In-class exercise: Regression

1. A store keeps track of purchase history. The manager wonders if there is an association between the amount of money a customer spends on their first visit to the store and their second visit. Below is data collected on 10 customers. Each column corresponds to one customer. For example, the first customer spend $20 on the first visit and $23 on the second visit. The second customer spend $32 on first visit and $34 on second, etc.

Money spent on first visit (in dollars): 20,32,35,34,40,51,52,56,57,68
Money spent on second visit (in dollars): 23,34,36,44,42,51,54,57,54,62

a. Display the relationship between first and second visit dollar amounts?
b. Describe the pattern in part (a) briefly. Is there a relationship? Is it positive or negative? Is it linear or non-linear? Is it weak or strong?
c. Calculate the correlation coefficient between the amount of money spent on the first visit and the second visit.
d. What does the standard error in part (c) refer to?
e. Calculate an approximate 95% confidence interval for $\boldsymbol{\rho}$.

2. Answer the following question using the data from question (2).

a. Adding $30 to each of the observations for the second visit. How is the correlation coefficient between first and second visits affected? What can you conclude about the effects on the correlation coefficient of *adding* a constant to one or both of the variables?
b. Convert the first visit to cents (i.e., multiply by 100). How does this affect the correlation between the first and second visits? What can you conclude about the effects on the correlation coefficient of *multiplying* one or both of the variables by a constant?

3. Some species seem to thrive in captivity, whereas others are prone to health and behavior difficulties when caged. Maternal care problems in some captive species, for example, lead to high infant mortality. Can these differences be predicted? The following data are measurements of the infant mortality (percentage of births) of 20 carnivore species in captivity along with the log (based-10) of the minimal home range sizes (in km2) of the same species in the wild (Clubb and Mason 2003). For example, -1.3 is the home range and 4 is the captive infant mortality percentage.

Log home range size: -1.3 (4),-0.5 (22),-0.3 (0),0.2 (0),0.1 (11),0.5 (13),1.0 (17),0.3 (25),0.4 (24),0.5 (27),0.1 (29),0.2 (33),0.4 (33),1.3 (42),1.2 (33),1.4 (20),1.6 (19),1.6 (19),1.8 (25),3.1 (65)

a. Draw a scatter plot of these data, with log of home range size as the explanatory variable. Describe the association between the two variables in words.
b. Estimate the slope and intercept of the least squares regression line, with the log of home range size as the explanatory variable. Add this line to your plot.
c. Does home range size in the wild predict the mortality of captive carnivores? Carry out a formal test. Assume that the species data are independent.

d. Outliers should be investigated because they might have a substantial effect on the estimate so of the slope and intercept. Recalculate the slope and intercept of the regression line from part (c) after excluding the outlier at large home range size (which correspond to the polar bear). Add the new line to your plot. By how much did it change the slope?