2028: Basic Statistical Methods
Homework 8 (Last Homework)

This homework is due Monday, Dec. 2nd in class **BEFORE class starts**. Late papers will not be accepted. Please do not turn in any papers to any mailbox.

- Please remember to staple if you turn in more than one page.
- Please make sure to **SHOW ALL WORK** in order to receive full credit.

Questions 1-3 are about ratings of NFL quarterback and points they gained.

1. **Simple linear regression:** 11-3, page 410.

   Data file name is data113.txt.

2. **Hypothesis testing:** 11-25, page 419.

3. **Confidence Interval:** 11-41 page 425 (a,b,d)

4. **Mercury pollution.**

   Mercury pollution is a serious problem in some waterways. Mercury levels often increase after a lake is flooded due to leaching of naturally occurring mercury by the higher levels of the water. Excessive consumption of mercury is well known to be deleterious to human health. It is difficult and time consuming to measure every persons mercury level; a quick procedure would be nice that could be used to estimate the mercury level of a person based upon the average mercury level found in fish and estimates of the person's consumption of fish. The following data was collected on the methyl mercury intake of subjects and the actual mercury levels recorded in the blood stream from a random sample of people around recently flood lakes.

Intake ($\mu gHg/day$)	Level (ng/g)
180	90
200	120
230	125
410	290
600	310
550	290
275	170
580	375
600	150
105	70
250	105
60	205
650	480

A R session produced the following incomplete output.

```
Coefficients:
             Estimate Std. Error
 (Intercept)  50.4441    47.7674
 intake        0.4529     0.1154
---
Residual standard error: 84.45 on 11 degrees of freedom
```

a Test whether the intercept and the slope coefficients are significant. What is your conclusion at $\alpha = 0.01$?

b Estimate the mercury level for the mercury intake of 200. Find the 90% confidence interval for the mean response at this intake value.

5. **Smoking and Cancer**

The data are per capita numbers of cigarettes smoked (sold) by 43 states and the District of Columbia in 1960 together with death rates per thousand population from various forms of cancer. (Nevada and the District of Columbia are outliers in the distribution of cigarette consumption (sale) per capita by states in 1960. The ready explanation for the outliers is that cigarettte sale are swelled by tourism (Nevada) and tourism and commuting workers (District of Columbia). There are 44 number of cases and the variable names in the data file are

• `CIG`: Number of cigarettes smoked (hds per capita)

• `BLAD`: Deaths per 100K population from bladder cancer

• `LUNG`: Deathes per 100K population from lung cancer

• `KID`: Deaths per 100K population from bladder cancer

• `LEUK`: Deaths per 100 K population from leukemia

Reference: J.F. Fraumeni, "Cigarette Smoking and Cancers of the Urinary Tract: Geographic Variations in the United States," *Journal of the National Cancer Institute*, 41, 1205-1211.

Getting the Data: The data file name is *smoking.txt*. Once you have saved the data file in the working directory, read the data in R using the command

```
data = read.table("smoking.txt",header=TRUE)
```

In the analysis below, we will investigate the association of `CIG` to `LUNG` and `LEUK` using linear regression. Define first

```
cig = as.numeric(data[,2])
lung = as.numeric(data[,4])
leuk = as.numeric(data[,6])
```

Question 1: Exploratory Data Analysis.

• Using a scatterplot describe the relationship between `CIG` and `LUNG`, and the relationship between `CIG` and `LEUK`. Describe the general trend (direction and form). (Use `plot` function in R with two input variables (e.g. cig and lung). Write down the commands you used.

• What is the value of the correlation coefficients? Please interpret. (Use `cor` function in R with two input variables - `cig` and `lung`; `cig` and `leuk`). Discuss the difference in the strength in correlation between the the number of cigarettes and the number of deaths from the two types of cancer.

• Based on this exploratory analysis, is it reasonable to assume the simple linear regression model for the relationship between `CIG` and `LUNG`? How about between `CIG` and `LEUK`?

Question 2: Fitting the Simple Linear Regression Model. Fit a linear regression to evaluate the relationship between `CIG` and `LUNG` using simple linear regression.

The function in R is `lm`. We perform a linear regression with R as follows

```
model = lm(lung~cig)
summary(model)
```

(i) What are the model parameters and what are their estimates?

(ii) Write down the equation for the least squares line;

(iii) Interpret the estimated value of the slope parameter in the context of the problem (include its standard error in your interpretation).

(iv) Find a 95% confidence interval for the slope parameter.

Question 3: Checking the Assumptions of the Model. To check whether the assumptions are met, we are going to use three visual displays:

i the scatterplot of the data,

ii a residual plot - a plot of the residuals, e_i, versus $\widehat{y}_i$ (also called the predicted or fitted values),

iii the normal probability plot of the residuals.

You can use the graphical tools in R to obtain each of the three plots after extracting the residuals and the fitted values using the `model` object defined above

```
resid = residuals(model)
fits = model$fitted
#split the display into 4 graphical panels
par(mfrow = c(2,2))
# plot 1
plot(lung,cig)
#plot 2
plot(fits,resid)
#plot 3
qqnorm(resid)
#plot 4
hist(resid,main="")
```

Interpret the three displays with respect to the assumptions of the linear regression model. In other words, comment on whether there are any apparent departures from the assumptions of linear regression model. Make sure you state the model assumptions and assess each one. Describe what graphical tool you used to evaluate each assumption. Also are there any extreme outliers in the data/residuals?

Question 4: Testing the significance of the linear relationship observed in the data. Test whether there is a significant linear relationship between `LUNG` and `CIG` (in other words, test whether the linear relationship we observe in the data can be generalized to the entire population). Recall from class, that the null and alternative hypothesis are stated in terms of the slope parameter β_1:

$$\begin{cases} H_0: & \beta_1 = 0 \text{ (X and Y are not linearly related)} \\ H_1: & \beta_1 \neq 0 \text{ (X and Y are linearly related)} \end{cases} \tag{1}$$

From the output in the session window, answer the following:

(i) What is the P-value of the test?

(ii) What does the actual value of the P-value tell you?

(iii) State your conclusion in the context of the problem.

Question 5: Prediction Use the regression equation to predict the number of lung cancer deaths per 100K population for a number cigarettes to be smoked equal to 10 (hds per capita) and obtain a 99% prediction interval. How did you determine this confidence interval? Show work for full credit. Is the prediction confidence interval wide? Why?