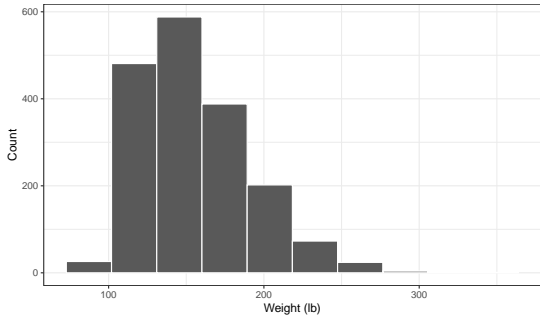Practice fitting, evaluating and making point estimates using multiple linear regression models (ch 6)

Justin Leinaweaver (Spring 2022)

What is the most "useful" model of weight in the Ross (1990) dataset?

Source: Ross (1990)

Predictors to Explore
- Height
- Exercise
- Age

For Today

Evaluate our four models using Wheelan ch 12

- Model 1: Height
- Model 2: Height and Exercise
- Model 3: Height and Age
- Model 4: Height, Exercise and Age

Common Regression Mistakes (Wheelan ch12)

1. Linear regression on nonlinear relationships
2. Correlation does not equal causation
3. Reverse causality
4. Omitted variable bias (too few variables)
5. Highly correlated explanatory variables (multicollinearity)
6. Extrapolating beyond the data
7. Data mining (too many variables)

Common Regression Mistakes (Wheelan ch12)

1. Nonlinearity: Residuals Plots
2. Correlation does not equal causation
3. Reverse causality
4. Omitted variable bias (too few variables)
5. Multicollinearity: Correlation Matrix
6. Outside Data: Scatterplot
7. Data mining (too many variables)

Common Regression Mistakes (Wheelan ch12)

Check the Data

1. Nonlinearity: Residuals Plots
2. Multicollinearity: Correlation Matrix
3. Outside Data: Scatterplot

Make an Argument

1. Correlation does not equal causation
2. Reverse causality
3. Omitted variable bias (too few variables)
4. Data mining (too many variables)

For Today

Evaluate our four models using Wheelan ch 12

- Model 1: Height
- Model 2: Height and Exercise
- Model 3: Height and Age
- Model 4: Height, Exercise and Age

	Weight (lb)			
	(1)	(2)	(3)	(4)
height	4.95*	5.21*	5.15*	5.29*
	(0.18)	(0.18)	(0.18)	(0.18)
exercise		−2.05*		−1.40*
		(0.30)		(0.31)
age			0.33*	0.27*
			(0.04)	(0.04)
Constant	−173.26*	−184.54*	−200.94*	−203.76*
	(11.91)	(11.88)	(12.14)	(12.09)
Observations	1,788	1,788	1,788	1,788
Adjusted R^2	0.30	0.32	0.33	0.33
Residual Std. Error	28.96 (df = 1786)	28.59 (df = 1785)	28.41 (df = 1785)	28.25 (df = 1784)
F Statistic	767.70* (df = 1; 1786)	417.44* (df = 2; 1785)	434.54* (df = 2; 1785)	299.65* (df = 3; 1784)

Note: *p < 0.05

One approach to building a "best" multiple regression model

1. Choose the logical predictors

2. Fit a simple OLS regression to each predictor

3. Fit a multiple regression with the "best" of those

4. Evaluate the model using all five steps

5. Use the model to make predictions

What is the "best" multiple regression model of earnings in the Ross (1990) dataset?

Outcome
Yearly income in thousands of dollars (earnk2021)

Predictors to Consider
Age, angry, education, ethnicity, exercise, father's education, height, male, mother's education, smoker, tense, walk and weight

What is the "best" multiple regression model of earnings in the Ross (1990) dataset?

Outcome
Yearly income in thousands of dollars (earnk2021)

Predictors to Consider
Age, ~~angry~~, **education**, ~~ethnicity~~, **exercise**, ~~father's education~~, **height**, ~~male~~, **mother's education**, ~~smoker~~, ~~tense, walk and weight~~

		Earnings (2021)		
	(1)	(2)	(3)	(4)	(5)
---	---	---	---	---	---
age	0.22* (0.06)				
education		5.57* (0.36)			
exercise			1.88* (0.42)		
height				3.08* (0.25)	
mother_education					6.21* (0.43)
Constant	33.17* (2.64)	−31.34* (4.89)	36.69* (1.62)	−162.54* (16.46)	−40.64* (5.87)
Observations	1,815	1,813	1,815	1,815	1,570
Adjusted R^2	0.01	0.11	0.01	0.08	0.12
Residual Std. Error	41.81 (df = 1813)	39.50 (df = 1811)	41.75 (df = 1813)	40.28 (df = 1813)	40.45 (df = 1568)

Note: *p < 0.05

	Earnings (2021 USD)
Age	0.47^*
	(0.05)
Education	5.42^*
	(0.35)
Exercise	0.95^*
	(0.41)
Height	2.84^*
	(0.24)
Mother's Education	-241.27^*
	(16.34)
Observations	1,813
Adjusted R^2	0.21
Residual Std. Error	37.43 (df = 1808)
F Statistic	118.04^* (df = 4; 1808)
Note:	$^*p < 0.05$
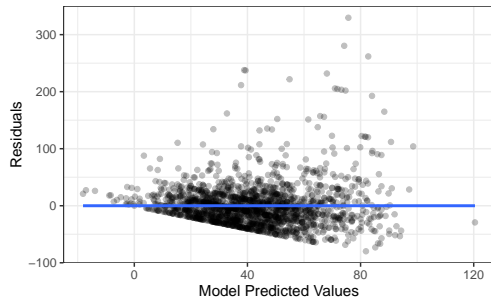

```
=========================================
          age  education  exercise  height
-----------------------------------------
age        1     -0.15      -0.33    -0.14
education -0.15    1         0.18     0.11
exercise  -0.33   0.18       1        0.22
height    -0.14   0.11       0.22     1
-----------------------------------------
```

| | Earnings (2021) | | | | |
	(1)	(2)	(3)	(4)	(5)
Age	0.22*				0.47*
	(0.06)				(0.05)
Education		5.57*			5.42*
		(0.36)			(0.35)
Exercise			1.88*		0.95*
			(0.42)		(0.41)
Height				3.08*	2.84*
				(0.25)	(0.24)
Constant	33.17*	−31.34*	36.69*	−162.54*	−241.27*
	(2.64)	(4.89)	(1.62)	(16.46)	(16.34)
Observations	1,815	1,813	1,815	1,815	1,813
Adjusted R^2	0.01	0.11	0.01	0.08	0.21
Residual Std. Error	41.81 (df = 1813)	39.50 (df = 1811)	41.75 (df = 1813)	40.28 (df = 1813)	37.43 (df = 1808)

Note: $^*p < 0.05$

For Tuesday - Make Predictions (w/ 95% PIs)

Establish a baseline prediction for a hypothetical person who is 25 (Age), Education 13, Exercise 1, Height 67

Then calculate the predicted income if:

1. Finish college? (16)

2. Finish grad school? (18)

3. Increase exercise to 1x / week? (4)

4. Increase exercise to 3x / week? (7)