Examples for Chapter 3 Stat286

1. Data set: **Huber.txt (**from the book website**)**
   1. Use YBad. Plot the data and get the least-square regression line.
   2. Get the residual plot. Do you think the linear model is appropriate?
   3. Find out the leverages for the data points.
   4. Fit the data with quadratic model.
   5. Look at the residual plot, do you think the quadratic model is appropriate?
2. The US Treasure bonds are among the least risky investments, in terms of the likelihood of your receiving the promised payment. Data set **bonds.txt** from the textbook website contains the couple rate and bid price for US Treasury bonds maturing between 1994 and 1998. The bid prices are listed per “face value” of $100 to be paid at maturity. Half of the coupon rate is paid every six months. We want to find the model that uses coupon rate to predict the bid price.
   1. Plot the data with the least-square regression line. Do you think the linear model is appropriate?
   2. What’s the least-square regression line? What’s the corresponding R2?
   3. Find the 95% confidence interval for the slope.
   4. Are there any leverage points? Outliers? What are they?
   5. Fit the linear model again after removing the leverage points 4, 13 and 35, as those are “flower bonds” that have tax advantage compared to the other bands. Do you think the linear model is appropriate after removing the leverage and outliers?
   6. Plot the Cook’s distance with all data included. Are there any observations with large Cook’s distance? If so, what are they?
   7. Do you think the normality assumption is satisfied?
3. A building maintenance company is planning to submit a bid on a contract to clean corporate offices scattered throughout an office complex. They study the relationship between the number of crews and the number of rooms cleaned by those crews. Data set **cleaning.txt** from the textbook website contains the number of rooms that were cleaned by varying number of crews for a sample of 53 days.
   1. In this data, which variable is the predictor? Which one is the dependent variable?
   2. Plot the data. Do you think the linear model is appropriate?
   3. Find the lease-square regression line that uses the number of crews to predict the number of rooms clean.
   4. What’s the R2 for this model?
   5. Using the linear model, predict the number of rooms cleaned by 4 and 16 crews, provide 95% prediction interval for your predictions. For the purpose of exercise, do NOT adjust for multiple intervals.
   6. Look at the diagnostic graphs using the **plot(out)** command. Which assumption is violated?
   7. Notice that both variables are counts, now transform them using the square root transformation. Re-fit the model.
   8. Do you think the assumptions are satisfied for the transformed model?
   9. What are the prediction intervals for the transformed model? Again, do NOT adjust the multiple intervals here.
4. TV case study
5. Data: **responsetransformation.txt (**from the textbook**)**
   1. Look at the scatter plot, do you think x and y are linearly related? Now try to fit the straight line regression model. Look at the residual plot. Is there any pattern in the residual plot? Do you think the constant variance assumption is satisfied?
   2. Get the Inverse Response Plot. What’s the best power transformation for the data?
   3. According to your transformation, what’s the model?
   4. Plot the original data, and add your fitted model to the graph. What do you think?
6. Data: **salarygov.txt (**from the textbook**)**

In this data, we are interested in two variables:

**Score**: score for job class based on difficulty, skill level, training requirements and level of responsibility as determined by a consultant to the government unit.

**MaxSalary**: contains the maximum salary (in $) for employees in the job class.

Find the appropriate model that uses Score to predict the maximum Salary.

* 1. The analysis starts by looking at the distribution of Score, MaxSalary, and the scatter plot of these two variables. What can you conclude?
  2. Find appropriate transformation for x first, so the translated x is approximately normal.
  3. Now, using the transformed x, find the power transformation of Y, using the Inverse Response Plot.
  4. What do you think about the transformed model?
  5. Now, use the box-cox method to find the translation in Y.
  6. Check the transformation of log(Y).