Unit 5: The General Linear Model

1. Multiple Regression
(Chapter 6.1)

3/30/2020

Recap from last time

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

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**

Now:

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

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

Practice Question 1: Interpreting Regression models

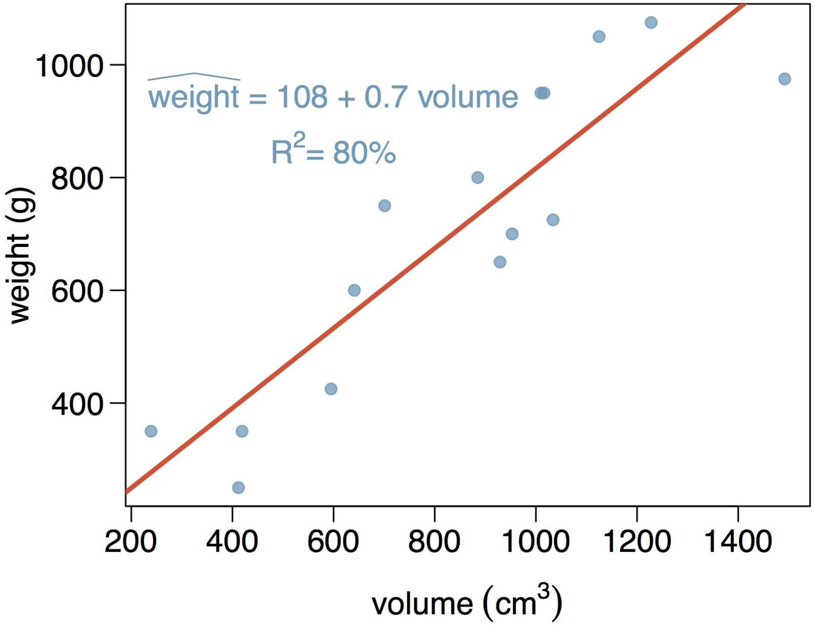
The scatterplot shows the relationship between weights and volumes of books as well as the regression output. Which of the following is correct?

(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10cm^3 over average are expected to weigh 7g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

(d) The model underestimates the weight of the book with the highest volume.



Practice Question 1: Interpreting Regression models

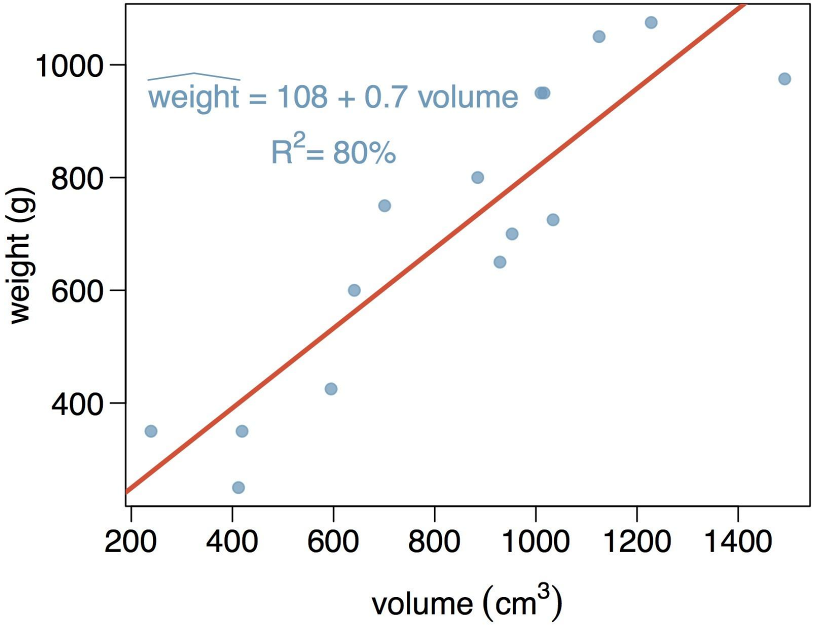
The scatterplot shows the relationship between weights and volumes of books as well as the regression output.
Which of the following is correct?

(a) Weights of 80% of the books can be predicted accurately using this model.

(b) Books that are 10cm^3 over average are expected to weigh 7g over average.

(c) The correlation between weight and volume is $R = 0.80^2 = 0.64$.

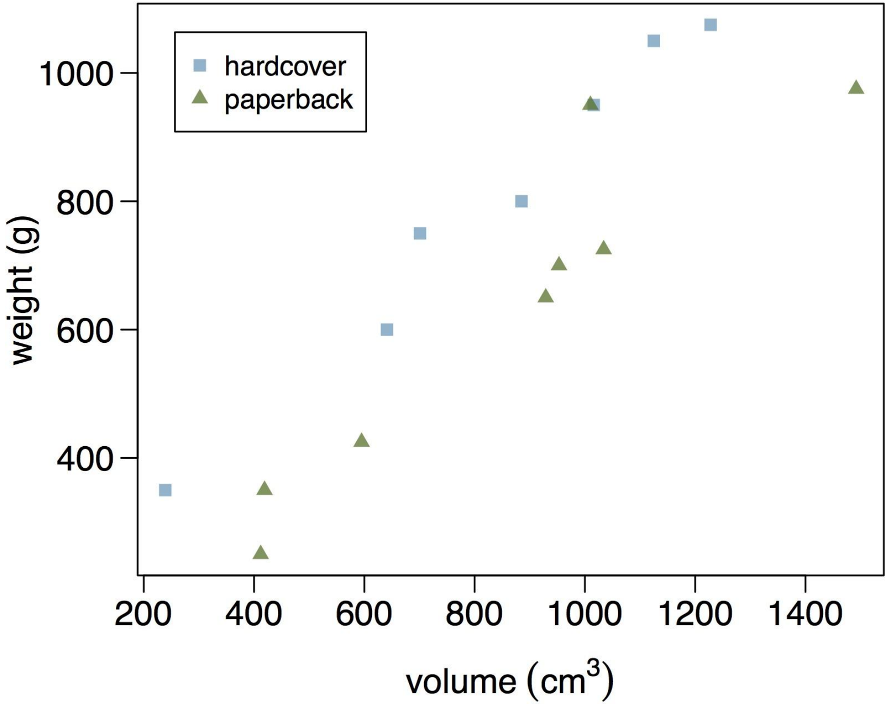
(d) The model underestimates the weight of the book with the highest volume.



Modeling weight using volume

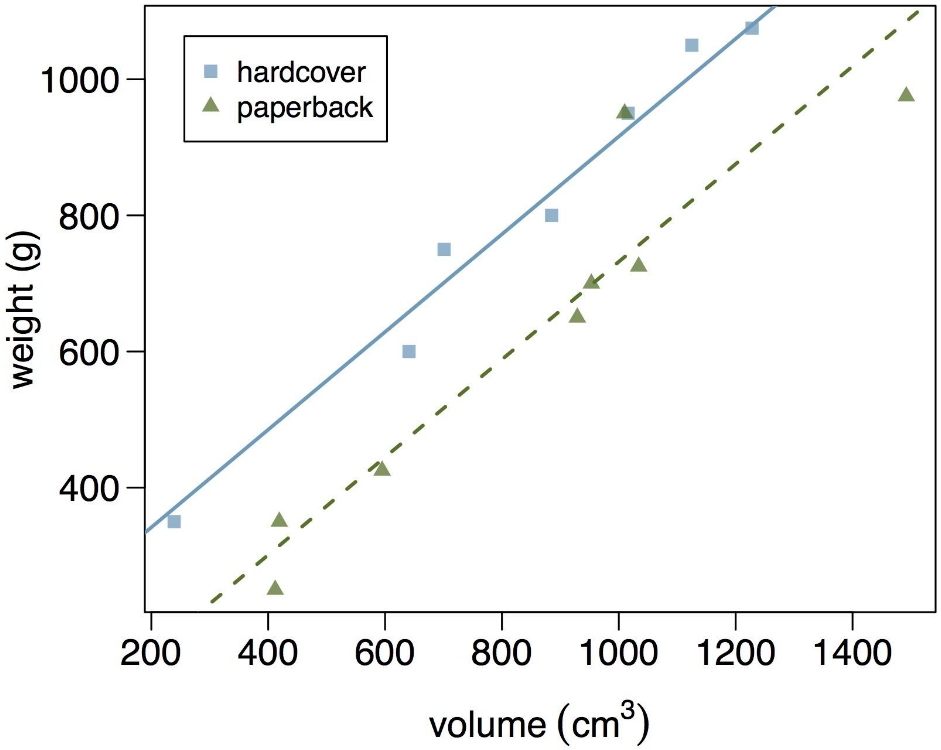
```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 107.67931   88.37758   1.218    0.245
volume        0.70864    0.09746   7.271 6.26e-06


Residual standard error: 123.9 on 13 degrees of freedom

Multiple R-squared: 0.8026,Adjusted R-squared: 0.7875
F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

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 ***

Residual standard error: 78.2 on 12 degrees of freedom
Multiple R-squared: 0.9275,Adjusted R-squared: 0.9154
F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```

What is the **reference level** for cover type?

hardcover

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

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

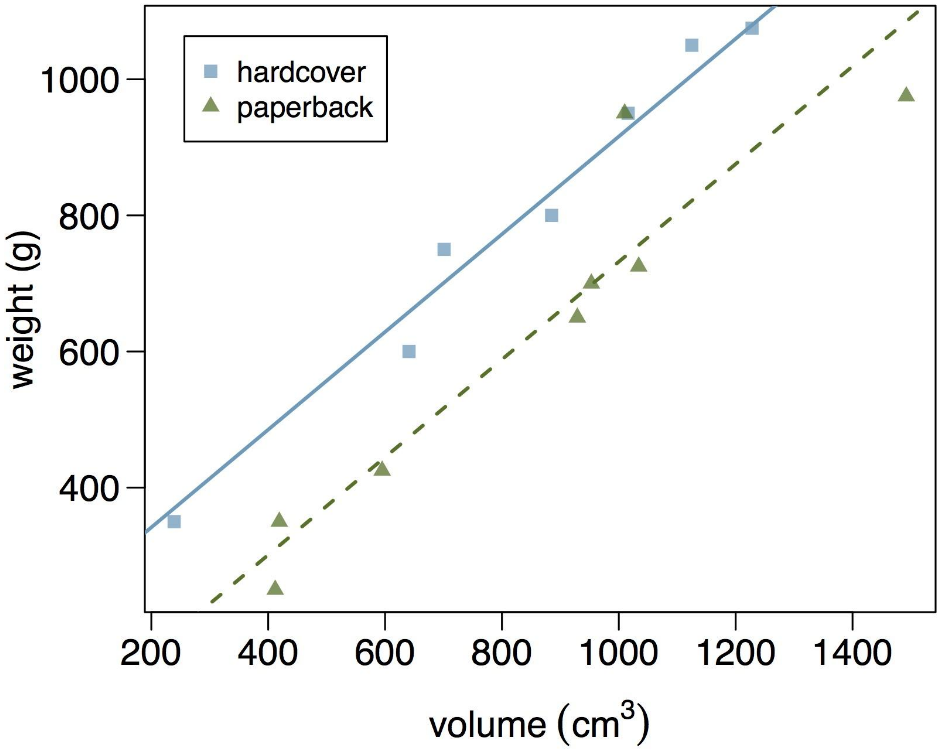
$$\widehat{weight} = 197.96 + .72volume$$

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

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

$$\widehat{weight} = 13.91 + .72volume$$

Visualizing the model

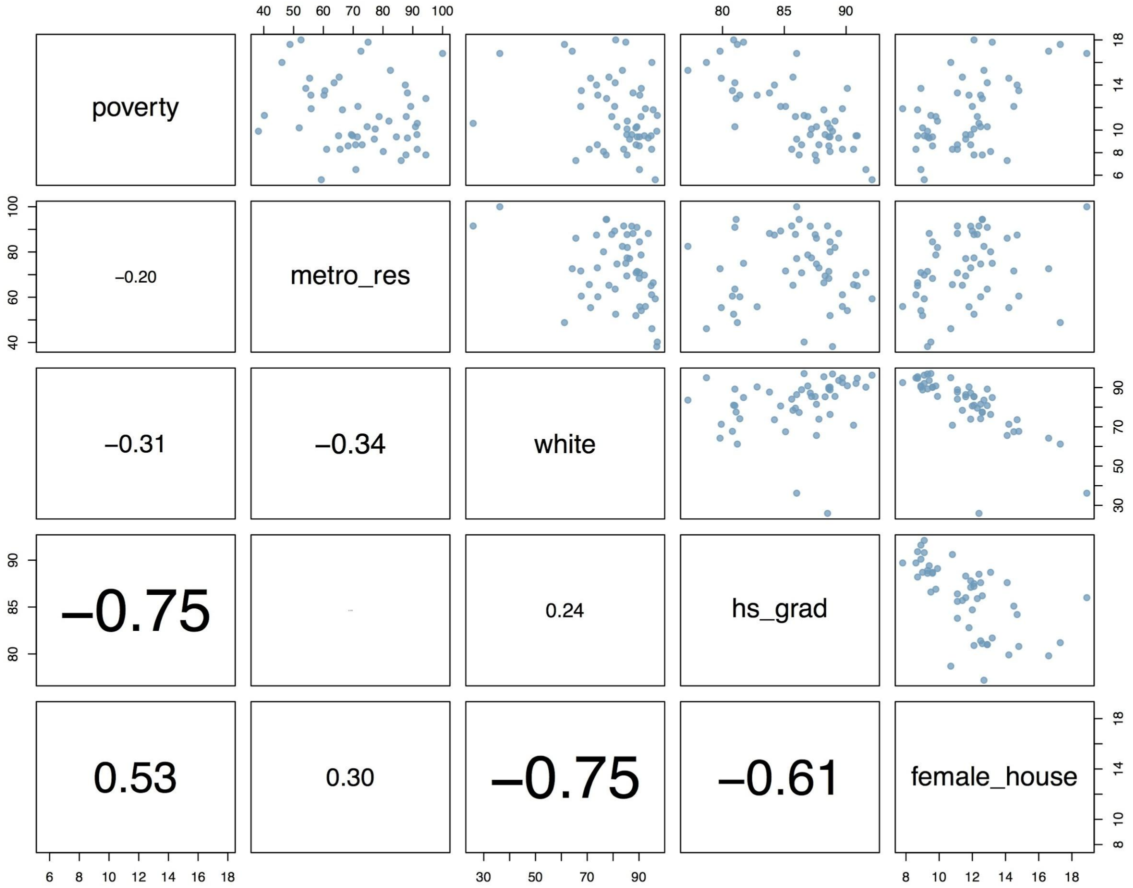


Interpreting the coefficients

```
               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 ***
```

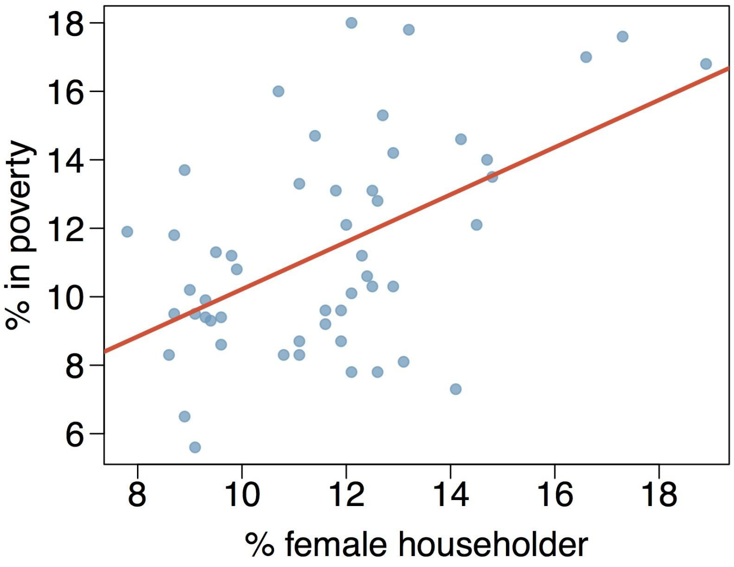
- **Slope of volume**: All else held constant, books that are 1 more cubic centimeter in volume tend to weigh about 0.72 grams more.
- **Slope of cover**: All else held constant, paperback books weigh 184 grams lower than hardcover books.
- **Intercept**: Hardcover books with no volume are expected on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

Revisiting poverty



Predicting poverty using % female householder

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00



$R = .53$

$R^2 = .53^2 = .28$

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:

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

This lets us use **ANOVA** to calculate the explained variability and total variability.

We're going to skip the details of this, just worry about understanding it conceptually

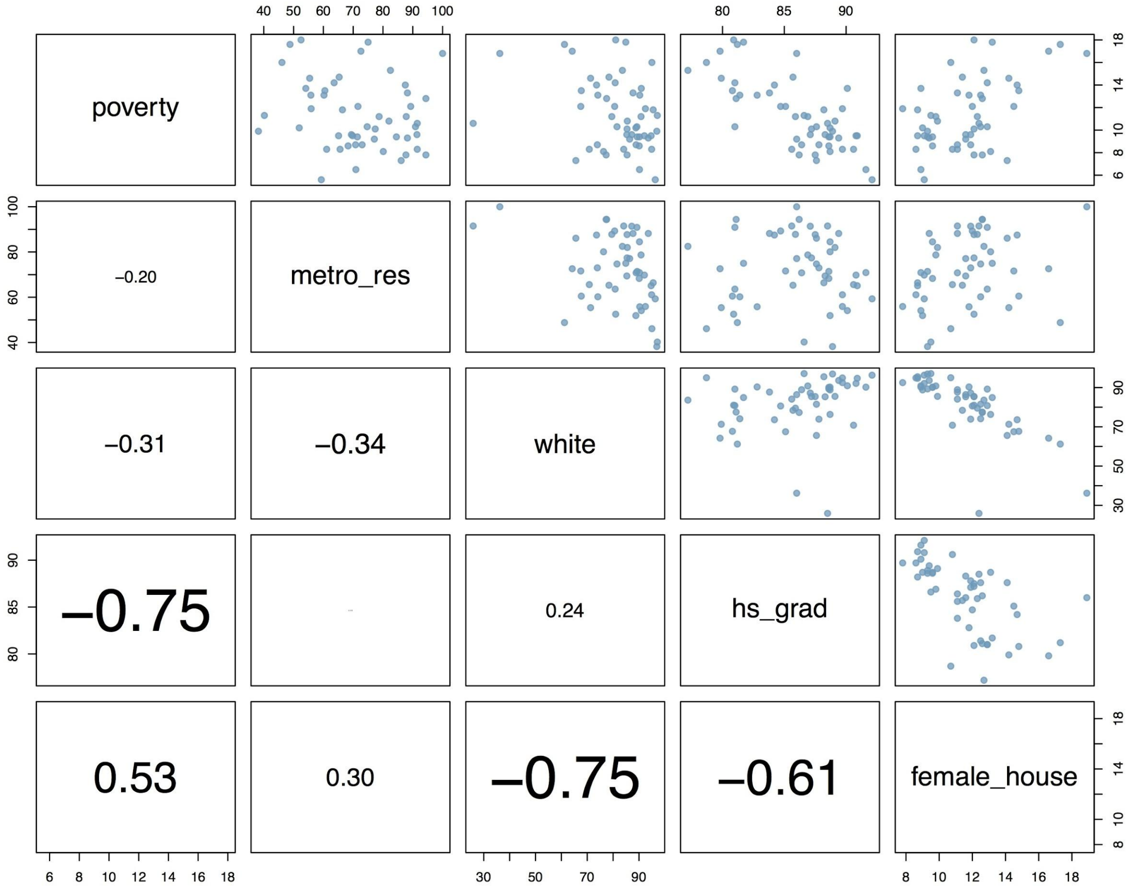
Why bother with a second calculation?

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 *x*s.

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

Does adding white to the model add any extra information?



Both are independently correlated, with income…

poverty vs. %female head of household

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.31	1.90	1.74	0.09
female_house	0.69	0.16	4.32	0.00

poverty vs. %female head of household and white

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.58	5.78	-0.45	0.66
female_house	0.89	0.24	3.67	0.00
white	0.04	0.04	1.08	0.29

Collinearity between explanatory variables

- Two predictor variables are said to be collinear when they are highly correlated, and this **collinearity** complicates model estimation.

- We don't like adding predictors that are associated with each other to the model, because often adding these variable brings nothing to the table. Instead, we prefer the simplest (**parsimonious**) model.

- While it's impossible to avoid collinearity from arising in observational data, experiments are usually designed to prevent correlation among predictors.

R^2 vs. adjusted R^2

	R^2	Adjusted R^2
Model 1 (Single-predictor)	0.28	0.26
Model 2 (Multiple)	0.29	0.26

- 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.

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