Examples for Chapter 6 Stat286

1. Work on the computer generated data **caution** in R library **alr3**.
   1. Read in the data by using the following command: **library(alr3)**, **attach(caution)**
   2. How many variables are there? What are they? How many cases?
   3. Make a matrix plot of each individual x to y by using the command: **pairs(y~x1+x2).** What can you say about the relationship between x1, x2 and y?
   4. What’s the correlation coefficient between x1 and x2?
   5. Find the least-square regression line for y on x1 and x2. Look at the following residual plots: (x1, standardized residual), (x2, standardized residual), (fitted value, standardized residual). What assumptions are violated?
2. Data set **manager.csv** contains average annual income of managers during the past two years (x1), a score measuring each manager’s risk aversion (x2), and the amount of life insurance carried (Y) for a sample of 18 managers in the 30-39 age group. Risk aversion was measured by a standard questionnaire administered to each manager: the higher the score, the greater the degree of risk aversion.
   1. Are the Income and Risk highly correlated?
   2. Find the least-square regression line for life insurance on income and risk. Does the linear model satisfy the assumptions?
   3. Look at the Added Variable Plot. What do the graphs suggest?
   4. Play with the data, try to find a model that improves the original linear model.
3. Use the Example of menu pricing in a new Italian Restaurant. Data: **nyc.cvs** (from the book website)
   1. Find the least square line with all the 4 predictors: Food, Décor, Service and East.
   2. Look at the residual plot. Does the plot suggest any violation of assumptions?
   3. Do you think we should keep Service in the model, with all other three variables already in? Carry-out a t-test.
   4. Look at the Added Variable Plot. What can you conclude?
   5. What’s your final model?
4. Data set named **defects.txt** contains the average defects per 1,000 parts produced (Denoted by Defective), Temperature, Density and Rate. (From textbook, page 167). Try to build a model that uses Temperature, Density and Rate to predict Defective.
   1. Look at the scatter plot matrix of the predictors and response. What can you say about the matrix plot?
   2. Start with a linear model with three predictors: Temperature, Density and Rate. What’s the model equation? What does the residual plot indicate?
   3. Use the Inverse Response Plot to find the appropriate transformation of Y.
   4. Use the Box-Cox method to find the transformation of Y.
   5. Get the scatter plot the each predictor with the transformed Y. What can you say?
   6. Find the model with three predictors and the transformed Y. How’s the residual plot look?
   7. Do you think all the three predictors should be included in the model? What’s your final model?
5. (From Textbook, data set: **magazine.csv**) An analyst is interested in understanding the relationship between revenue from magazine sales and that from advertising. The analyst has obtained some US data from *Advertising Age’s* 14th annual Magazine 300 report. (<http://www.adage.com>) which was released in September 2003. Data are available for 204 US magazines for the following variables:

Y=AdRevenue = Revenue from advertising (in thousands of $)

X1= AdPages = # of pages of paid advertising

X2=SubRevenue = Revenue from paid subscriptions (in thousands of $)

X3 = NewsRevenue = Revenue from newsstand sales (in thousands of $)

Build a model that predicts AdRevenue from Ad pages, Sub Revenue and News Revenue.

* 1. First look at the scatter plot matrix, what can you say about the possible relationship?
  2. What’s the linear model that predicts AdRevenue from AdPages, SubRevenue and News Revenue? Is it a good model?
  3. First try to find the appropriate transformation for the predictors, and then using the transformed predictors to find the transformation of Y. What’s your model? Is it good?
  4. Try to find a good model by transforming predictors and response simultaneously. Are your results in (c) and (d) the same?

1. Consider the Newspaper circulation data introduced in Chapter 1. Data: **circulation.txt** (from the book website). Try to build a model to predict Sunday circulation, using the Weekday circulation and whether the newspaper is a tabloid with s serious competitor. Using your model to predict the Sunday circulation for both tabloid and non-tabloid newspapers when the weekday circulation is 210,000. Give out the 95% prediction interval.
2. Data set **profsalary.txt** from the book website contains the salary (in thousands of dollars) and the numbers of years of experience. Find a model that descripts the relationship between salary and experience. (Chapter 5 example)
3. The file named **Bodyfat.csv** contains the data for a study of the relationship of amount of body fat to several possible predictor variables, based on a sample of 20 healthy females 25-34 years of age. The possible predictors are triceps skinfold thickness (*x1*), thigh circumference (*x2*), and midarm circumference (*x3*). The amount of body fat was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water. It would therefore be very helpful if a regression model with some or all of these predictor variables would provide reliable estimate of the amount of body fat since the measurements needed for the predictor variables are easy to obtain.
   1. Find the regression line of body fat on triceps skinfold thickness (*x1*).
   2. Find the regression line of body fat on thigh circumference (*x2*).
   3. Find the regression line of body fat on triceps skinfold thickness and thigh circumference (*x*1 and *x*2).
   4. Find the regression line of body fat on triceps skinfold thickness, thigh circumference and midarm circumference (*x*1, *x*2 and *x*3).
   5. Do you think the variable Midarm Circumference (*x*3) should be included in the model after triceps skinfold thickness and thigh circumference are already included? Perform a t-test and partial F-test to answer the question. Let α = 0.05.
   6. Do you think the variable Thigh Circumference (*x*2) should be included in the model after triceps skinfold thickness and Midarm Circumference are already included? Perform a t-test and partial F-test to answer the question. Let α = 0.05.
   7. Can you drop both *x*2 and *x*3 simultaneously?
   8. Compare the coefficients of triceps skinfold (*x1*) thickness in models in (a), (c) and (d). Do those coefficients change a lot?
   9. Compare the coefficients of thigh circumference (*x2*) in models in (b), (c) and (d). Do those coefficients change a lot?
   10. Why those coefficients change a lot in different models?
   11. Find the best model to predict the body fat.
4. Case Study: Wine Critics’ Rating on Prices: **Bordeaux.csv** (from the book website)
   1. Total there are 7 predictors in the data. How many are dummy variables?
   2. Look at the scatter matrix plot for response (Price) and non-dummy variables. What can you conclude?
   3. Work on the Y (Price) and the two continuous predictors (ParkerPoints and CoatsPoints). Try to find appropriate transformation for these variables.
   4. Now, add the dummy variables to the model. Find the best model that predict price.
   5. Is your final model easy to interpret?
   6. The book suggests using the log transform to the continuous variables (Price, ParkerPoints and CoatesPoints). Now, use the log transform to find the appropriate model.
   7. How to interpret the coefficients?
   8. Any outliers?