Statistics 347, Homework 2, due Thursday January 23

Discussion of homework problems among students is encouraged. However, all material handed in for credit must be your own work.

Please hand in each problem in a separate file.

1. The paper by Case and Deaton (2015) discusses the increase in mortality rates for white non-hispanic Americans between ages 45 to 54. Some of the data used in the paper is in the file Mortality.Rdata which you can load into R.

   (a) Read the paper and summarize the conclusions in a brief paragraph.

   (b) Plot the distribution of ages in the 45 to 54 range for both years. What do you observe? How could that affect the conclusion that mortality rates in the 45-54 age range increased among white non-hispanic Americans between 1999 and 2013.

   (c) Propose a way to control for the age distribution using a generalized linear model. Write out your model clearly, what model formula are you using, what distribution family, and explain how you would use it to test the hypothesis of the paper. Fit the model to the data and analyze the results. Do you agree with the authors' conclusions?

   (d) Perform the same analysis on deaths from suicides and from substance abuse. What are your conclusions.

2. A biologist analyzed an experiment to determine the effect of moisture content on seed germination. Eight boxes of 100 seeds each were treated with the same moisture level. Four boxes were covered and four left uncovered. The process was repeated at six different moisture levels. (Dataset `seeds` in faraway package).

   germ moisture

   (a) Plot the germination percentage against the moisture level on two side-by-side plots according to the coverage of the box. What relationship do you see?

   (b) Create a new factor describing the box (the data are ordered in blocks of 6 observations per box). Add lines to your previous plot that connect observations from the same box. Is there an indication of a box effect?

   (c) Fit a binomial response model with main effects including the coverage, box and moisture predictors. There is an NA appearing for the BOX8 factor level. Can you explain why?

   (d) Test for the significance of a box effect in your model. Deviance

   (e) If the box factor is not significant, how could you aggregate the data? Will that change the estimated parameters? Will it change the deviance and degrees of freedom?

   (f) Use the plots to determine an appropriate choice of model beyond main effects.

   (g) At what value of moisture does the predicted maximum germination occur for non-covered seeds? For covered seeds?

   (h) Produce a plot of the residuals against the fitted values and interpret.

   (i) Plot the residuals against moisture while distinguishing the covering. Interpret.

   (j) Compute the hat matrix explicitly - show the r-commands you use to obtain it and obtain the individual leverages.

   (k) Plot the residuals against the leverages. Are there any influential points?

3. Faraway Chapter 4, problem 1.