STAT 561 - Homework 1

Instructions

1. *Due Monday, April 6th at the 11:59pm. Any submission after that (24hrs after) will be graded out of 50%.*

2. *Your submission should include a* ***pdf*** *from your generated R markdown, and your R markdown file.*

3. *Make sure to highlight the part of the output that have the information you will be providing as answers.*

4. This work should be done as a group. Only one person in the group should submit the work with the names of all the people in the group on the document

5. Send me an email directly if you feel one or more members in your group are not pulling their weights in the homeworks.

---

1. **Patient satisfaction data** . A hospital administrator wished to study the relation between patient satisfaction (Y) and patient's age (X_1, in years), severity of illness (X_2, an index), and anxiety level (X_3, an index). The administrator randomly selected 46 patients and collected the data presented below, where larger values of Y, X_2, and X_3 are, respectively, associated with more satisfaction, increased severity of illness, and more anxiety.

```
#load data using
pat_sat=read.data("pat_sat.txt",header=T)

#Have a look at the data
View(pat_sat)

#Check the columns of the data
str(pat_sat)
```

a. Prepare a histogram and box plot for each of the predictor variables using the hist() and boxplot() functions in R. Also use summary() to generate summaries for each of the predictor variables (do not produce the summary results). Are any noteworthy features revealed by these plots and your exploration?

b. Obtain the scatter plot matrix and the correlation matrix using the pairs() and cor() functions respectively. Interpret these and state your principal findings. Do you see any observations that are extreme (separated from the others)?

c. Fit a multiple linear regression model for three predictor variables to the data and state the estimated regression function. How is $\hat{\beta}_2$ interpreted here?

d. Conduct a test to check if the overall model is significant; use $\alpha = .05$. State the null and alternative hypotheses, p-value decision whether to reject H_0 or to fail to reject H_0, and your conclusion (*Hint : Use the F-test.*).

e. Obtain a 90% confidence interval for β_1 using the code below. Interpret your results.

```
model <- lm(...)
confint(model,level=0.9) #95% confidence intervals for the model coefficients
```

f. What is the coefficient of multiple determination value produced by your model (this is same as R^2). What does this indicate about the model?

g. Predict the patient satisfaction for a new patient with $X_1 = 35, X_2 = 45$, and $X_3 = 2.2$. Also give a 90 percent prediction interval for this new observation. Interpret your prediction interval.
You can utilize the following code:

```
attach(pat_sat)
model <- lm(...)
# New data for prediction
new_data <- data.frame(X1=...,X2=...,X3=...)

# Predict mpg with prediction intervals
predicted_values <- predict(model, newdata = new_data, interval = "prediction")
```

Note: *A prediction interval is a range within which a future observation is expected to fall with a certain probability, given a specific level of confidence. It accounts for both the uncertainty in estimating the underlying model and the natural variability of the data.*

h. Use both forward and backward selection criteria to select a final model. Do the 2 criteria produce the same model? If not, which of the models offers the best balance between complexity and explanatory power and produces the lowest AIC? (*Hint: use trace=1 in the step() function.*)

Polynomial regression

2. **Refer to Muscle mass data**. A person's muscle mass is expected to decrease with age. To explore this relationship in women, a nutritionist randomly selected 15 women from each 10-year age group, beginning with age 40 and ending with age 79. The results follow; X is age, and Y is a measure of muscle mass.

a. What is the correlation between age and muscle mass measure? Interpret this value.

b. Fit a first-order regression model to the data :

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Plot the fitted regression function and the data. Does the this regression function appear to be a good fit here? Record the value, R^2.

c. Fit a second-order regression model

$$Y_i = \beta_0 + \beta_1 X_i + \beta_{11} X_i^2 + \varepsilon_i$$

d. Plot the fitted regression functions in a) and b) on the same scatterplot of the data using different colors. Which of the regression functions appears to be a better fit?.

e. Test whether or not there is a significant regression relation for the model in b); use $\alpha = .05$. (Just give the conclusion of the test and report the p-value).

f. Test whether the quadratic term can be dropped from the regression model; use $\alpha = .05$. (*Hint: This is where you use the p-value for the quadratic term produced in the summary. Your null hypothesis is* $H_0 : \beta_{11} = 0$ *against the alternative* $H_a : \beta_{11} \neq 0$)

g. Fit a third-order model and test whether or not $\beta_{111} = 0$: use $\alpha = .05$. Just state your conclusion and p-value.

Qualitative predictors

3. **Refer to the CDI data set**. The number of active physicians (Y) is to be regressed against total population (X_1), total personal income (X_2), and geographic region (X_3, X_4, X_5).

a. Fit a multiple linear regression model. Write the regression equation, specify what X_3, X_4 and X_5 are and how they are encoded.

b. Briefly explain what the coefficients β_2 and β_3 in the context of the model.

0.1 CDI Data Description

This data set provides selected county demographic information (CDI) for 440 of the most populous counties in the United States. Each line of the data set has an identification number with a county name and state abbreviation and provides information on 14 variables for a single county. Counties with missing data were deleted from the data set. The information generally pertains to the years 1990 and 1992 . The 17 variables are:

Variable Number	Variable Name	Description
1	Identification number	$1 - 440$
2	County	County name
3	State	Two-letter state abbreviation
4	Land area	Land area (square miles)
5	Total population	Estimated 1990 population
6	Percent of population aged $18 - 34$	Percent of 1990 CDI population aged 18-34
7	Percent of population 65 or older	Percent of 1990 CDI population aged 65 years old or older
8	Number of active physicians	Number of professionally active nonfederal physicians during 1990
9	Number of hospital beds	Total number of beds, cribs, and bassinets during 1990
10	Total serious crimes	Total number of serious crimes in 1990 , including murder, rape, robbery, aggravated assault, burglary, larceny-theft, and motor vehicle theft, as reported by law enforcement agencies
11	Percent high school graduates	Percent of adult population (persons 25 years old or older) who completed 12 or more years of school
12	Percent bachelor's degrees	Percent of adult population (persons 25 years old or older) with bachelor's degree
13	Percent below poverty level	Percent of 1990CDI population with income below poverty level
14	Percent unemployment	Percent of 1990 CDI labor force that is unemployed
15	Per capita income	Per capita income of 1990CDI population (dollars)
16	Total personal income	Total personal income of 1990CDI population (in millions of dollars)
17	Geographic region	Geographic region classification is that used by the U.S. Bureau of the Census, where: $1 = NE, 2 = NC, 3 = S, 4 = W$