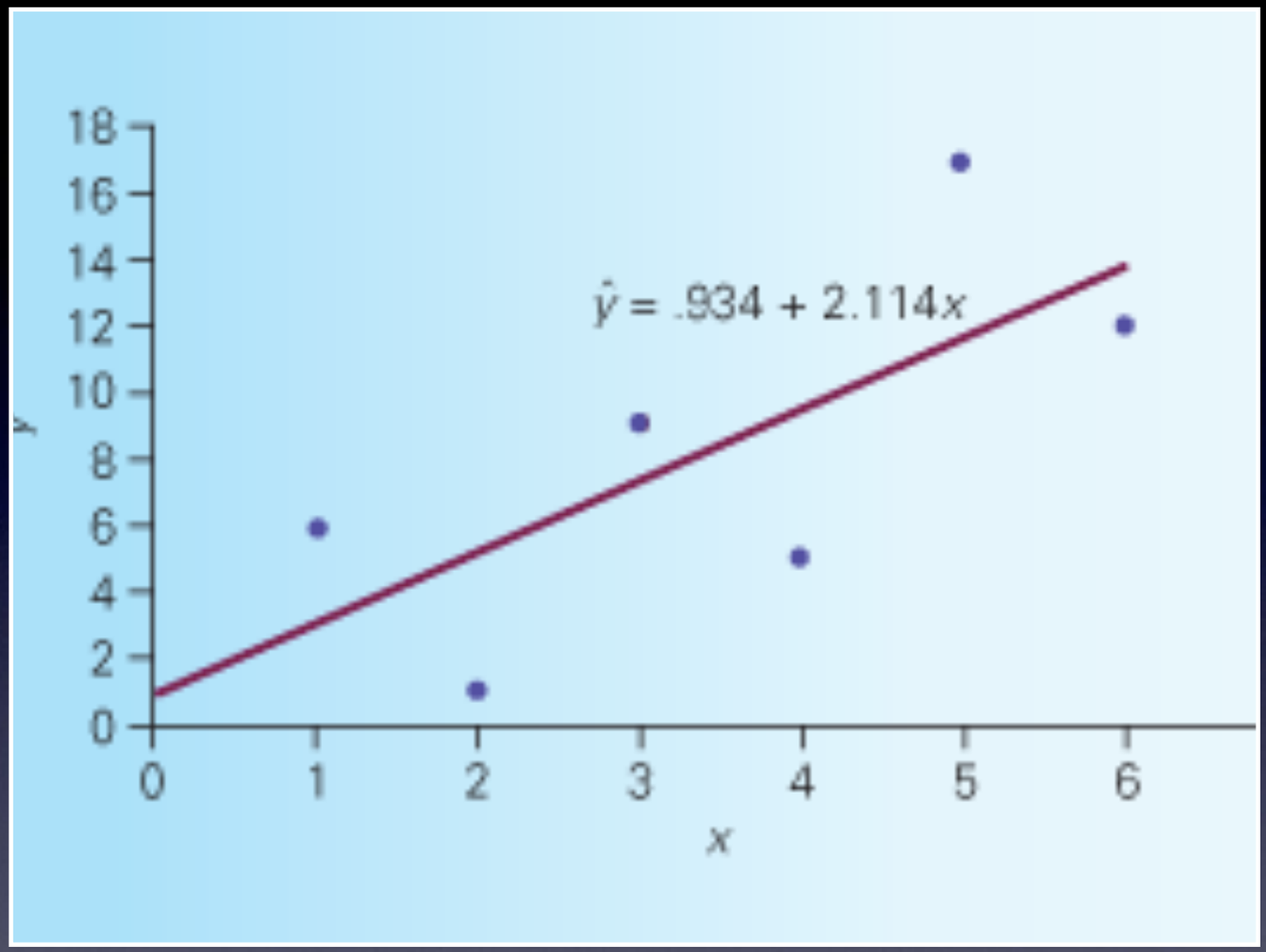


Regression analysis

Today

- Introduction to regression analysis and models
- Simple linear regression / OLS
- Assumptions
- Hypothesis testing
- Fit
- Correlation and regression and plotting

What can regression do?

Correlation:
Is there a relationship?
Direction/sign of relationship?

Regression:
Is there a relationship?
How are they related?
Strength of relationship

Increase one variable. What happens to the other?

Roles of regression

Regression can do 2 things:

Prediction/forecast

Relation between X and Y

Models

All models are wrong

– but some are useful

George Box

Y vs X

What goes where?

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

Y: dependent, response, outcome

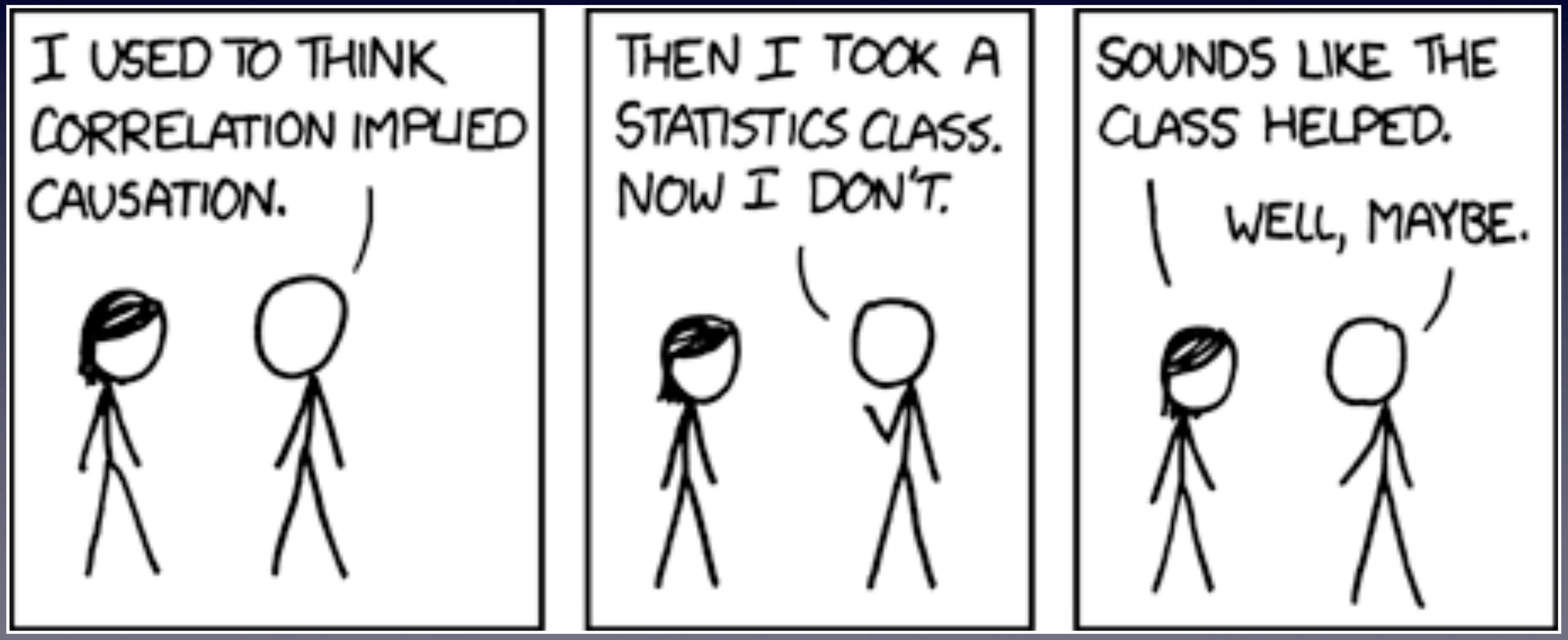
X: independent, explanatory, control, covariate

β_0 and β_1: parameters to be estimated (b_0,b_1)


Population
β_0 and β_1
Sample
b_0 and b_1

What is the question?

Role of theory



Some simple rules

Must make sense

If observed at different times: Y after X

Should not measure the same thing

Least squares

Aka Ordinary Least Squares or OLS

$$\hat{y} = b_0 + b_1 x$$

Linear model for which minimize sum of

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

Why this model?

Easy to use

Estimated line closest to the data (min squared residuals)

"Useful characteristics"

Sum of residuals = 0

"Best" estimator

(under some conditions)

Estimating

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

$$b_1 = \frac{s_{xy}}{s_x^2}$$

$$s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

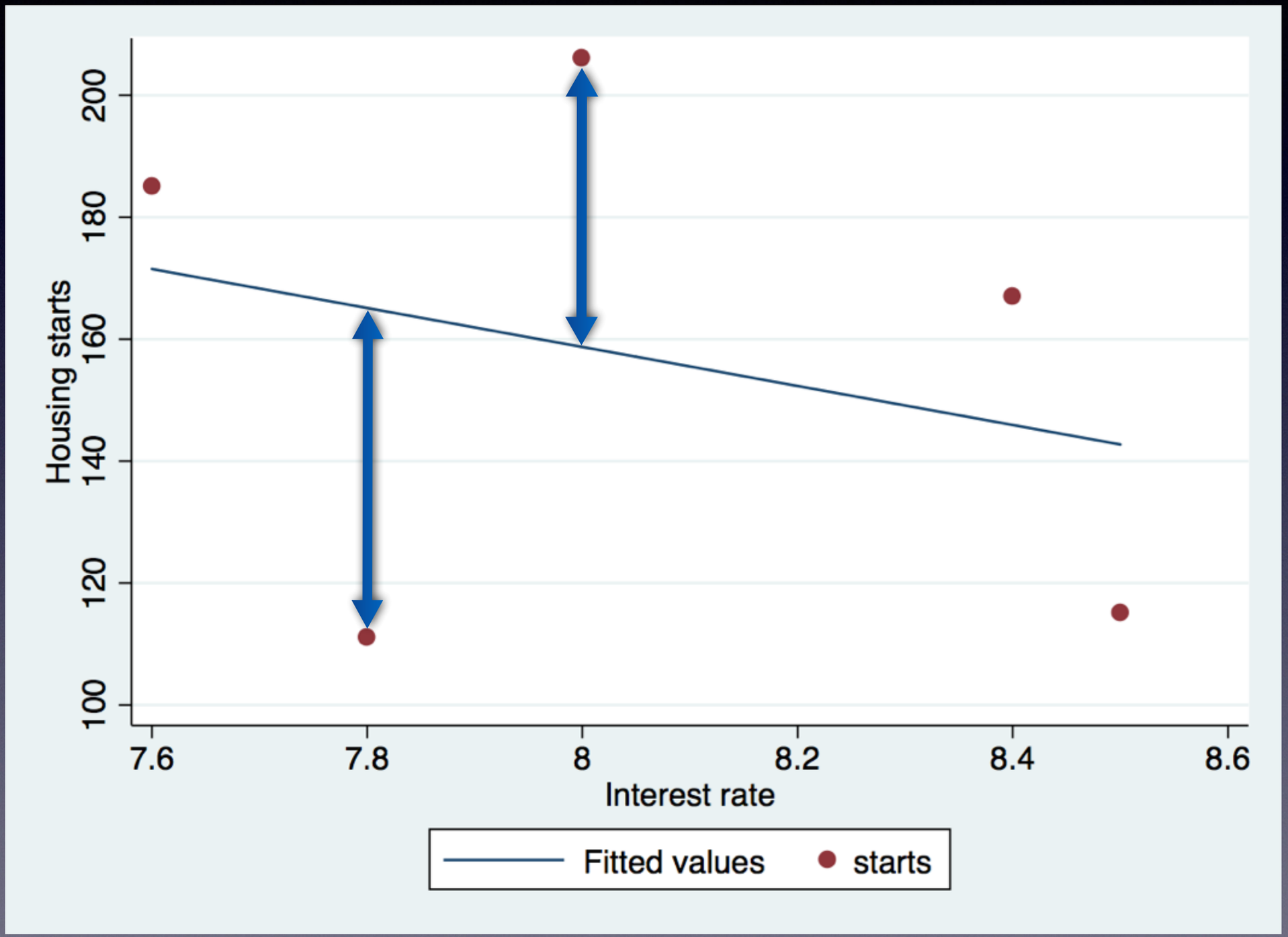
Starts (Y)	Rate (X)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
115	8.5	0.44	-41.8	-18.392	0.1936
111	7.8	-0.26	-45.8	11.908	0.0676
185	7.6	-0.46	28.2	-12.972	0.2116
206	8.0	-0.06	49.2	-2.952	0.0036
167	8.4	0.34	10.2	3.468	0.1156
Average	Average			Sum	Sum
156.8	8.1			-18.940	0.5920

s_{xy}	-18.94	/	(5-1)	=	-4.735
s_x^2	0.592	/	(5-1)	=	0.148
b_1	-4.735	/	0.148	=	-32.0
b_0	156.8	-	-257.9	=	414.7

R results for same problem

```
Call:
lm(formula = starts ~ rate, data = housing_starts)

Coefficients:
(Intercept)         rate
     414.67       -31.99
```

Residuals / errors



Example

Son's height = 33.73 + 0.516 x Father's height

a. Interpret the coefficient 0.516
b. What does the regression line tell you about the height of sons of tall fathers?
c. What does the regression line tell you about the height of sons of short fathers?
d. What can you learn from the intercept?

Why this effect?

"The child inherits partly from his parents, partly from his ancestry. Speaking generally, the further his genealogy goes back, the more numerous and varied will his ancestry become, until they cease to differ from any equally numerous sample taken at haphazard from the race at large. Their mean stature will then be the same as that of the race; in other words, it will be mediocre.

Regression towards Mediocrity in Hereditary Stature
Francis Galton

The role of average

A 1-inch increase in father's height increases the expected height of his son by, **on average**, 0.516 inches.

OLS estimates the **average** expected effect

Also:

$$\bar{y} = b_0 + b_1 \bar{x}$$

Your turn

```
modelName <- lm(yVar ~ xVar,
                data = dataFrame)

summary(modelName)
```

Use Alumni.csv

classesIt20: Percent classes with fewer than 20 students
sfratio: student/faculty ratio
alumnigivingrate: Percent alumni who donate to school

Assumptions

Assumptions about errors

$$y = \beta_0 + \beta_1 x + \varepsilon$$

1. Zero population mean
2. Not correlated with each other
3. Constant variance
4. Normally distributed

Draw sample

OLS to find b_0 & b_1

Population

β_0 and β_1

Sample

b_0 and b_1

Draw inference

This is where the 4 assumptions fit in!

Why these assumptions?

Best Linear Unbiased Estimator (BLUE)

Allows hypothesis testing!!

If not – look for other methods

Zero population mean

Error term for each observation
determined entirely by chance

What if mean is not zero?

That is why we have the intercept
No impact on b_1 if we have b_0

Uncorrelated errors

(Independence)

Why? Accurate estimates of standard errors

No impact on parameters

Constant variance

(Homoskedasticity)

Why? Accurate estimate of standard errors

No impact on parameters

Example: Variance of error increases with X

Normally distributed

Remember:
errors are already independent and zero mean

Why then?

Not required for OLS estimation

Needed for hypothesis testing:
Without it most small sample test are invalid

Distribution of Y given X

$$E(Y) = \beta_0 + \beta_1 X$$

For each X, expected/mean value of Y is E(Y)

Standard deviation is σ_ε

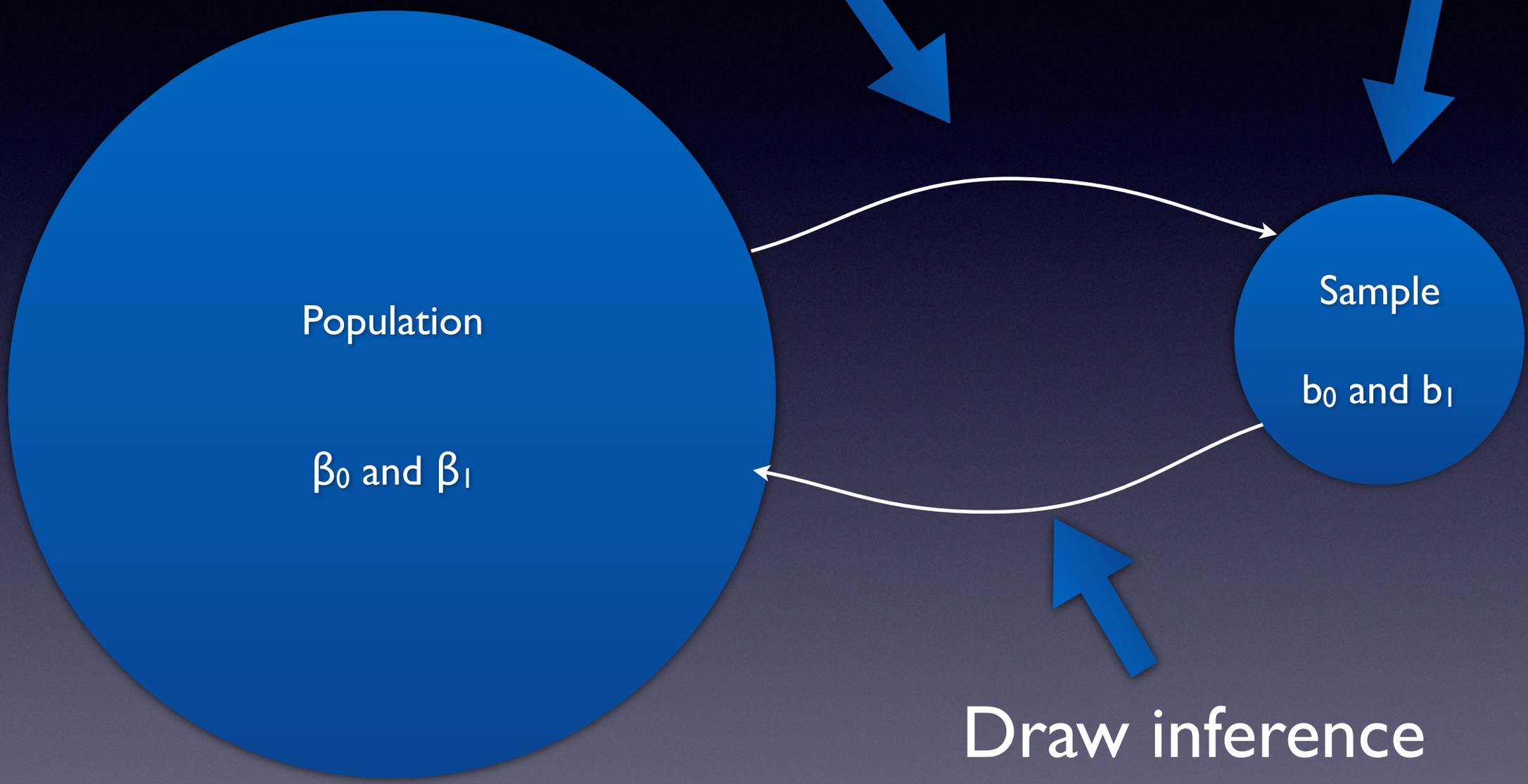
What affect b_0 and b_1?

- Model should be correctly specified, linear model with additive error term
- X must be uncorrelated with error term
- If more than one X, any 2 or more of them cannot be perfectly correlated

"correctly specified" and "X not correlated with ε"!

Draw sample

OLS to find b_0 & b_1



Draw inference

Places where we might run into problems

Question	Data	Problem
Effect of mother's education on child health	Individual with some siblings	Independence
Determinants of states' education spending	State level	Constant variance
What drives sale of our product	Weekly	Independence
Cost of patient care	Expenditure per patient	Normality Constant variance

Hypothesis testing

t test

Why do we test for the slope parameter in the simple linear regression being equal to zero?

The testing of the slope parameter in simple linear regression is conducted to know if variables have a linear relationship to each other. If they are not equal to zero, there is a relationship between the y and x variables.

Reasons cannot reject slope parameter is zero

No relationship between X and Y

Curved relationship between X and Y

Using the formula to figure out the last possibilities

t test components

$$t = \frac{b_1 - \beta_1}{s_{b_1}}$$

b_1 : parameter estimate

β_1 : null hypothesis (normally zero)

s_{b1} : standard error of b_1

Taking the t test apart

$$t = \frac{b_1 - \beta_1}{\frac{s_\varepsilon}{\sqrt{(n-1)s_x^2}}} = \frac{b_1 - \beta_1}{\frac{\sqrt{\frac{SSE}{(n-2)}}}{\sqrt{(n-1)s_x^2}}} = \frac{b_1 - \beta_1}{\frac{\sqrt{\frac{(n-1)}{(n-2)}\left(s_y^2 - \frac{s_{xy}^2}{s_x^2}\right)}}{\sqrt{(n-1)s_x^2}}}$$

Sample too small

Small sample variance in X (s_x^2)

Large sample variance in Y (s_y^2)

How to do the t test

How to do the t test

SSE (Sum of Squares for error)

=

sum of squared residuals

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

$b_1 = -31.99, b_0 = 414.67, s_x^2 = 0.148$

Starts (Y)	Rate (X)	$\hat{y}$	$(y_i - \hat{y})^2$	$(x_i - \bar{x})^2$
115	8.5	142.76	770.34	0.1936
111	7.8	165.15	2932.01	0.0676
185	7.6	171.55	181.01	0.2116
206	8.0	158.75	2232.56	0.0036
167	8.4	145.95	442.93	0.1156
Average	Average		Sum (SSE)	Sum
156.8	8.1		6558.85	0.5920

S_ε	Square root		(6558.85/(5-2))	=	46.8
s_{b1}	46.75771	/	sqrt((5-1)s_x^2)	=	60.8
t	-31.99	/	60.8	=	-0.5

Options for testing

Against critical value(s)

Find p-value

Confidence interval

Degrees of freedom = n - 2 = n - k - 1

k = number of explanatory variables

Options for testing

Against critical value(s)

p-value

Degrees of freedom = n - 2 = n - k - 1

k = number of explanatory variables

R output

```
> RegModel.1 <- lm(start~rate, data=housing)

> summary(RegModel.1)

Call:
lm(formula = start ~ rate, data = housing)

Residuals:
     1      2      3      4      5
-27.72 -54.12  13.48  47.28  21.08

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   414.67     490.26   0.846    0.460
rate          -31.99      60.77  -0.526    0.635

Residual standard error: 46.76 on 3 degrees of freedom
Multiple R-squared:  0.08457, Adjusted R-squared:  -0.2206
F-statistic: 0.2772 on 1 and 3 DF,  p-value: 0.635
```

Critical value

For 2 tailed test:

Look up $t_{\alpha/2}$ with n-2 degrees of freedom

For 1 tailed test:

Look up t_{α} with n-2 degrees of freedom

For $\alpha = 0.10$

$t_{\alpha}(3) = 1.638$

$t_{\alpha/2}(3) = 2.353$

$t = -0.53$

Confidence interval

$b_1 \pm t_{\alpha/2}\, s_{b1}$

Is zero in the interval?

$-31.99 \pm 2.353 \times 60.77$

P value

Only works for H_0: $\beta = 0$ (two-tailed test)!

Compare p-value against 3 levels: 0.1, 0.05, 0.01

If > 0.1: not statistically significant

If < 0.1: statistically significant at 10% level

If < 0.05: statistically significant at 5% level

If < 0.01: statistically significant at 1% level

R output

```
> RegModel.1 <- lm(start~rate, data=housing)

> summary(RegModel.1)

Call:
lm(formula = start ~ rate, data = housing)

Residuals:
     1      2      3      4      5
-27.72 -54.12  13.48  47.28  21.08

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   414.67     490.26   0.846    0.460
rate          -31.99      60.77  -0.526    0.635

Residual standard error: 46.76 on 3 degrees of freedom
Multiple R-squared:  0.08457, Adjusted R-squared:  -0.2206
F-statistic: 0.2772 on 1 and 3 DF,  p-value: 0.635
```

R^2 and model fit

What does R^2 do?

Measure how much variability in Y we explain

What role does R^2 play?

Forecasting

↓

Large is better

X's effect on Y

↓

Does not matter much

What if we had no X?

Best estimate of Y? Mean of Y

Total sum of squares (SST)

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

How far is mean Y from predicted Y?

Sum of squares due to regression (SSR)

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

Final part is SSE

SSE (Sum of Squares for error)

=

sum of squared residuals

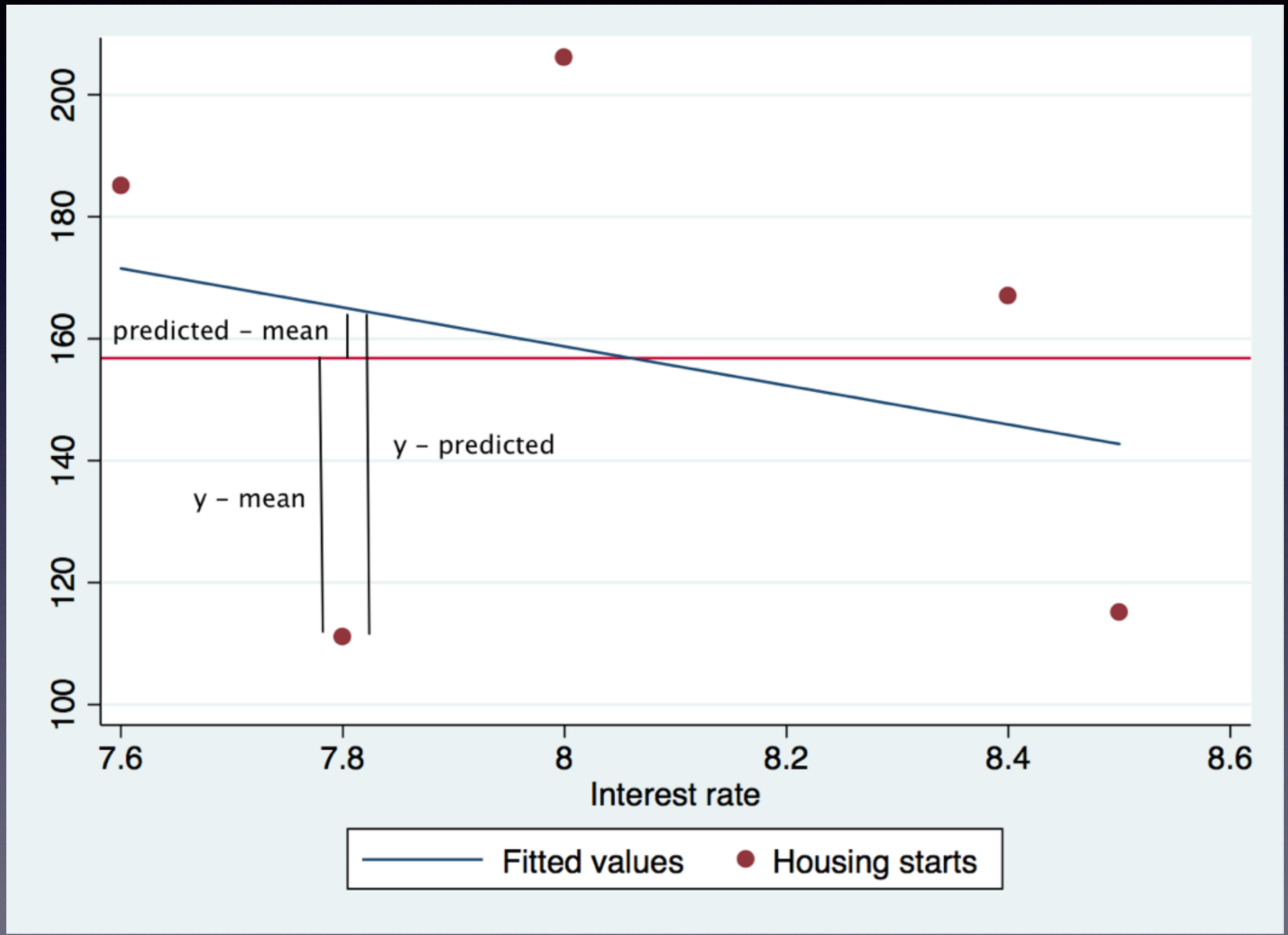
$$SSE = \sum_{i=1}^{n} (y_i - \hat{y})^2$$

Relationships

$$SST = SSR + SSE$$

$$R^2 = SSR/SST = 1 - (SSE/SST)$$

Graphical


200
180
160
140
120
100
predicted – mean
y – predicted
y – mean
7.6
7.8
8
8.2
8.4
8.6
Interest rate
Fitted values
Housing starts

Wording…

R^2 = 0.08547 in our housing starts example

8.5 percent of the variation in housing starts is explained by the variation in interest rate.

Remember: It is variation we are explaining and to use the variable names

Correlation vs Regression

Use housing start data
(from the data page on Canvas)

How to find correlation

R:

```
cov(data$var1,data$var2)
```

R tasks

- Load data
- Find standard deviation for starts and rate
- Find correlation between starts and rate
- Run regression analysis with Y: starts and X: rate

Correlation

$$r = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{X_i - X}{s_X} \right) \left(\frac{X_i - Y}{s_Y} \right)$$

r = -0.29

s_y = 42.32

s_x = 0.3847

Regression results

$b_0 = 414.67$ $b_1 = -31.99$

$r = b_1 s_x / s_y$

$r = -31.99 * (0.3847/42.32) = -0.29$

What if you do it in reverse?

Try to run the opposite regression in R

Y: rate

X: starts

Redo the calculation

Regress toward mean

Regression toward the mean will occur as long r<1(one)!!

Remember our example on height of father and sons

Symmetric:

A tall son will have a shorter father

A short son will have a taller father

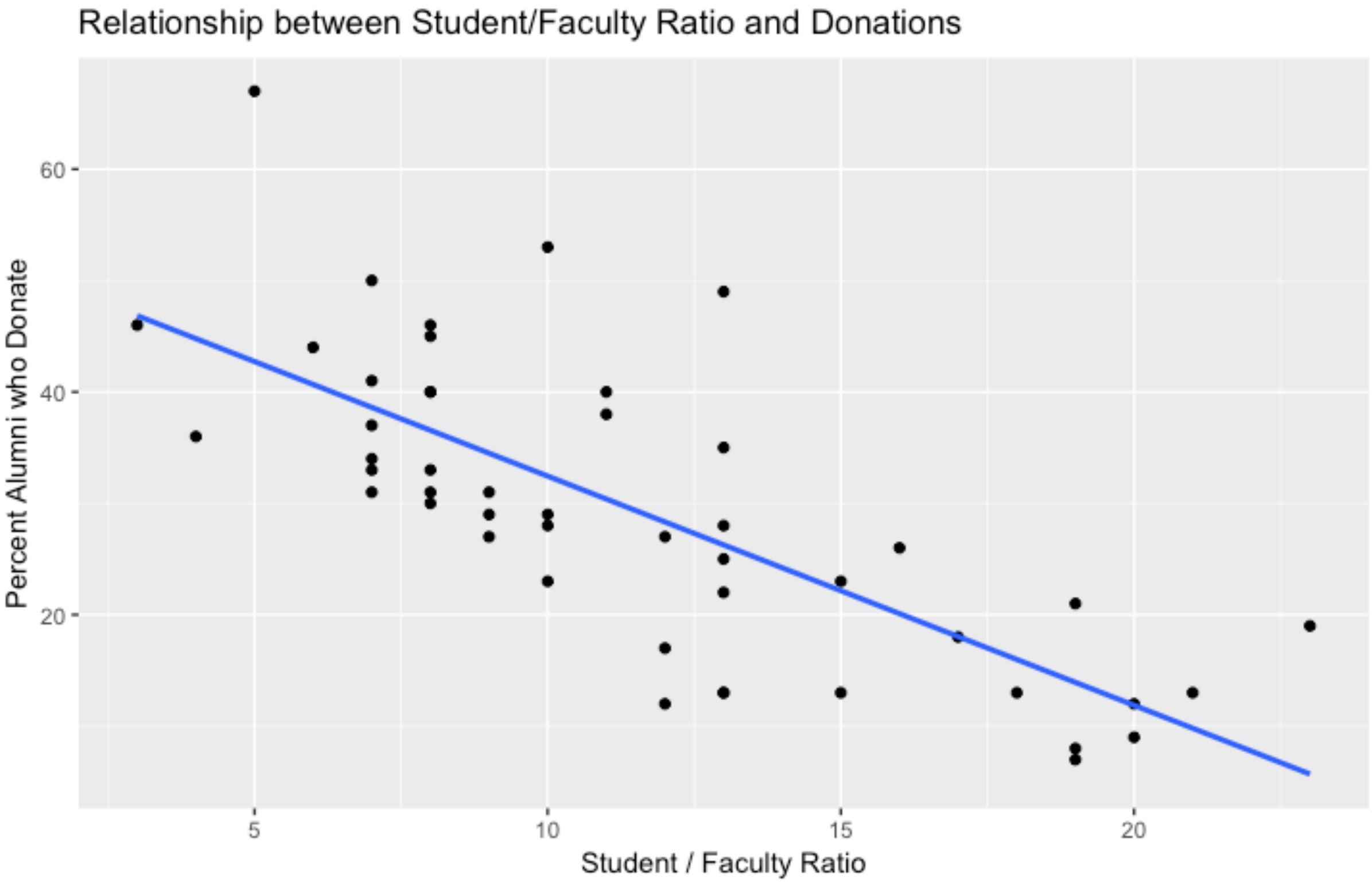
More on the role of average

A 1-inch increase in father's height increases the expected height of his son by, **on average**, 0.516 inches.

OLS estimates the **average** expected effect

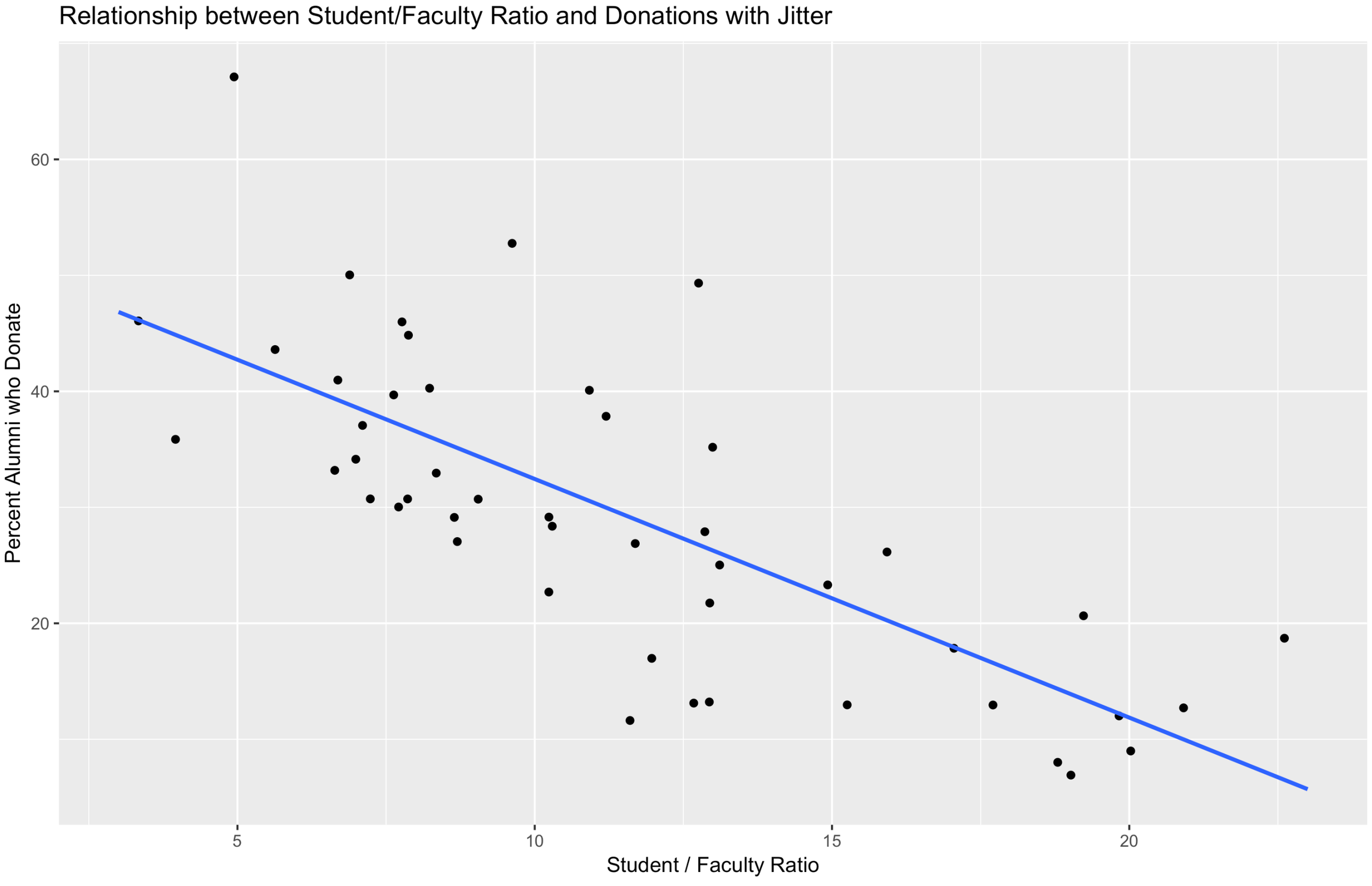
Plotting data in R

```
ggplot(alumni,
  aes(x = sfratio, y = alumnigivingrate)) +
  geom_point() +
  geom_smooth(method = 'lm', formula = y ~ x)+
  xlab("Student / Faculty Ratio") +
  ylab("Percent Alumni who Donate")+
  ggtitle("Relationship between Student/
Faculty Ratio and Donations")

ggsave(here("figures",
  "alumni_regression.png"))
```


Relationship between Student/Faculty Ratio and Donations
Percent Alumni who Donate
60
40
20
5
10
15
20
Student / Faculty Ratio

Plotting data in R

```
ggplot(alumni,
  aes(x = sfratio, y = alumnigivingrate)) +
  geom_jitter() +
  geom_smooth(method = 'lm', formula = y ~ x) +
  xlab("Student / Faculty Ratio") +
  ylab("Percent Alumni who Donate") +
  ggtitle("Relationship between Student/Faculty
Ratio and Donations")

ggsave(here("figures", "alumni_jitter.png"))
```


Relationship between Student/Faculty Ratio and Donations with Jitter
Percent Alumni who Donate
60
40
20
5
10
15
20
Student / Faculty Ratio

More graphing

png or jpeg

Both good for inserting in documents

Often need to play around with
graphs to get them to look good

For Next Time

- Introduction to Statistical Learning: Chapter 3.2 and 3.6.3
- Wooldridge: Chapters 3 and 4