Simple Linear Regression: Conditions and Transformations

Sleuth3 Chapter 8

Example 1: Exercise 8.17 in Sleuth3

Quote from book:

> In a study of the effectiveness of biological control fo the exotic weed tansy ragwort, researchers manipulated the exposure to the ragwort flea beetle on 15 plots that had been planted with a high density of ragword. Havesting the plots the next season, they measured the average dry mass of ragword remaining (gams/plant) and the flea beetle load (beetles/gram of ragword dry mass) to see if the ragwort plants in plots with high flea beetle loads were smaller as a result of herbivory by the beetles. (Data from P. McEvoy and C. Cox, "Successful Biological Control of Ragwort, *Senecio jacobaea, by Introduced Insects in Oregon,"*, Ecological Applications* 1 (4) (1991): 430-42.)

Here is the data:

```
## # A tibble: 6 x 2
##    Load  Mass
##   <dbl> <dbl>
## 1  12.2 18.2
## 2  14.6 17.5
## 3  15.8  7.22
## 4  25.3 30.6
## 5  38.6  6.66
## 6  76.4  6.14
```

Our explanatory variable is `Load`, and the response is `Mass`.

1. Make a suitable plot of the data.

2. Through trial and error, find a suitable transformation of the data so that the linear regression conditions are satisfied as well as possible. (Let's assume the plots were in different areas so that they can be regarded as independent.)

For the purpose of this lab, I'll give you a little more instruction than I would on a homework assignment:

- You should start by transforming the response variable, aiming to fix the problem with unequal standard deviations of the residuals. I ended up with a transformation of mass to the power of 0.25 – your goal for today is to understand why and see how you could arrive at that transformation yourself.
- Then, once you have pretty constant standard deviation of the residuals, if necessary you can try transformations of the explanatory variables until you have a relationship that looks closer to linear. I chose a log transformation for the load variable. Again, your goal is to understand why and how you could identify that transformation yourself through experimentation.
- For each transformation you try, you should make a scatter plot of the transformed variables, and a scatter plot of the transformed explanatory variable vs. the residuals from a simple linear regression model fit to the transformed variables.

3. Conduct a test of the claim that there is no association between the beetles load and the mean dry mass of ragweed harvested (after transformation).

4. Find an estimate of the median ragweed mass when the beetle load is 50, and when it is 250. Additionally, report a confidence interval for the median ragweed mass when the beetle load is 250.

Example 2: Adapted from Exercise 8.24 in Sleuth3

Quote from the book:

A high respiratory rate is a potential diagnostic indicator of respiratory infection in children. To judge whether a respiratory rate is truly "high", however, a physician must have a clear picture of the distribution of normal respiratory rates. To this end, Italian researchers measured the respiratory rates of 618 children between the ages of 15 days and 3 years. Analyze the data and provide a statistical summary. The following R code reads the data in.

```
## # A tibble: 6 x 2
##     Age  Rate
##   <dbl> <dbl>
## 1   0.1    53
## 2   0.2    38
## 3   0.3    58
## 4   0.3    52
## 5   0.3    42
## 6   0.4    62
```

Our explanatory variable is `Age` (in months), and the response is `Rate` (breaths per minute).

1. Make a suitable plot of the data.

2. Through trial and error, find a suitable transformation of the data so that the linear regression conditions are satisfied as well as possible. (Let's assume the measurements for different children in the sample can be regarded as independent.)

I selected transformations of $-1/\sqrt{Rate}$ for the response and $Age^{0.75}$ for the explanatory variable. Again, your goal is to understand the process for getting to those (or similar) transformations.

3. Obtain 95% prediction intervals for the respiratory rates *on the original data scale* for children of ages 5 months, 10 months, 20 months, and 30 months. Add a display of your prediction intervals to a scatter plot of the original data (you can copy/paste your plot code from part 1 and add a layer for the prediction intervals).