STAT 140: Homework 3

Linear Regression, Part I

YOUR NAME HERE 2/14/2019

BOOK EXERCISES (OpenIntro Statistics, Fourth Edition)

8.2) Trends in the residuals. Shown on pg. 312 are two plots of residuals remaining after fitting a linear model with two different sets of data. Describe important features and determine if a linear model would be appropriate for these data. Explain your reasoning.

a)

b)

8.7) Match the correlation, Part I. Match each correlation with the corresponding scatterplot (pg. 314).

- a)
- b)
- c)
- d)
- 1)
- e)

8.10) Guess the correlation. Eduardo and Rosie are both collecting data on number of rainy days in a year and total rainfall for the year. Eduardo records rainfall in inches and Rosie in centimeters. How will their correlation coefficients compare?

8.16) **Correlation, Part II.** What would be the correlation between the annual salaries of males and females at a company if for a certain type of position men always made

- a) \$5,000 more than women?
- b) 25% more than women?
- c) 15% less than women?

Note: if you want to build your intuition for a problem like this, you can always make up some data that follows each of these patterns (or patterns like these) and find the correlation in R.

8.17) Units of regression. Consider a regression predicting weight (kg) from height (cm) for a sample of adult males. What are the units of the correlation coefficient, the intercept, and the slope?

Answer:

8.20) Over-under, Part II. Suppose we fit a regression line to predict the number of incidents of skin cancer per 1,000 people from the number of sunny days in a year. For a particular year, we predict the incidence of skin cancer to be 1.5 per 1,000 people, and the residual for this year to be 0.5. Did we over or under estimate the incidence of skin cancer? Explain your reasoning.

Answer:

8.26) Cats, Part I. The following regression output is for predicting the heart weight (in grams) of cats from their body weight (in kg). The coefficients are estimated using a dataset of 144 domestic cats.

- a) Write out the linear model.
- b) Interpret the intercept.
- c) Interpret the slope.
- d) Interpret R^2 .
- e) Calculate the correlation coefficient.

R EXERCISE

The setup chunk loads packages and the data we need. Run it before you run any other chunks.

```
## Load packages
library(ggplot2)
```

```
## Load data
bac <- read.csv("https://www.openintro.org/stat/data/csv/bac.csv")</pre>
```

Make a scatterplot of bac and beers (use ggplot).

make scatterplot

Use the cor() function to find the correlation between bac and beers.

find correlation

Does the order of the argument in cor() matter?

Using the lm() function, fit a linear model that predicts blood alcohol concentration (BAC) from beers. Name your model bac_model.

fit bac linear model

Use the summary() function to summarize the bac_model fit.

get the summary of the fit

Create a residual plot for the model. Are there any patterns present?

create a residual plot

Find R^2 and interpret it in the context of the problem (No R required - a text answer is fine).