Homework 4

Due October 1 at 3:00

Searching for solutions is forbidden.

(1) Get the cat data using `data(cats)`. Fit the model:

```
out = lm(Hwt ~ Sex + Bwt,data=cats)
```

(a) Plot the data. Add the fitted line for male cats and the fitted line for female cats.

(b) Now add an interaction between Sex and Bwt. Plot the data again. Add the fitted line for male cats and the fitted line for female cats. What differences, if nay, are there compared to the previous plot?

(c) Add an interaction What is the estimated slope for Males? What is the estimated slope for Females?

(d) Another approach is to fit two separate regressions, one for Male and one for female. Do this. What is the estimated slope for Males? What is the estimated slope for Females? Plot the data and add both lines.

(e) The advantage of the second method is that it makes a weaker assumptions than the first method. What is the weaker assumption?

(2) Download the dataset SENIC.txt from Canvas. The variables on 113 hospitals are:

hospital ID, average Length of stay in hospital (in days), average age, probability of infection, culturing ratio (cultures performed divided by number of patients with no infections, chest x-ray ratio, number of beds, medical school affiliation (yes=1 or no=2), geographic region, number of patients, number of nurses, available facilities divided by services. Use the following code to get the data:

```
df = read.table("SENIC.txt")
str(df)
colnames(df)
df = df[,-1]
colnames(df) = c("Stay","Age","Infection","Culture","Xray",
      "Beds","MedSchool","Geo","Patients","Nurses","Facilities")
names(df)
df$Geo = factor(df$Geo)
is.factor(df$Geo)
```

(a) Fit a linear model to predict length of stay from the other variables. What variables appear to be significant?

(b) What is the R^2 value? How do you interpret this?

(c) Examine the diagnostics (residuals etc). Comment on your findings. No need to include every plot. Just comment on what you conclude from the diagnostics.

(d) Test whether the variable Geo is significant. (Careful. This is a factor with 3 levels.)

(e) Add an interaction between Infection and Age. Is this new variable significant?

(f) Now fit a linear model to predict length of stay from just Infection. Summarize the fit (fitted model, tests and diagnostics).

(g) Repeat (f) but try using a cubic polynomial. Did this improve the model in any way?