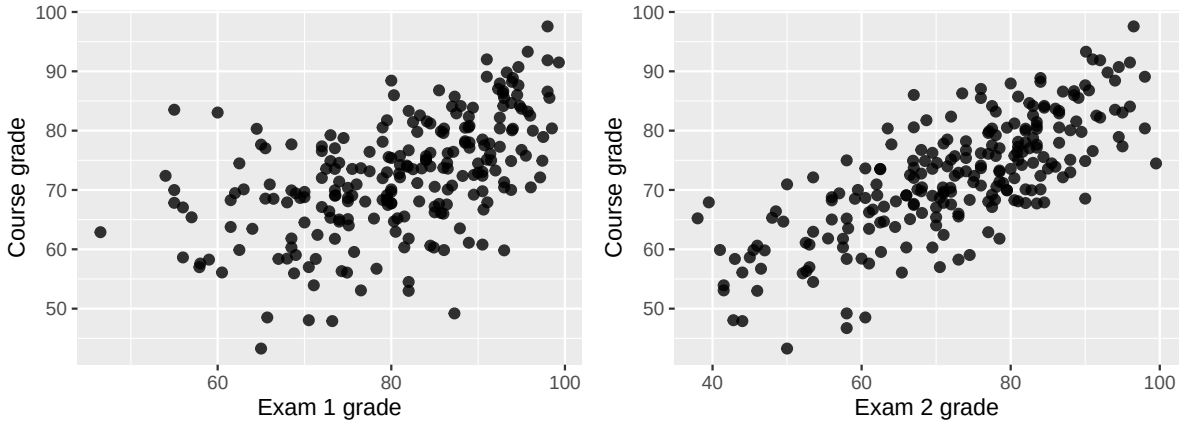
Problem Set 12

Simple Linear Regression

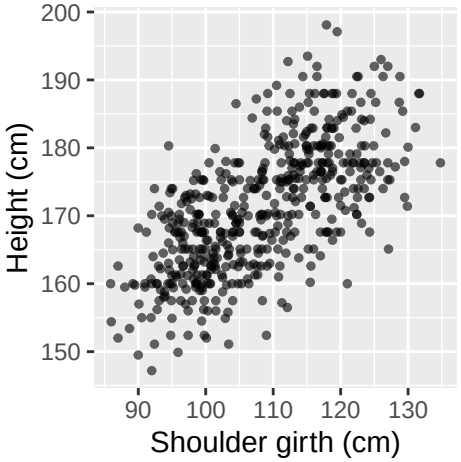
1. The two scatterplots below show the relationship between the overall course average and two midterm exams (Exam 1 and Exam 2) recorded for 233 students during several years for a statistics course at a university.



   a. Based on these graphs, which of the two exams has the strongest correlation with the course grade? Explain.

   b. Can you think of a reason why the correlation between the exam you chose in part (a) and the course grade is higher?

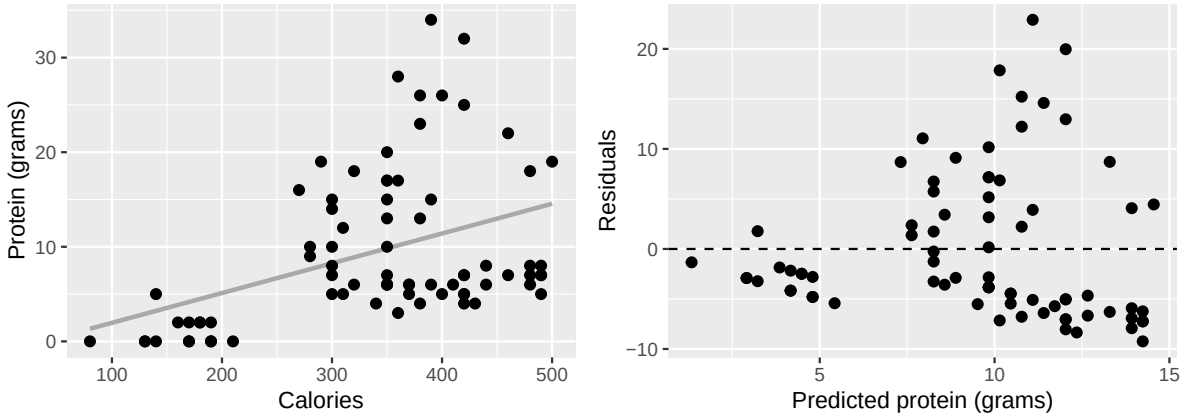
2. Researchers studying anthropometry collected body and skeletal diameter measurements, as well as age, weight, height and sex for 507 physically active individuals. The scatterplot below shows the relationship between height and shoulder girth (circumference of shoulders measured over deltoid muscles), both measured in centimeters.



   a. Describe the relationship between shoulder girth and height.

   b. How would the relationship change if shoulder girth was measured in inches while the units of height remained in centimeters?

3. *The scatterplot below shows the relationship between the number of calories and amount of protein (in grams) Starbucks food menu items contain. Since Starbucks only lists the number of calories on the display items, we might be interested in predicting the amount of protein a menu item has based on its calorie content.



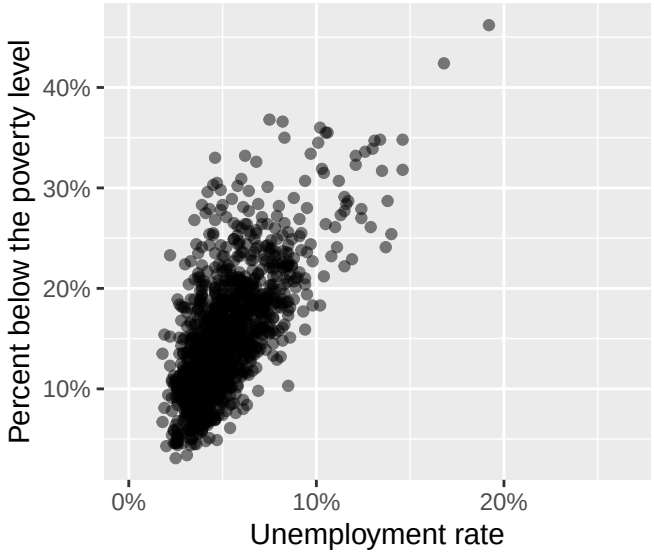
   a. Describe the relationship between number of calories and amount of protein (in grams) that Starbucks food menu items contain.

   b. In this scenario, what are the predictor and outcome variables?

   c. Why might we want to fit a regression line to these data?

   d. What does the residuals vs. predicted plot tell us about the variability in our prediction errors based on this model for items with lower vs. higher predicted protein?

4. *The following scatterplot shows the relationship between percent of population below the poverty level (`poverty`) from unemployment rate among those ages 20-64 (`unemployment_rate`) in counties in the US, as provided by data from the 2019 American Community Survey. The regression output for the model for predicting `poverty` from `unemployment_rate` is also provided.



term	estimate	std.error	statistic	p.value
`(Intercept)`	4.604	0.349	13.182	<0.0001
`unemployment_rate`	2.054	0.062	33.110	<0.0001

a. Write out the linear model.

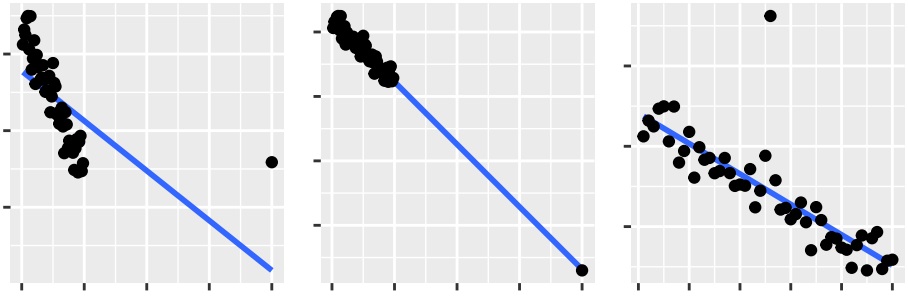
b. Interpret the intercept.

c. Interpret the slope.

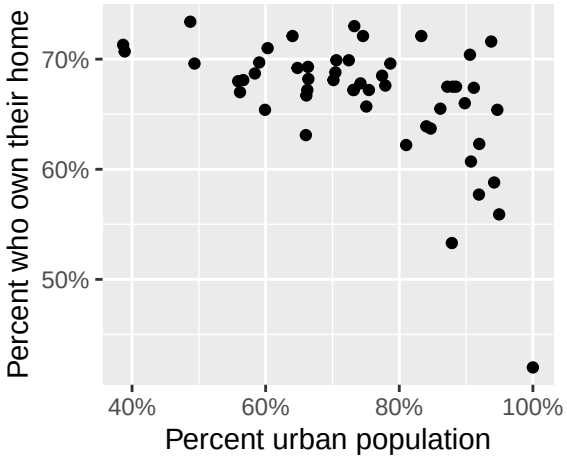
d. For this model R^2 is 46%. Interpret this value.

e. Calculate the correlation coefficient.

5. Identify the outliers in the scatterplots shown below, and determine what type of outliers they are. Explain your reasoning.



6. The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas. There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.



a. Describe the relationship between the percent of families who own their home and the percent of the population living in urban areas.

b. The outlier at the bottom right corner is District of Columbia, where 100% of the population is considered urban. What type of an outlier is this observation?