Unit 4: Regression and Prediction

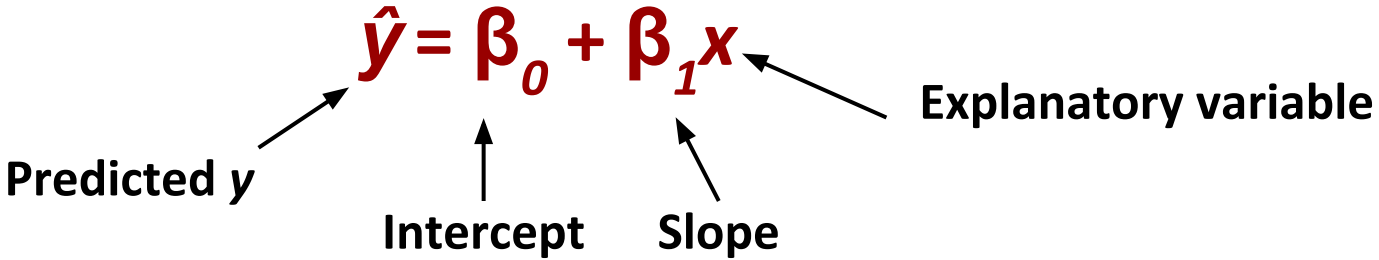
4. Multiple Regression

(Chapter 6.1)

11/28/2018

Recap from last time

1. A regression's slope codes the relationship between two measures
2. Correlation (R) is equivalent to slope (β_1) for standardized values
3. Inference for regression parameters uses t-tests



Predicting the weights of books

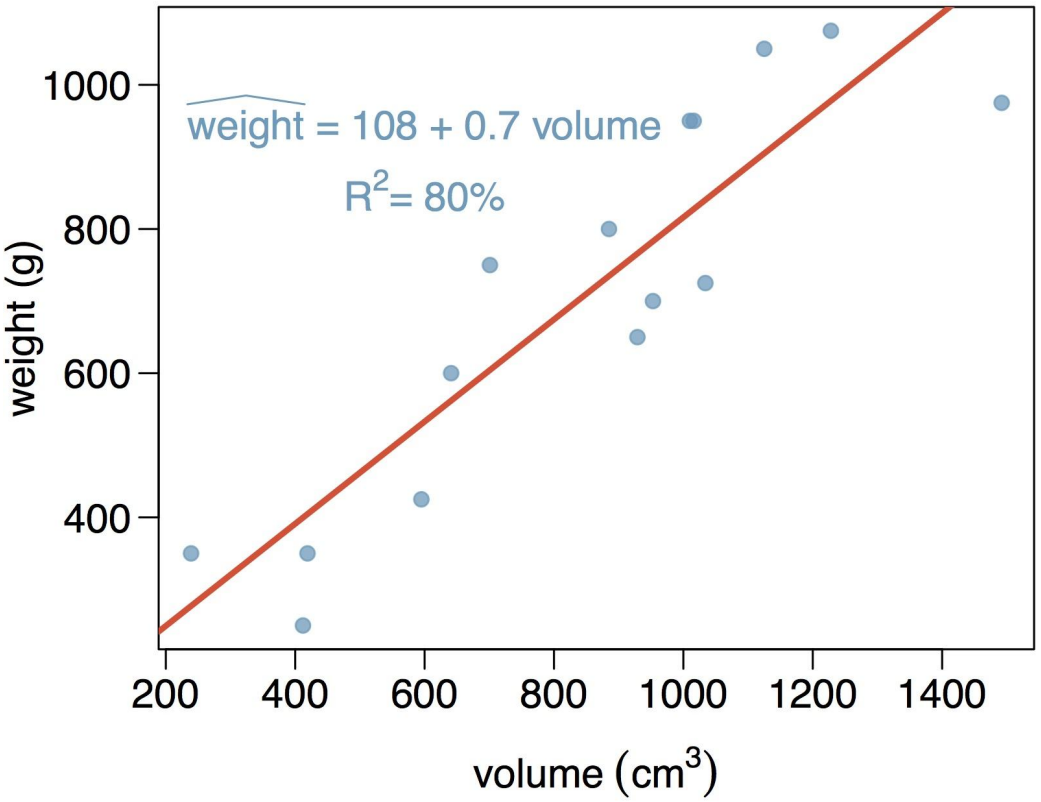
	weight (g)	volume (cm^3)	cover
1	800	885	hc
2	950	1016	hc
3	1050	1125	hc
4	350	239	hc
5	250	412	pb
6	700	953	pb
7	650	929	pb
8	975	1492	pb

From: Maindonald, J.H. and Braun, W.J. (2nd ed., 2007) “Data Analysis and Graphics Using R”

Modeling weight using volume

```
Coefficients:
             Estimate
(Intercept) 107.67931
volume        0.70864
```

$$\hat{y} = \beta_0 + \beta_1 x$$



Practice Question 1: Interpreting Regression models

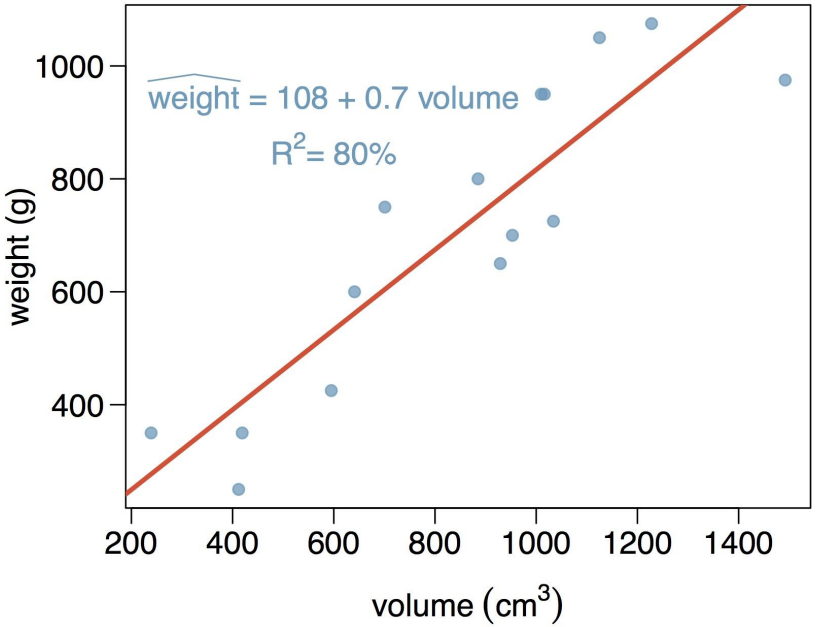
The scatterplot shows the relationship between *weights* and *volumes* of books as well as the regression output.
Which of the following is *wrong*?

(a) We can explain 80% of the variance of book weights if we know the book's volume

(b) According to our model, a book with 0 cm^3 volume has a weight of 0 grams

(c) The correlation (R) between book volume and book weight is close to 0.9

(d) We would estimate a 1000 cm^3 book to weigh 808 grams



Practice Question 1: Interpreting Regression models

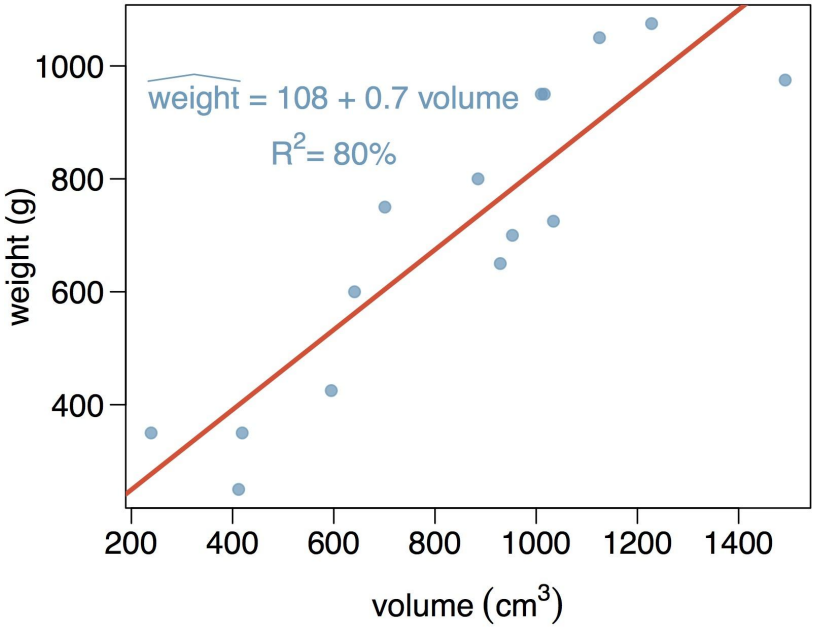
The scatterplot shows the relationship between *weights* and *volumes* of books as well as the regression output.
Which of the following is *wrong*?

(a) We can explain 80% of the variance of book weights if we know the book's volume

(b) According to our model, a book with 0 cm^3 volume has a weight of 0 grams

(c) The correlation (R) between book volume and book weight is close to 0.9

(d) We would estimate a 1000 cm^3 book to weigh 808 grams



Inference for linear regression

Inference for the slope for a single-predictor linear regression model:

Hypothesis test: $T = \frac{b_1 - null\ value}{SE_{b_1}}$ $\quad df = n - 2$

Confidence interval: $b_1 \pm t^{\star}_{df=n-2} SE_{b_1}$

The null value is often 0 since we are usually checking for **any** relationship between the explanatory and the response variable.

The regression output gives b_1, SE_{b1}, and **two-tailed** p-value for the t-test for the slope where the null value is 0.

We rarely do inference on the intercept, so we'll focusing on the slope.

Hypothesis Test Output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06
```

$$\hat{y} = \beta_0 + \beta_1 x$$

$$T = \frac{b_1 - \text{null value}}{SE_{b_1}} \qquad df = n - 2$$

Key ideas

1. In multiple regression, every variable is conditional on every other variable
2. For inference, we care about both the whole model and the individual variables
3. We use adjusted R^2 to account to penalize additional variables

Intro to multiple regression

So far:

- Simple linear regression: Ask if **y** is predicted by **x**

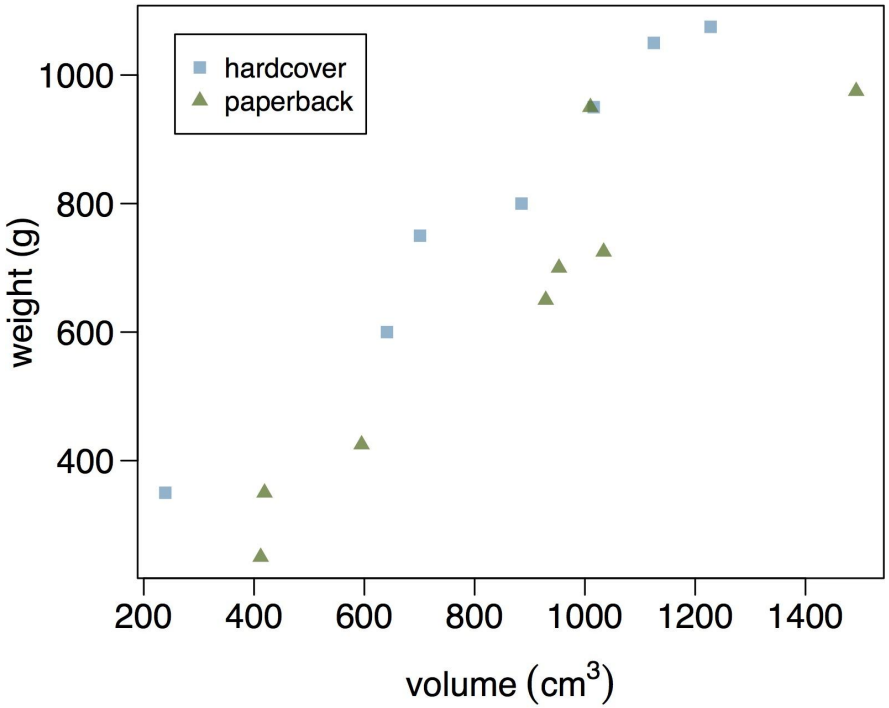
$$\hat{y} = \beta_0 + \beta_1 x_1$$

Now:

- Multiple linear regression: Ask if **y** is predicted by a combination of many variables x_1, x_2, x_3...

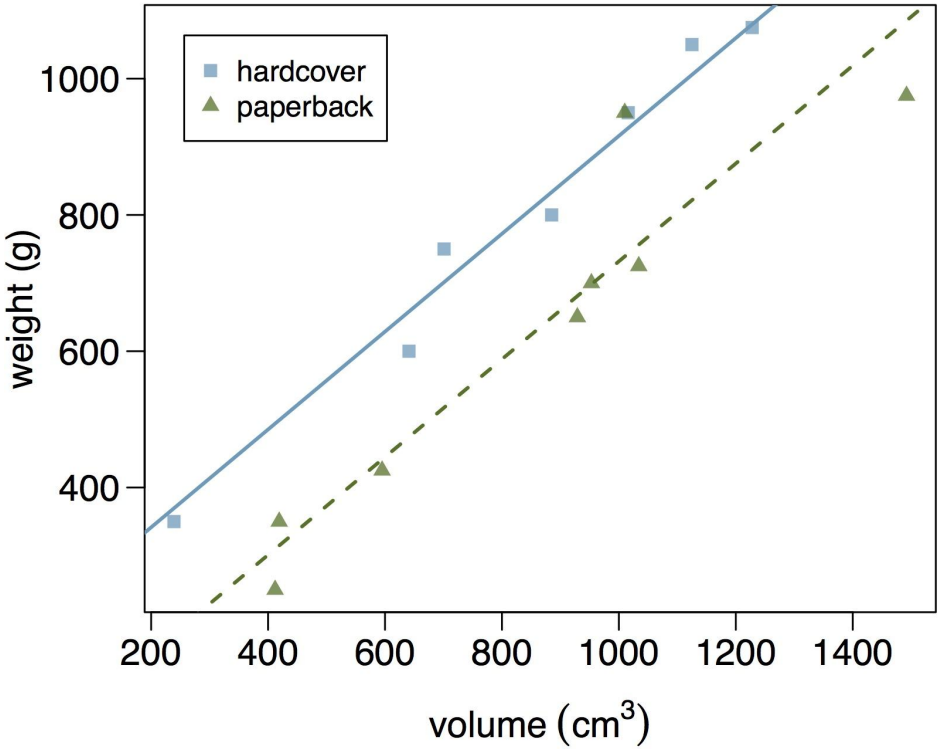
$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

What about cover type?



Paperbacks tend to weigh less than hardcovers ***controlling for volume***.

Two different effects



Modeling weight using volume and cover type

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***
```

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Practice Question 2: Understanding the regression equation

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***
```

Which of these correctly describes the role of the variables in this model?

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) response: weight, explanatory: volume, cover type

Practice Question 2: Understanding the regression equation

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  197.96284   59.19274   3.344 0.005841 **
volume         0.71795    0.06153  11.669  6.6e-08 ***
cover:pb    -184.04727   40.49420  -4.545 0.000672 ***
```

Which of these correctly describes the role of the variables in this model?

(a) response: weight, explanatory: volume, paperback cover

(b) response: weight, explanatory: volume, hardcover cover

(c) response: volume, explanatory: weight, cover type

(d) response: weight, explanatory: volume, cover type

The Linear Model

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197.96284	59.19274	3.344	0.005841	**
volume	0.71795	0.06153	11.669	6.6e-08	***
cover:pb	-184.04727	40.49420	-4.545	0.000672	***

$$\widehat{weight} = 197.96 + .72volume - 184.05cover{:}pb$$

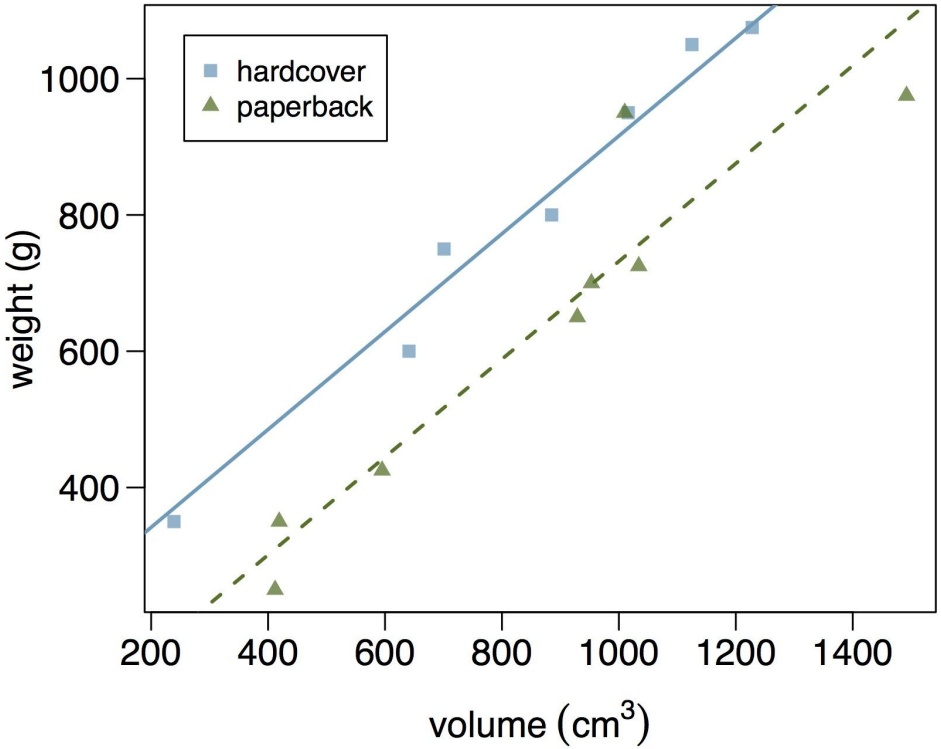
For **hardcover** books: plug in **0** for cover (reference level)

$$\widehat{weight} = 197.96 + .72volume - 184.05\times\mathbf{0}$$

For **paperback** books: plug in **1** for cover

$$\widehat{weight} = 197.96 + .72volume - 184.05\times\mathbf{1}$$

Two different effects



Interpreting the coefficients

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	197.96284	59.19274	3.344	0.005841	**
volume	0.71795	0.06153	11.669	6.6e-08	***
cover:pb	-184.04727	40.49420	-4.545	0.000672	***

- **Slope of volume**: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about **0.72** grams more.
- **Slope of cover**: All else held constant, paperback books weigh **184** grams lower than hardcover books.
- **Intercept**: Hardcover books with no volume are expected on average to weigh **198** grams. *Does this make sense?*

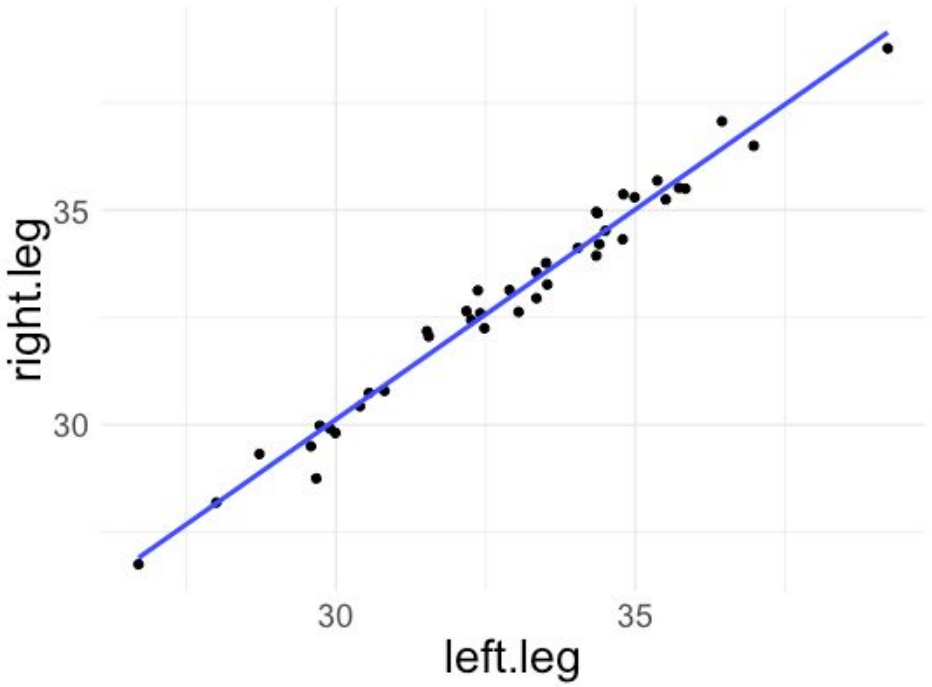
Including additional parameters

So should we just add parameters for everything we could possibly measure?

- **No**, the more complex a model is the harder it is to reason about
- There are dangers in interpreting parameters that are highly correlated (**collinearity**)
- We prefer the simplest (**parsimonious**) model

A collinearity example

	height (in)	Left leg (in)	Right leg (in)
1	63.2	32.26	32.44
2	62.3	28.00	28.19
3	70.4	36.45	37.07
4	66.2	33.35	32.95
5	65.0	32.37	33.13
6	71.7	39.22	38.77
7	71.7	35.51	35.25
8	62.4	28.72	29.32



Example from: *Statistical Rethinking,* Richard McElreath

Inference for parameters

Heights predicted with **just** left leg:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.9961	4.6578	7.299	9.73e-09 ***
left.leg	0.9771	0.1413	6.914	**3.22e-08 *** **

Heights predicted with **both** left and right leg:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	32.8617	4.6401	7.082	2.2e-08 ***
left.leg	-0.4241	0.9325	-0.455	**0.652**
right.leg	1.4329	0.9429	1.520	**0.137**

Example from: *Statistical Rethinking,* Richard McElreath

Inference for an entire model

Heights predicted with **just** left leg:

Residual standard error: 2.341 on 38 degrees of freedom
Multiple R-squared: **0.5571**, Adjusted R-squared: 0.5455
F-statistic: 47.8 on 1 and 38 DF, p-value: **3.218e-08**

Heights predicted with **both** left and right leg:

Residual standard error: 2.301 on 37 degrees of freedom
Multiple R-squared: **0.5831**, Adjusted R-squared: 0.5606
F-statistic: 25.88 on 2 and 37 DF, p-value: **9.334e-08**

Example from: *Statistical Rethinking,* Richard McElreath

Parameter vs model inference

- **Model test**: Is my model reliably useful?
  - Can I predict your height if I know the length of your legs?

- **Parameter test**: How much more do I learn by including a variable?
  - How much more do I gain by knowing the length of your right leg if I already know the length of your left leg?

Practice Question 3: Understanding Regression Output

Here is some R output of a model I ran on the nycflights data in an attempt to predict arrival delay:

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.0336374  0.1746381  -17.37   <2e-16 ***
dep_delay    1.0124888  0.0024376  415.36   <2e-16 ***
distance    -0.0026087  0.0001343  -19.43   <2e-16 ***

Residual standard error: 17.82 on 32732 degrees of freedom
Multiple R-squared:  0.841,   Adjusted R-squared:  0.841
F-statistic: 8.657e+04 on 2 and 32732 DF,  p-value: < 2.2e-16
```

Is this model useful?

If I know dep_delay, is it still useful to know distance?

Another look at R^2

R^2 can be calculated in two ways:

1. Squaring the correlation coefficient of standardized *x* and *y* (*R*)
2. Based on the definition:

$$R^2 = \frac{\text{explained variability in } y}{\text{total variability in } y}$$

Why bother with a second calculation?

For a simple linear regression, you don't need to worry about it.

But for multiple regression, you can't compute the correlation between *y* and *x* because you have multiple *xs*.

And also, we want to use this second method to compute **adjusted R^2**

R^2 vs. adjusted R^2

- When **any** variable is added to a model, R^2 increases.
- But if the added variable doesn't really provide any new information, or is completely unrelated, adjusted R^2 does not increase.

Adjusted R^2 example

Heights predicted with **just** left leg:

Residual standard error: 2.341 on 38 degrees of freedom
Multiple R-squared: **0.5571**, Adjusted R-squared: **0.5455**
F-statistic: 47.8 on 1 and 38 DF, p-value: 3.218e-08

Adjusted R^2 example

Heights predicted with **everything** I could have measured:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	34.46065	5.76696	5.976	1.04e-06 ***
left.leg	-0.58290	1.02812	-0.567	0.575
right.leg	1.61296	1.05255	1.532	0.135
gender	-0.63707	0.83762	-0.761	0.452
IQ	-0.01879	0.02585	-0.727	0.472
native.english	0.07218	1.03086	0.070	0.945
fun.at.parties	-0.29258	1.04315	-0.280	0.781

Residual standard error: 2.386 on 33 degrees of freedom
Multiple R-squared: **0.6004**, Adjusted R-squared: **0.5278**
F-statistic: 8.265 on 6 and 33 DF, p-value: 1.68e-05

Key ideas

1. In multiple regression, every variable is conditional on every other variable
2. For inference, we care about both the whole model and the individual variables
3. We use adjusted R^2 to account to penalize additional variables