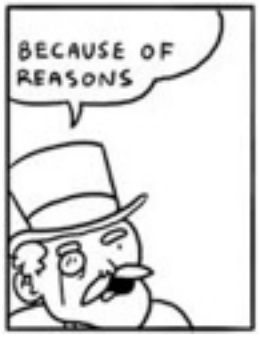

College endowments
Frequency
0
5
10
15
0
100
200
300
400
Endowment, 2013 (in millions of GBP)


Size of endowment, 2013 (millions of GBP)
0
100
200
300
400
500
0
200
400
600
800
Years since establishment
St_John's
Christchurch
All_Souls
The_Queen's
Merton
New
Magdalen
Nuffield
Jesus
University
Trinity
Brasenose
Corpus_Christi
Lincoln
Wadham
Balliol
Exeter
Pembroke
Oriel
St_Edmund_Hall
St_Catherine's
Worcester
St_Peter's
Linacre
Mansfield
Harris_Manchester

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


**Descriptive/predictive** questions:

- How are age and size of endowment related?
- What is the average endowment of a college that is X years old?

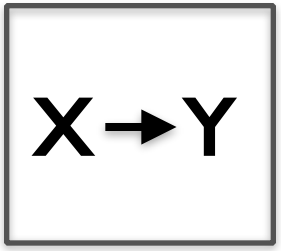



**Explanatory** questions:

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

**Causal** questions:

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



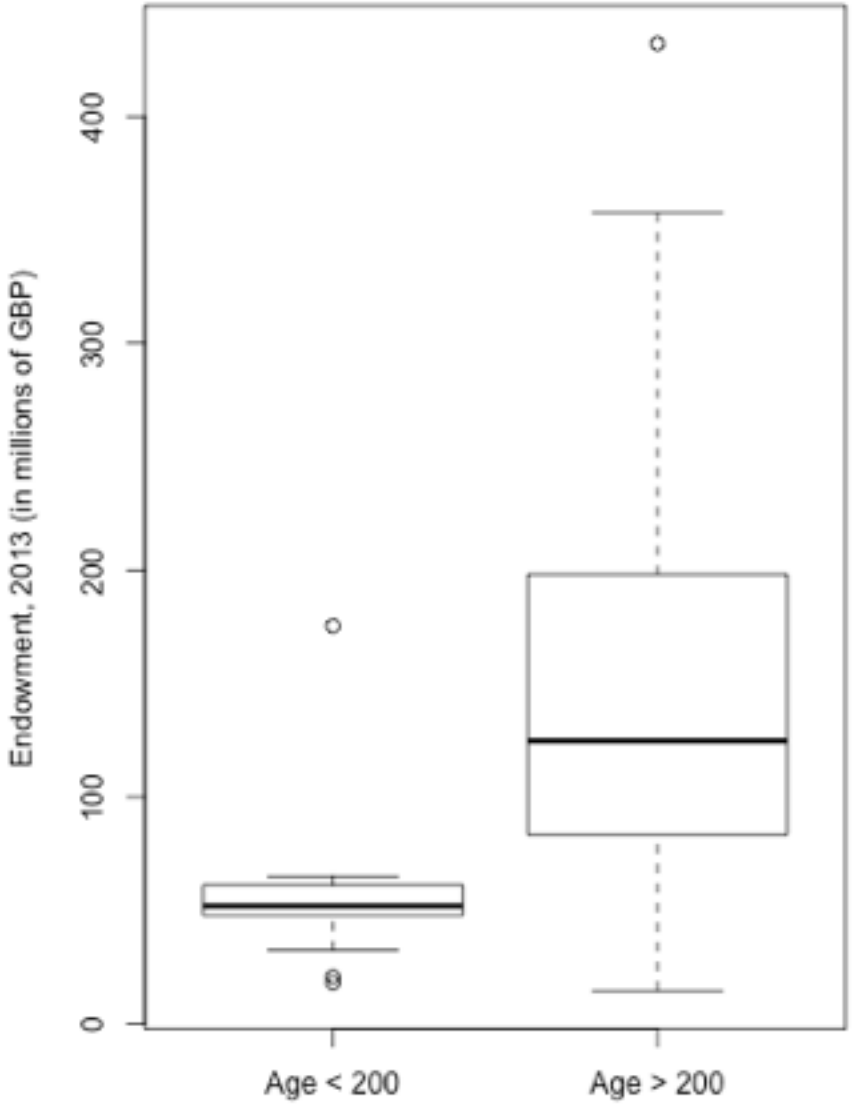
# Various ways to summarize a relationship between two variables

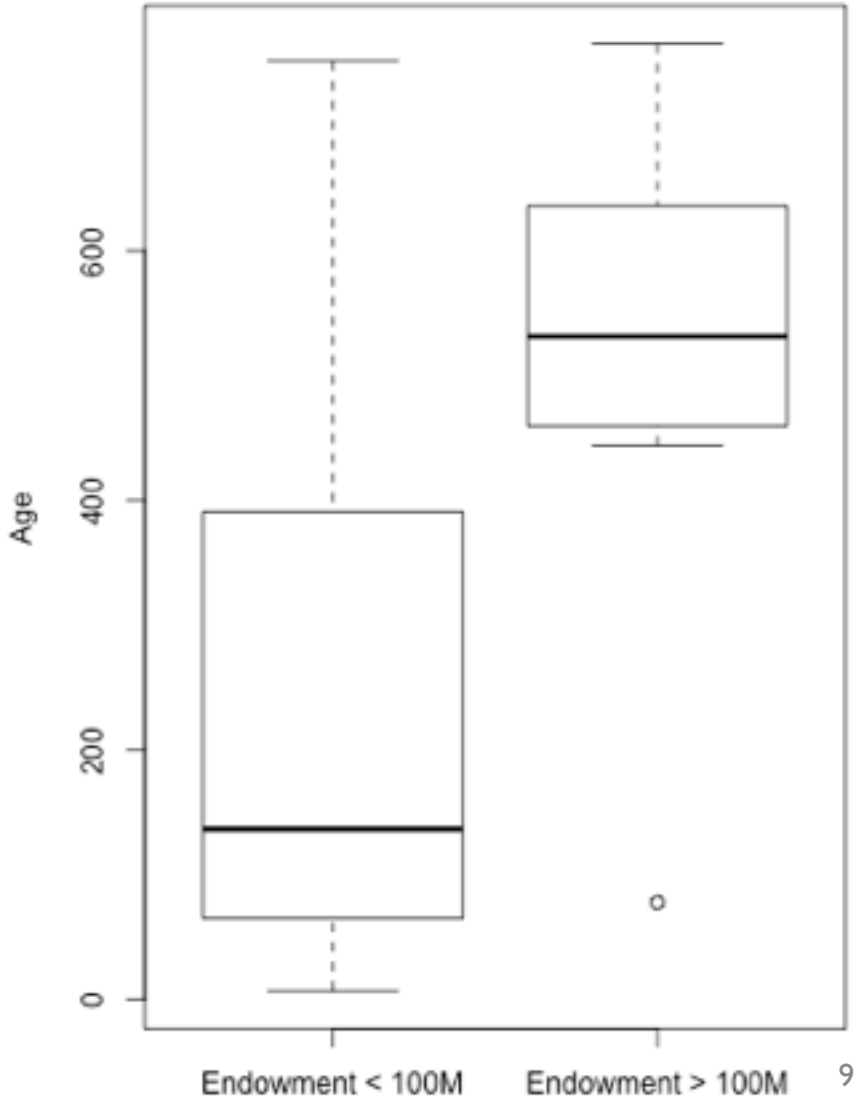
- Show the scatterplot!
- Compare boxplots of one variable across categories of the other
- Show/report mean of one variable across categories of the other (binned means or kernel average smoother)
- Report covariance/correlation
- Report regression coefficient(s)
- Report predicted values of one variable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

**But first: one minute to provide us with some data!**

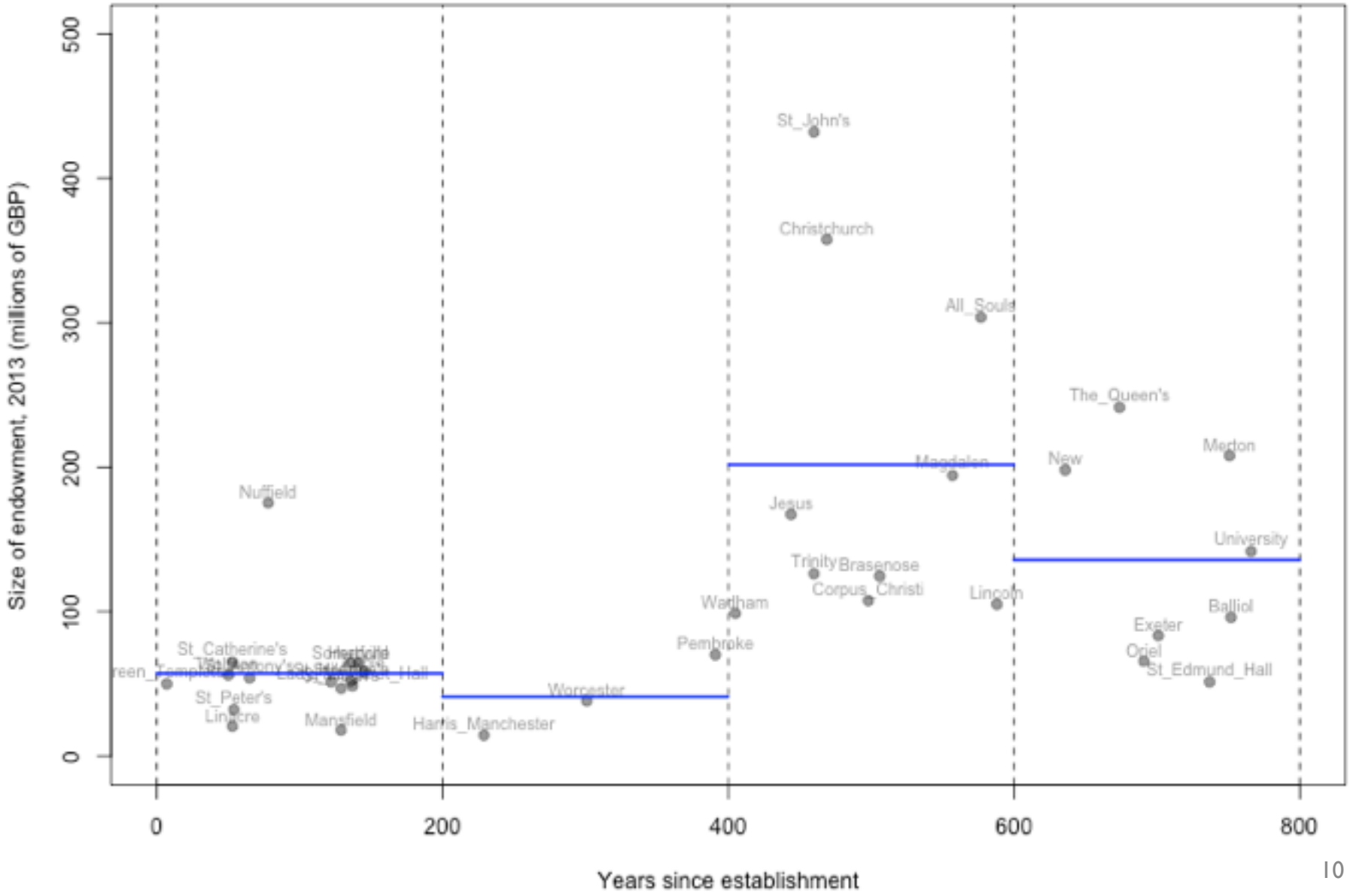
- get out a device (smartphone, laptop, tablet)
- go to http://andy.egge.rs/form.html
- click on the link
- fill out the survey (5 questions)
- put away your device and look up

# Boxplots: which do you prefer?

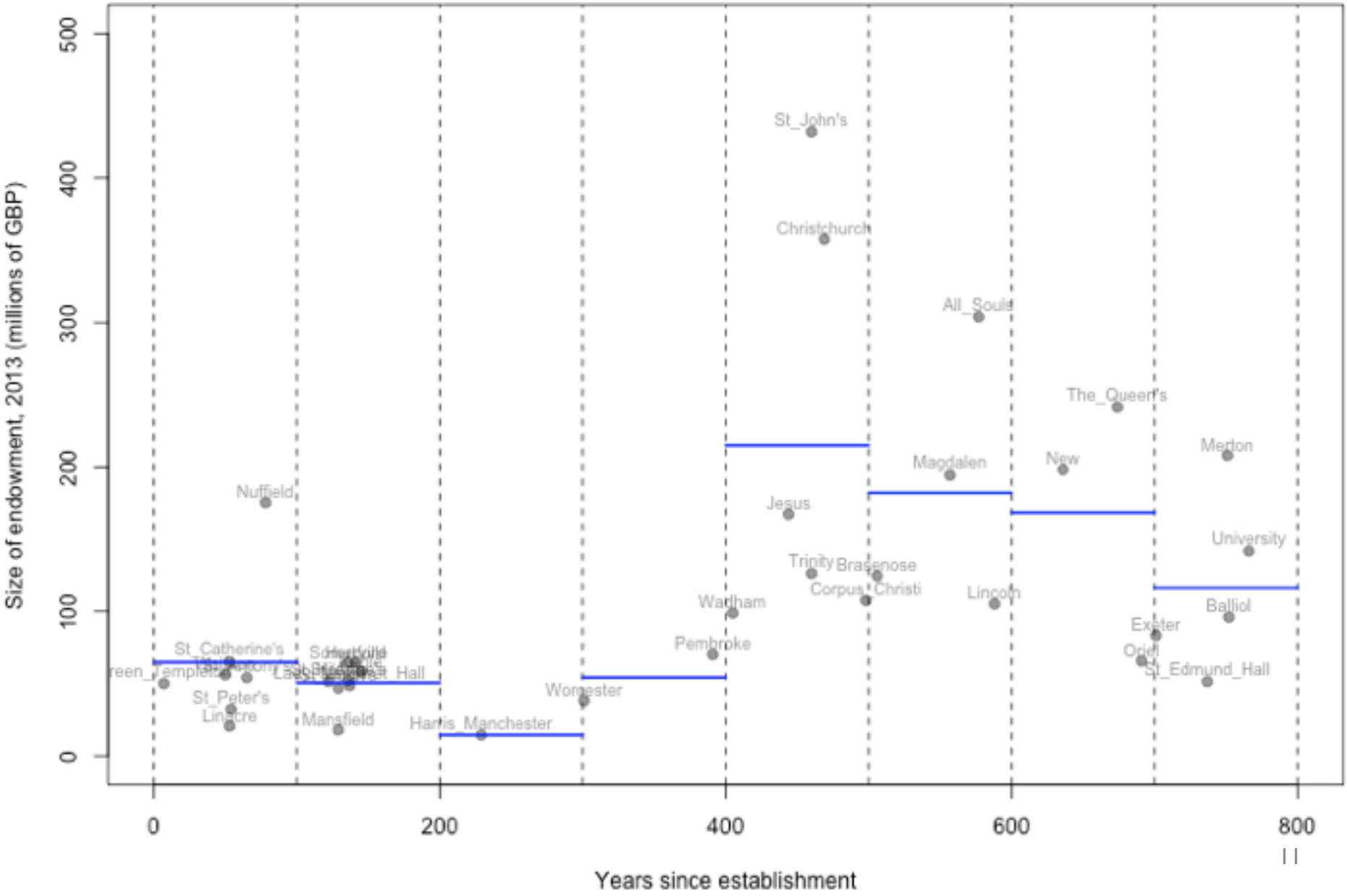


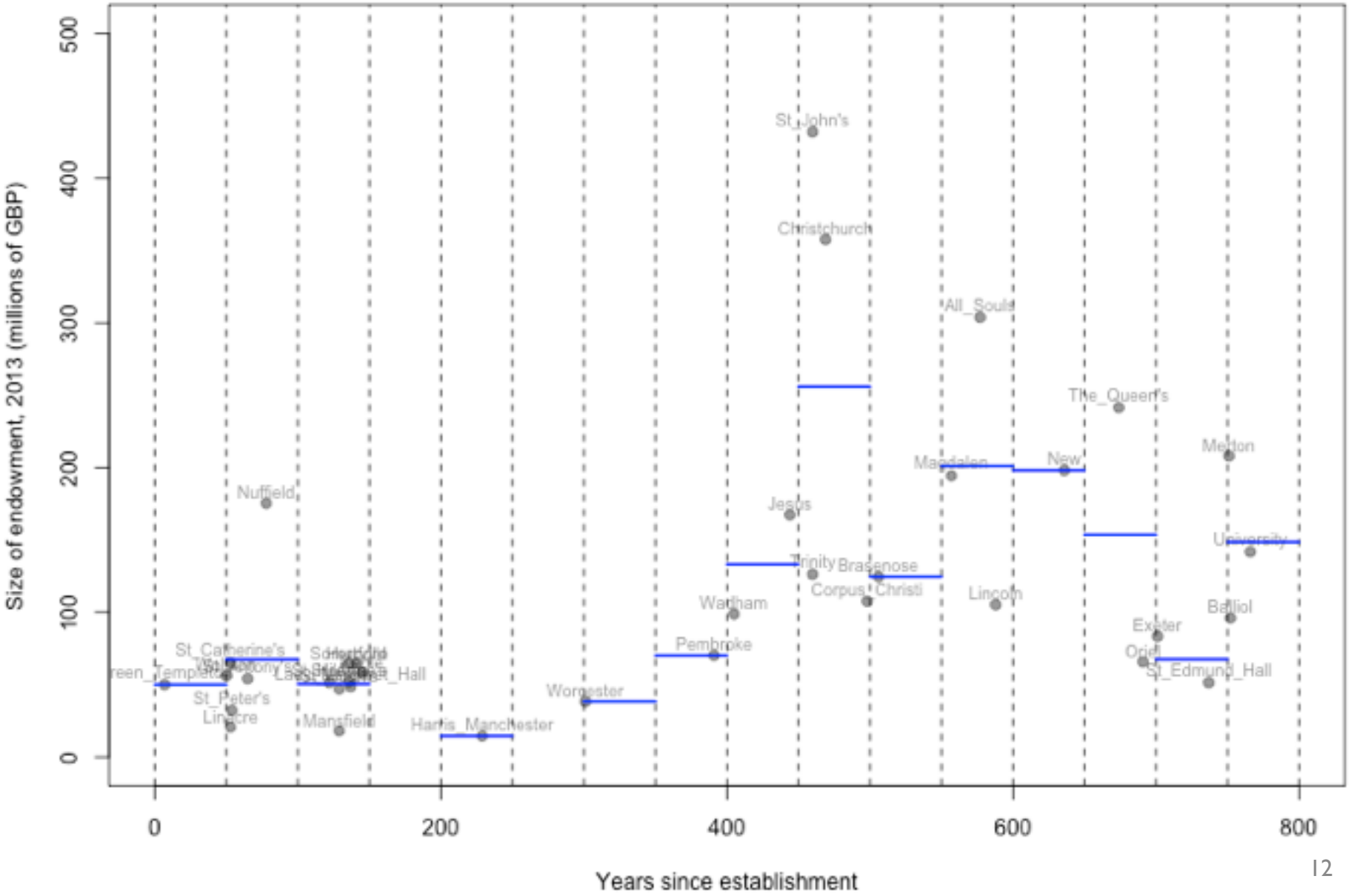
# Mean endowment within 200-year intervals



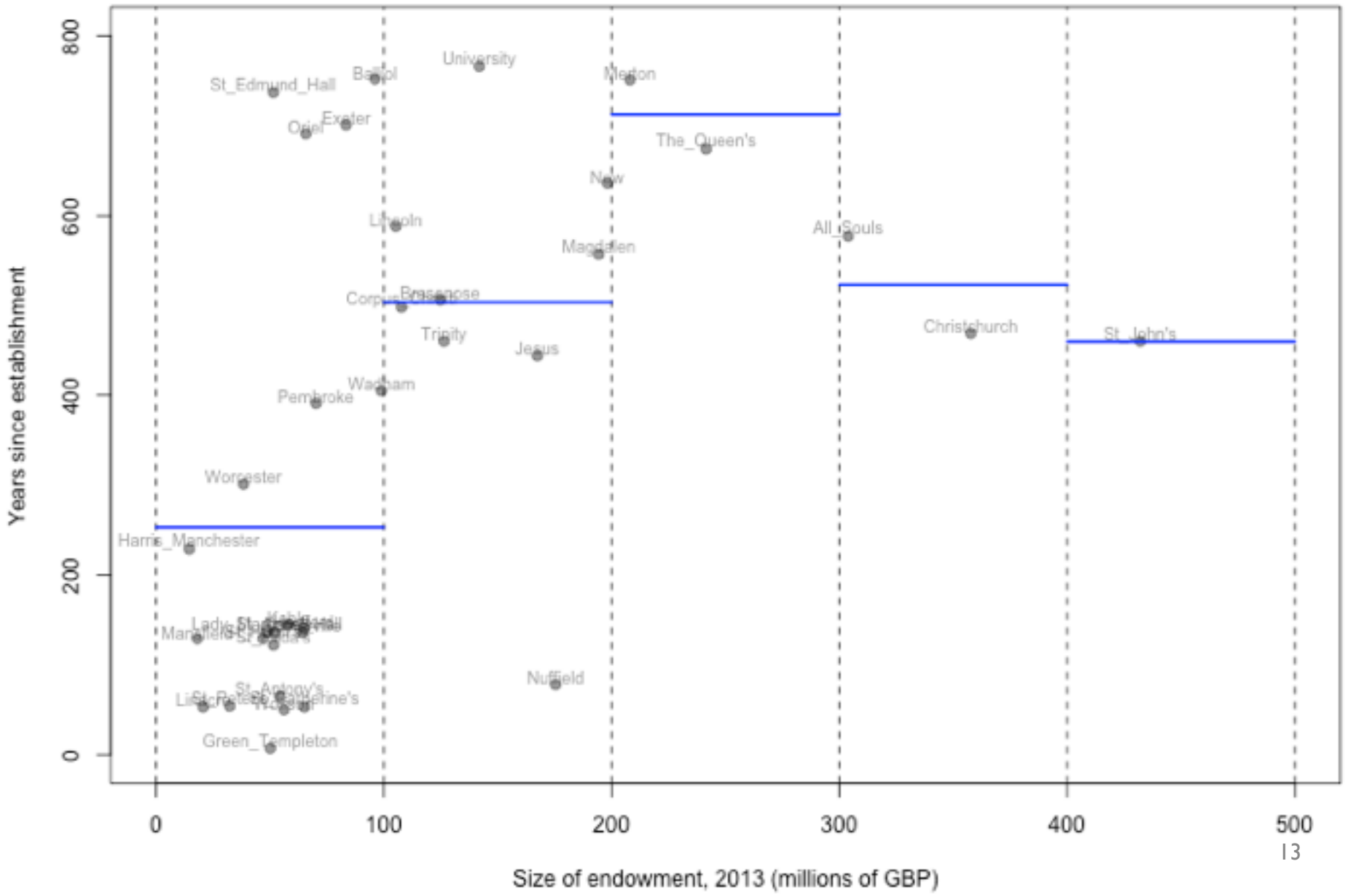
# Mean endowment within 100-year intervals



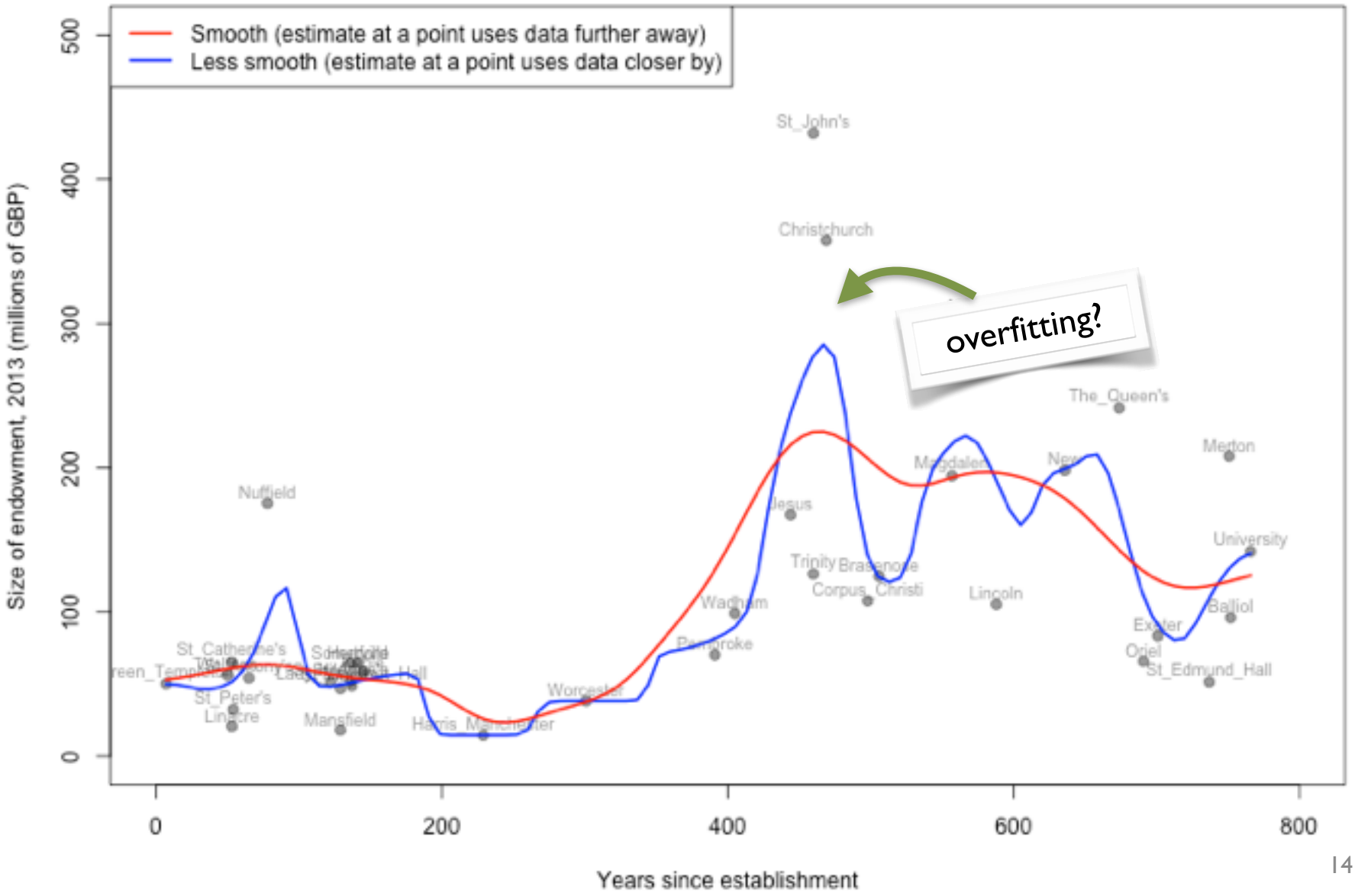
# Mean endowment within 50-year intervals



# Mean age within intervals of endowment



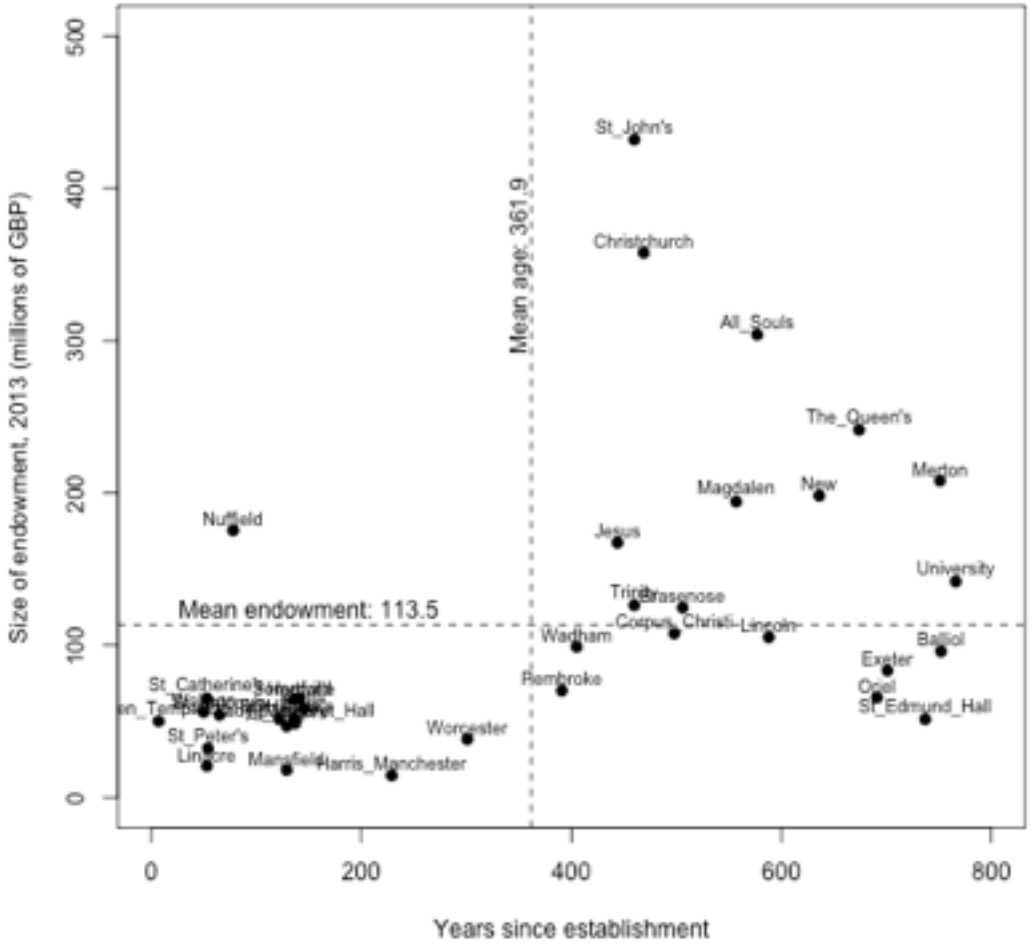
# Kernel smoother: estimate at a point is (weighted) average of nearby points



# Covariance: a measure of linear association

$$\mathrm{Cov}(x, y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

"mean of y"




| College | $(x_i - \overline{x})$ | $(y_i - \overline{y})$ | $(x_i - \overline{x})(y_i - \overline{y})$ |
|---|---|---|---|
| Mansfield | -232.9 | -95.4 | 22,218.1 |
| Nuffield | -283.9 | 61.9 | -17,579.2 |
| St. John's | 98.1 | 318.6 | 31,256.5 |
| Wadham | 43.1 | -14.6 | -627.6 |

etc., for all the colleges, and take the average of the final column. Or, more simply:

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

# Two facts about covariance

$$\mathrm{Cov}(x, y) = \frac{\Sigma_i (x_i - \overline{x})(y_i - \overline{y})}{n - 1}$$

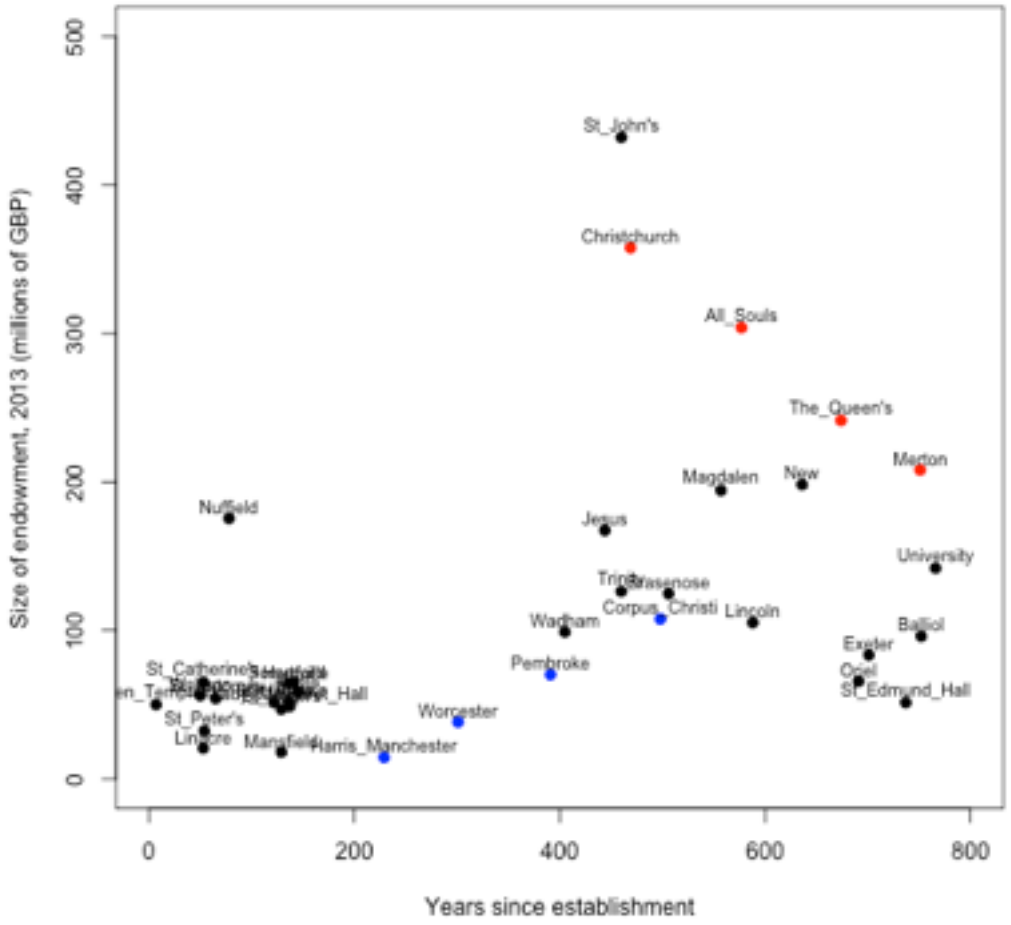
**FACT 1:** $\mathrm{Cov}(x, y) = \mathrm{Cov}(y, x)$

**FACT 2:** $\mathrm{Cov}(x, x) = \mathrm{Var}(x)$

# Correlation: a scale-invariant measure of linear association

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

"standard deviation of y"




| *Sample* | *Cov(x,y)* | *Cor(x,y)* |
|---|---|---|
| All | 11,838.5 | 0.471 |
| All, with y in 1000s of GBP | 11,828,503.0 | 0.471 |
| Red only | -8093.7 | -0.99785 |
| Blue only | 4697.3 | 0.99992 |
| Red and blue | 15083.7 | 0.658 |

# Five facts about correlation

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

**FACT 1:** $\mathrm{Cor}(x,y) \in [-1,1] \quad \forall \quad x,y$

**FACT 2:** $\mathrm{Cor}(x,y) = \mathrm{Cov}\left(\frac{x}{\mathrm{sd}(x)}, \frac{y}{\mathrm{sd}(y)}\right)$

**FACT 3:**

$\mathrm{Cor}(x,y) = \mathrm{Cor}(y,x)$

**FACT 4:**

The x,y correlation for a line of points will be 1 or -1 unless it is vertical or horizontal, in which case it is undefined.

**FACT 5:**

# Finally: regression!

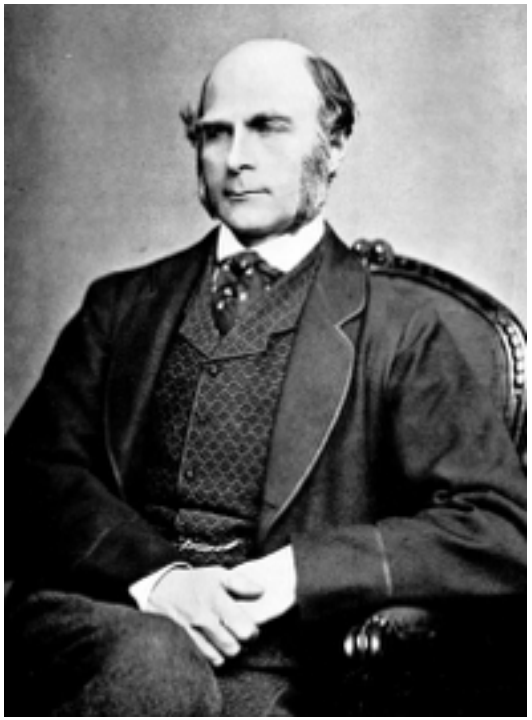
## ANTHROPOLOGICAL MISCELLANEA.

### Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.

[With Plates IX and X.]

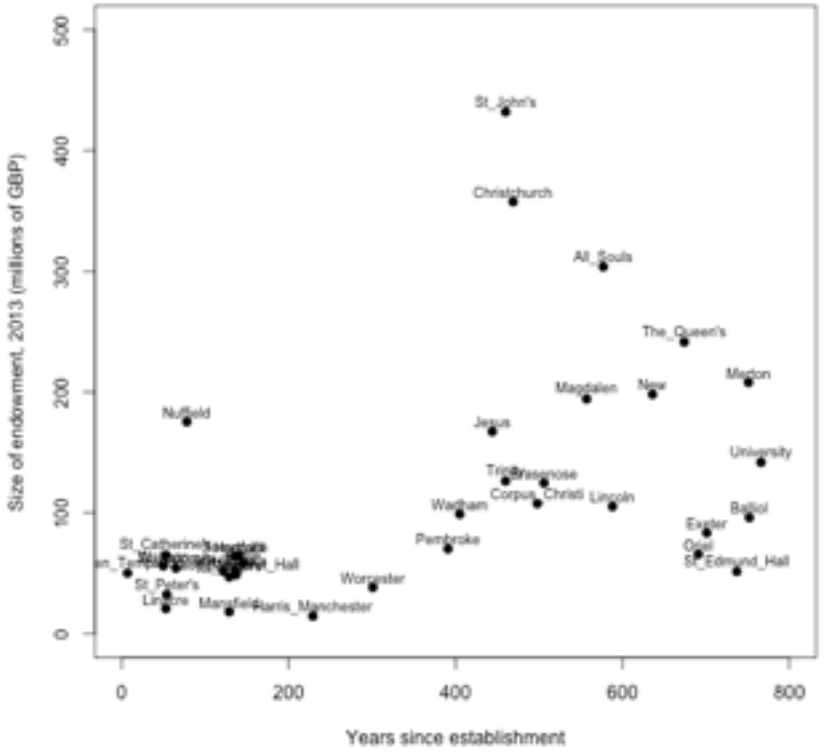
This memoir contains the data upon which the remarks on the Law of Regression were founded, that I made in my Presidential Address to Section H, at Aberdeen. That address, which will appear in due course in the Journal of the British Association, has already been published in "Nature," September 24th. 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.



Francis Galton, *Journal of the Anthropological Institute of Great Britain and Ireland*, 1886.

# Bivariate regression: drawing a line **near** some points

$$y_i = \alpha + \beta x_i$$




No way to draw a line **through** all of these points!

How can we produce a line "close" to the points?

- "Eyeball it"
- Draw a line that minimizes distance from the points to the line
  - Minimize the **total** distance from the points to the line (or equivalently, **average** distance)
  - Minimize the **median** distance from the points to the line
- Draw a line that minimizes total/average or median **vertical** distance from the points to the line (i.e. "prediction error", if we think of **x** predicting **y**)

Linear regression (ordinary least squares) minimizes the total/average **squared vertical distance** from the points to the line.

# Bivariate regression: more formally

$$y_i = \hat{\alpha} + \hat{\beta} x_i + \hat{u}_i$$

- Greek letters without hats: parameters in the **population.**
- Greek letters with "hats"*: estimates of those parameters produced from the sample.

* In Andrea's lecture, Roman letters.

Linear regression (OLS) solves:

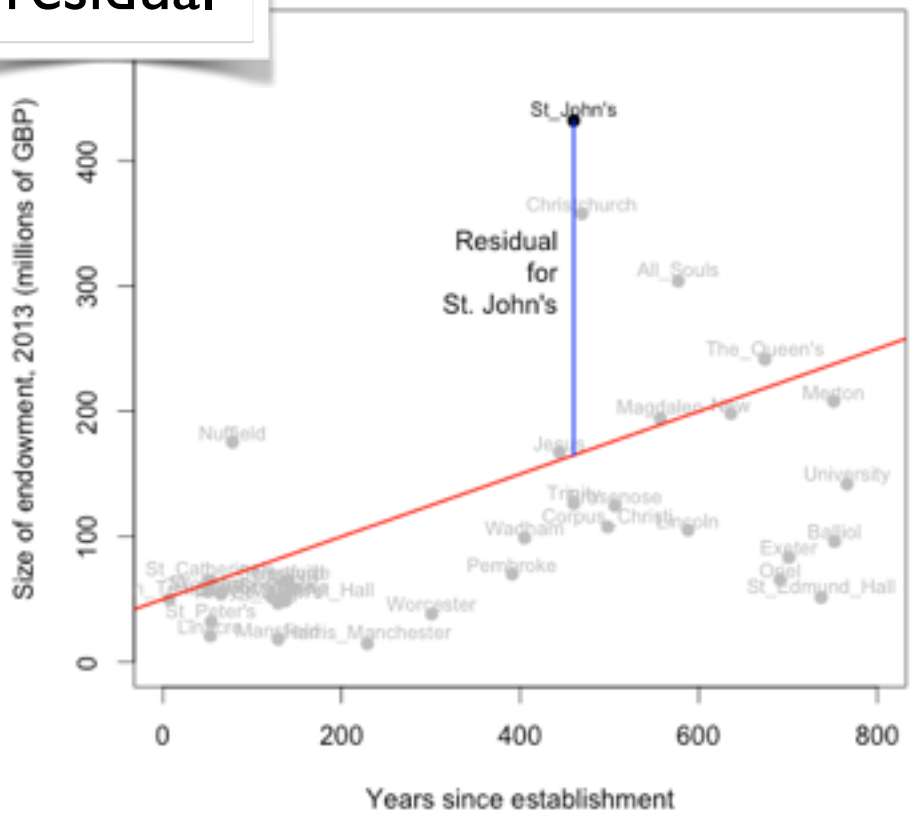
$$\min_{\hat{\alpha},\hat{\beta}} \sum_i \hat{u_i}^2$$

"residual"

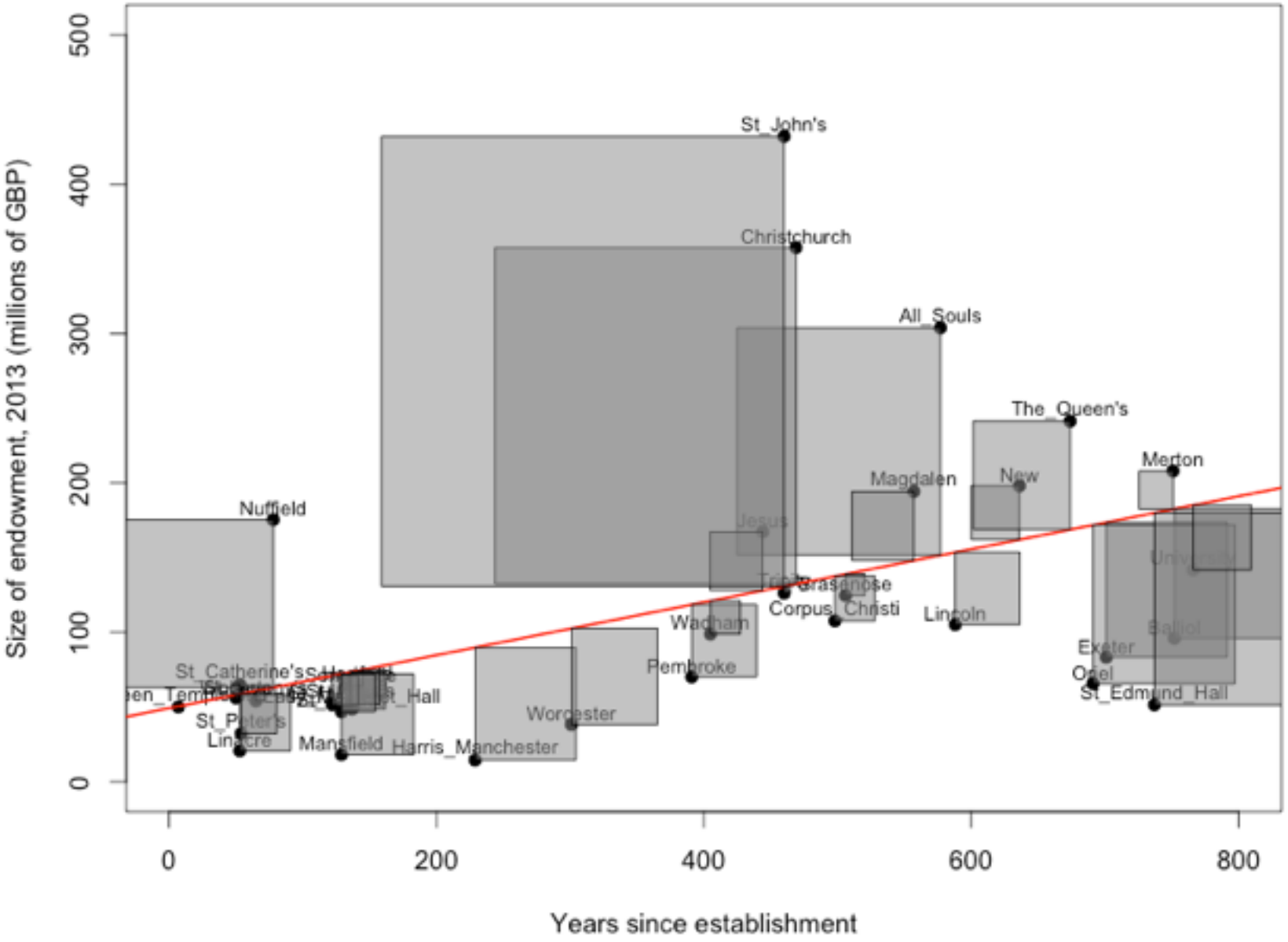
Minimization (remember calculus?) gives:

$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

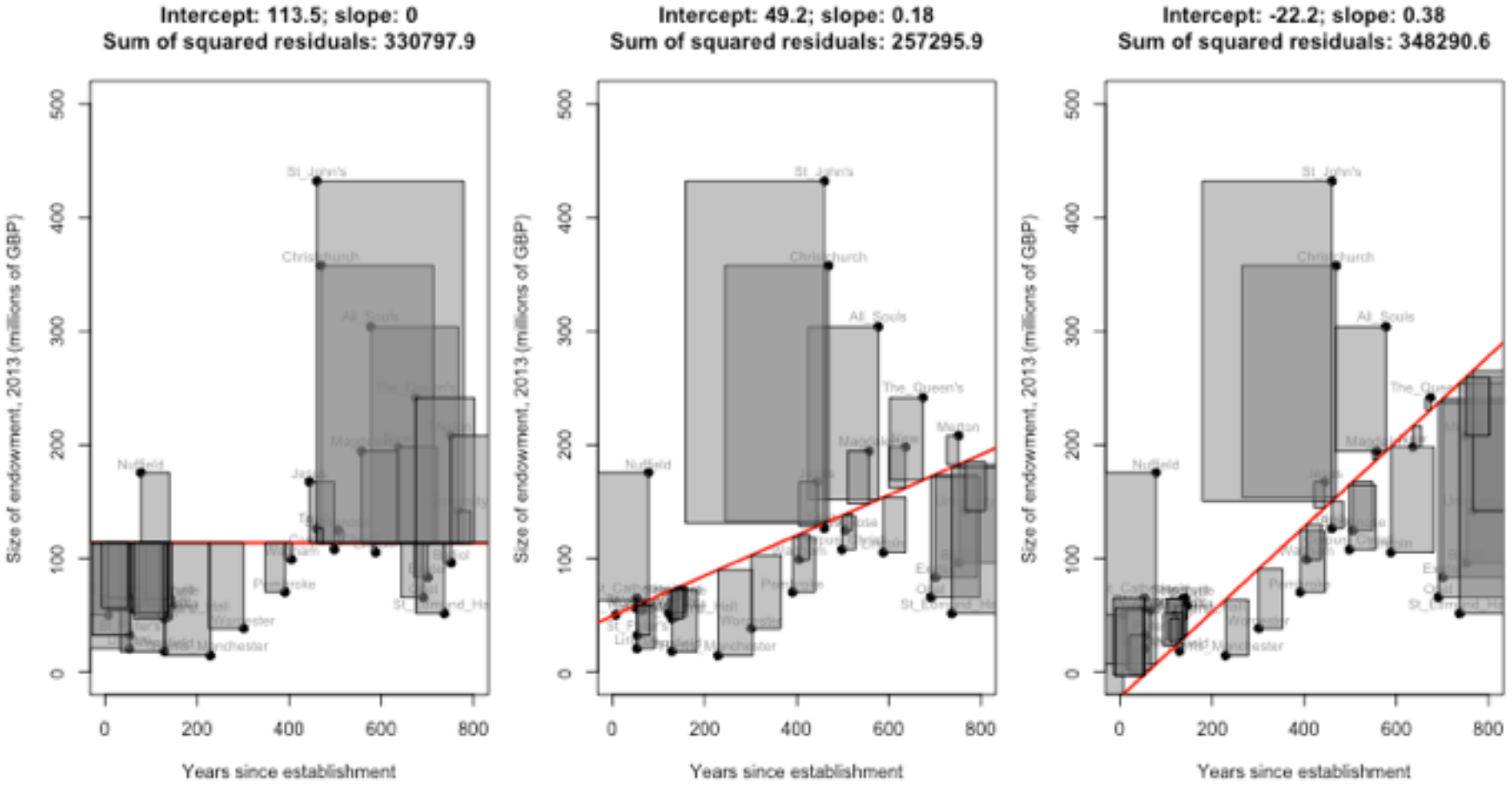
$$\hat{\beta} = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$



# Visualizing the sum of squared residuals



# Visualizing the minimization of the sum of squared residuals



# Visualizing minimization, continued

Recall that the OLS solution for the **intercept** is:

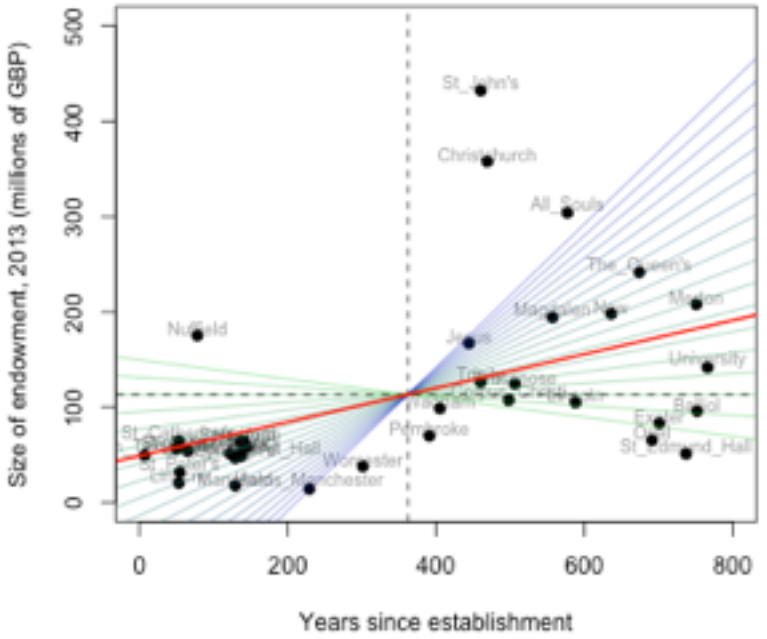
$$\hat{\alpha} = \overline{y} - \hat{\beta}\overline{x}$$

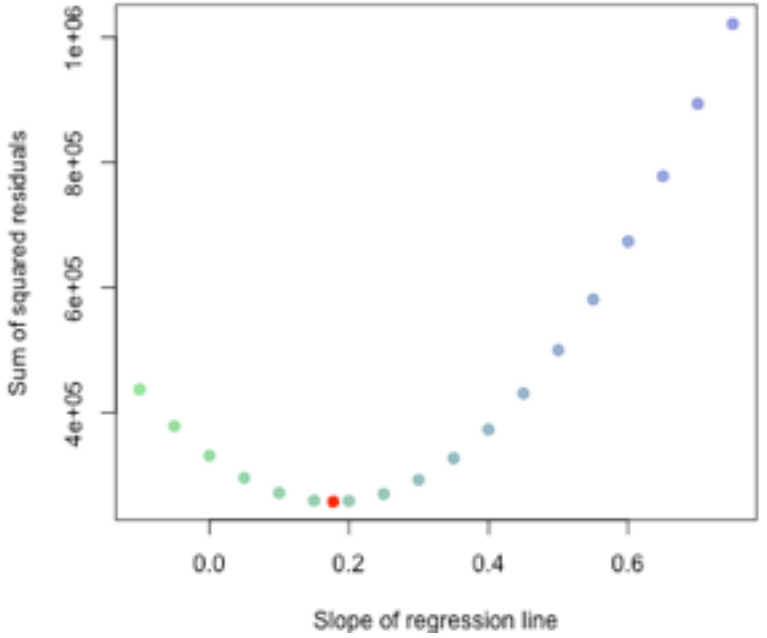
So what is the y value when we are at the mean of x?

$$\hat{\alpha} + \hat{\beta}\overline{x} = \overline{y}$$

So the bivariate regression line always goes through the point

$$(\overline{x}, \overline{y})$$



# Relationships among covariance, correlation, and slope of regression

Recall that the OLS solution for the **slope** is: $\hat{\beta} = \frac{\text{Cov}(x, y)}{\text{Var}(x)}$

Variances are always positive, so Cov(x,y), Cor(x,y) and $\hat{\beta}$ always have the same **sign**. For example:

```
> cov(d$age, d$endow.mill)
[1] 11828.5
> cor(d$age, d$endow.mill)
[1] 0.4713768
> coef(lm(d$endow.mill ~ d$age))[2]
    d$age
0.1775421
```

In one (important) special case, Cov(x,y), Cor(x,y) and $\hat{\beta}$ are **exactly the same:**

```
> d$std.age = d$age/sd(d$age)
> d$std.endow.mill = d$endow.mill/sd(d$endow.mill)
> cov(d$std.age, d$std.endow.mill)
[1] 0.4713768
> cor(d$std.age, d$std.endow.mill)
[1] 0.4713768
> coef(lm(d$std.endow.mill ~ d$std.age))[2]
d$std.age
0.4713768
```

**Correlation** tells you **how close to linear** a relationship is. Correlation of 0.99 means "very linear!" no matter what units we are talking about.

**Slope coefficient in a regression** tells you what a slope usually tells you: how much does the line go up (or down) when you move to right by one unit? It depends on units.

# Interpreting our regression coefficients

Suppose we had randomly left out a college.

We can use our regression estimate to predict its endowment!

- If it was created this year: 49.2 million
- If it was created 100 years ago: 49.2 + 100*0.178 = 67 million

(Note correlation does not yield a prediction.)

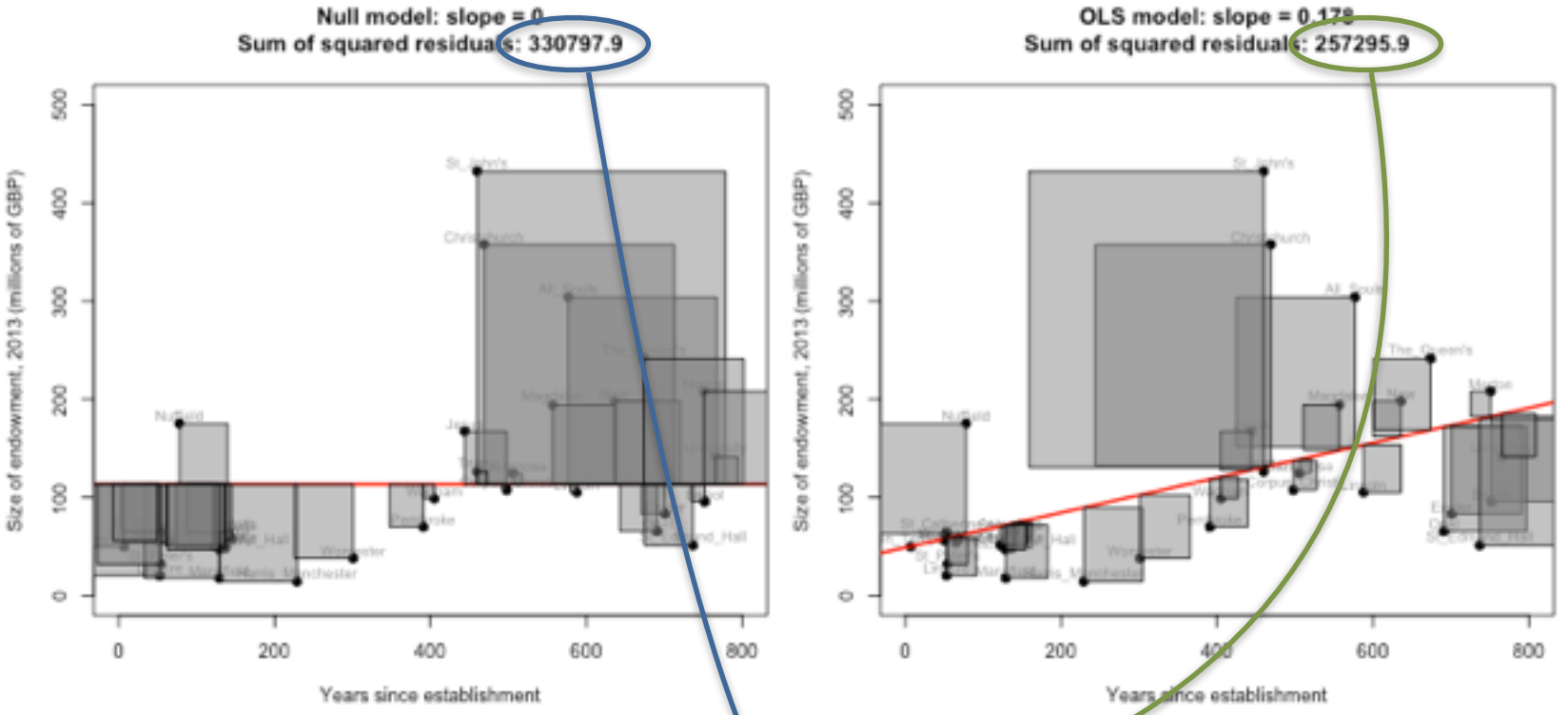
What about statements like these:

- "The effect of age on endowment is 0.178."
- "A 100-year increase in age leads to an increase of 17.8 million in endowment."

| | ***Dependent Variable: Endowment in millions*** |
|---|---|
| Intercept | 49.2<br>(25.2) |
| Age of college | 0.178**<br>(0.057) |
| Number of observations | 36 |
| R^2 | 0.222 |

# Interpreting R^2

R^2 measures how much of the variation in y is explained by x.




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

TSS is closely related to the variance of y:

```
> var(d$endow.mill)*(nrow(d) - 1)
[1] 330797.9
```

# Looking ahead

- Rest of this week: data labs, part 3!
- Next two weeks: multivariate regression and statistical inference