Simonson_HW10

Martin Simonson

April 10, 2019

1. Bear Weight

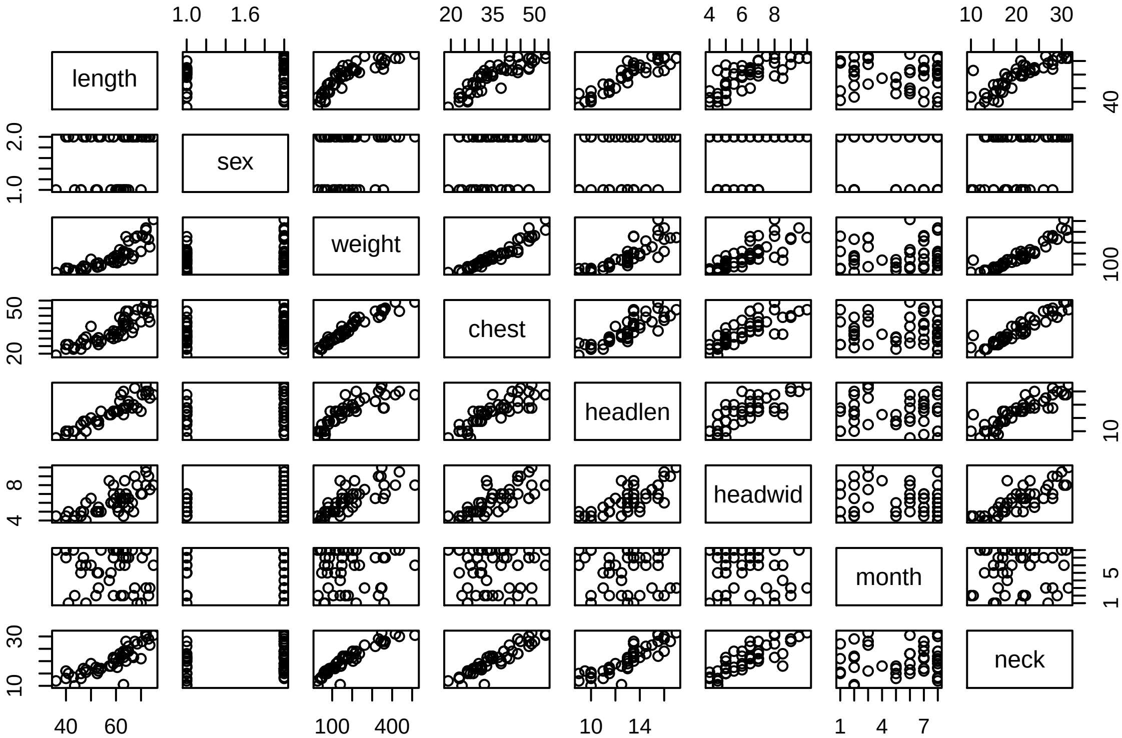
Animal ecologists wish to track the health of bear populations in the wild. By obtaining physical measurements of bears, researchers can gauge the effects of changes taking place in and around bear habitat (e.g., nearby housing and business development, recreational activities, introduction of non-native plant and animal species, severe storms, harsh winters, etc.). Researchers have studied bears by anesthetizing them in order to obtain vital measurements, such as age, gender, length, and width. A bear's weight is another important variable that is quite difficult to obtain in the wild because most bears are heavy and difficult to lift. The scientific problem is to develop a method for predicting the weight of a bear, given other more easily obtained measurements. A good method might alleviate the need to weigh bears in the wild and greatly simplify the data collection process.

In one detailed study, researchers were able to obtain weight measurements along with several other variables for each of 48 bears. The file *bears.txt* contains the data for 48 bears. There is one row for each bear that was anesthetized and measured carefully using a tape measure and scale. There is one column for each variable in the data set. The variables (from left to right) are:

- length: length of body (in inches)
- sex: male or female
- weight: weight (in pounds)
- chest: chest circumference (in inches)
- headlen: length of head (in inches)
- headwid: width of head (in inches)
- month: month of capture
- neck: neck circumference (in inches)

A)

Construct a scatterplot matrix of all the variables in the data set. Include the graph obtained and comment on the relationships between Y and the explanatory variables. When possible, describe the strength, direction, and type of relationship.



- **Answer:** We see that weight (Y) has a strong positive linear relationship with chest and neck circumference, and a weaker (but still positive and linear) relationship with body length, head width and head length.

B)

Estimate the parameters in the multiple regression model with weight as the response variable and chest, length, and neck as explanatory variables. Specifically, provide an estimate of the intercept, an estimate of the partial regression coefficient for each explanatory variable, and an estimate of the standard deviation of bear weights for any given values of the explanatory variables (i.e. $\hat{\sigma}$).

- **Answer:**

C)

Conduct one test of the null hypothesis that says that the partial regression coefficients for chest, neck and length are all zero against the alternative that at least one coefficient is not. State the hypotheses, the test statistic, the degrees of freedom, p-value and conclusion.

- **Answer:**

D)

Provide an interpretation of the partial regression coefficient associated with the variable *chest*.

- **Answer:**

E)

Compute a 95% confidence interval for the partial regression coefficient associated with the variable length. Is the partial regression coefficient associated with the variable length significantly different from zero? Explain how your confidence interval can be used to answer this question.

- **Answer:**

2. Bear Weight, revisited

Using the same model as in problem 1 (multiple linear regression of weight on *chest*, *length*, and *neck*):

A)

What proportion of variation in bear weights is explained by the multiple regression of weight on *chest*, *length*, and *neck*?

- **Answer:**

B)

Provide an estimate of the mean weight of a captured bear that is 60 inches long, with chest circumference of 35 inches, and neck circumference 24 inches.

- **Answer:**

C)

Provide a 95% confidence interval for the mean weight estimated in part (2b).

D)

Provide a 95% prediction interval for the weight of the bear described in part (2b).

- **Answer:**

E)

Examine the residual plot and the normal probability plot of the residuals from the fit of the multiple regression model with weight as the response variable and *chest*, *length*, and *neck* as explanatory variables. Which of the assumptions of multiple linear regression is questionable based on these plots?

- **Answer:**

F)

i)

How many categorical variables are provided in the data set? What are these variables?

- **Answer:**

ii)

How many dummy/indicator variables would you need to construct to incorporate the variable *month* in a regression model? Define these variables here, using the Indicator coding used by the textbook. What level of this factor was the reference level?

- **Answer:**

G)

Add the variable *sex* to the model in question 1b. Use R to fit the new model to the data. Interpret the estimated coefficient associated with the new *sex* variable within the context of the data.

- **Answer:**

3. Mpi genotypes

The 1989 Nature article titled "Selection component analysis of the Mpi locus in the amphipod" presented a study of data collected on the amphipod crustacean *Platorchestia platensis* on a beach near Stony Brook, Long Island, in April, 1987. The research group counted the number of eggs each female was carrying, freeze-dried them, weighed them, then used protein electrophoresis to determine the genotype at the locus for mannose-6-phosphate isomerase (Mpi). The data in *MPI.dat* consist of four columns: the identification number of the individual, the weight (in mg), fertility (quantified by the number of eggs), and the genotype. The biological question is whether the Mpi genotypes differ in size-adjusted fecundity.

A)

To answer the biological question, fit a parallel regression lines model to all genotypes. Write the estimated regression equation here, clearly explaining any notation or coding used.

- **Answer:**

[

]

B)

Using the parallel lines model, provide an interpretation of the relationship between weight and fertility. Make sure to include in your interpretation a number quantifying the magnitude of the relationship.

- **Answer:**

C)

Using the parallel lines model, estimate the mean difference in fertility between genotypes *ff* and *ss*, after adjusting for the effects of weight. Provide an s.e. of the estimate.

- **Answer:**

D)

Using the parallel lines model, test the null hypothesis of no difference between genotypes *ff* and *ss*, after adjusting for the effects of weight.

- **Answer:**

E)

Using the parallel lines model, test the hypothesis of no differences in fertility between the three genotypes, after adjusting for the effects of weight. Write down the null and alternative hypothesis, give the name of the test used, the p-value, and a one sentence conclusion.

- **Answer:**

4. Mpi genotypes, revisited

Consider the same biological question as in Poblem 3, but now fit a regression model that allows the slope of the three regression lines (one for each genotype group) to vary.

A)

a) Write out the estimated regression equation, clearly explaining any notation or coding used.

- **Answer:**

[

]

B)

Using this model, report the three estimates of the slope for the regression lines.

- **Answer:**

C)

Test the hypothesis that all three regression lines have the same slope. Write the null hypothesis, the p-value and a conclusion.

- **Answer:**

D)

Now you are ready to answer the biological question formulated in problem 3. Use any/all of the results obtained before this question to conclude whether the Mpi genotypes differ in size-adjusted fecundity. Be brief but clearly explain your argument.

- **Answer:**