2028: Basic Statistical Methods
Homework 6 (Last Homework)

This homework is due **Tuesday, Nov. 24** in class BEFORE class starts. Late papers will not be accepted.

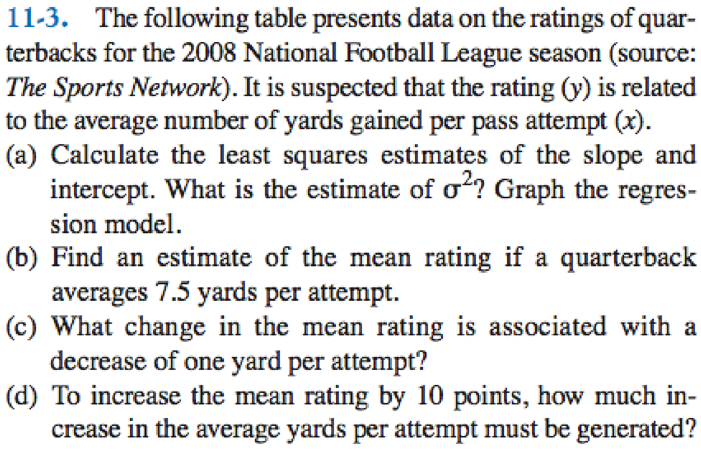
- Please remember to staple if you turn in more than one page.

- Please make sure to SHOW ALL WORK in order to receive full credit.

1. **Simple linear regression.** (100 points: 30 points + 30 points + 20 points.) Questions 11-3, 11-25, 11-41 on pages 410, 419 and 425. Data is contained in file data113.txt. You may also use data113.csv, and read the data using

   ```
   data = read.csv("data113.csv",header=TRUE)
   ```

   They are about analyzing the NFL player performance. The first question finds the simple linear regression model. The second question performs hypothesis testing. And the third question finds the confidence intervals of the parameters.



11-3. The following table presents data on the ratings of quarterbacks for the 2008 National Football League season (source: *The Sports Network*). It is suspected that the rating (y) is related to the average number of yards gained per pass attempt (x).

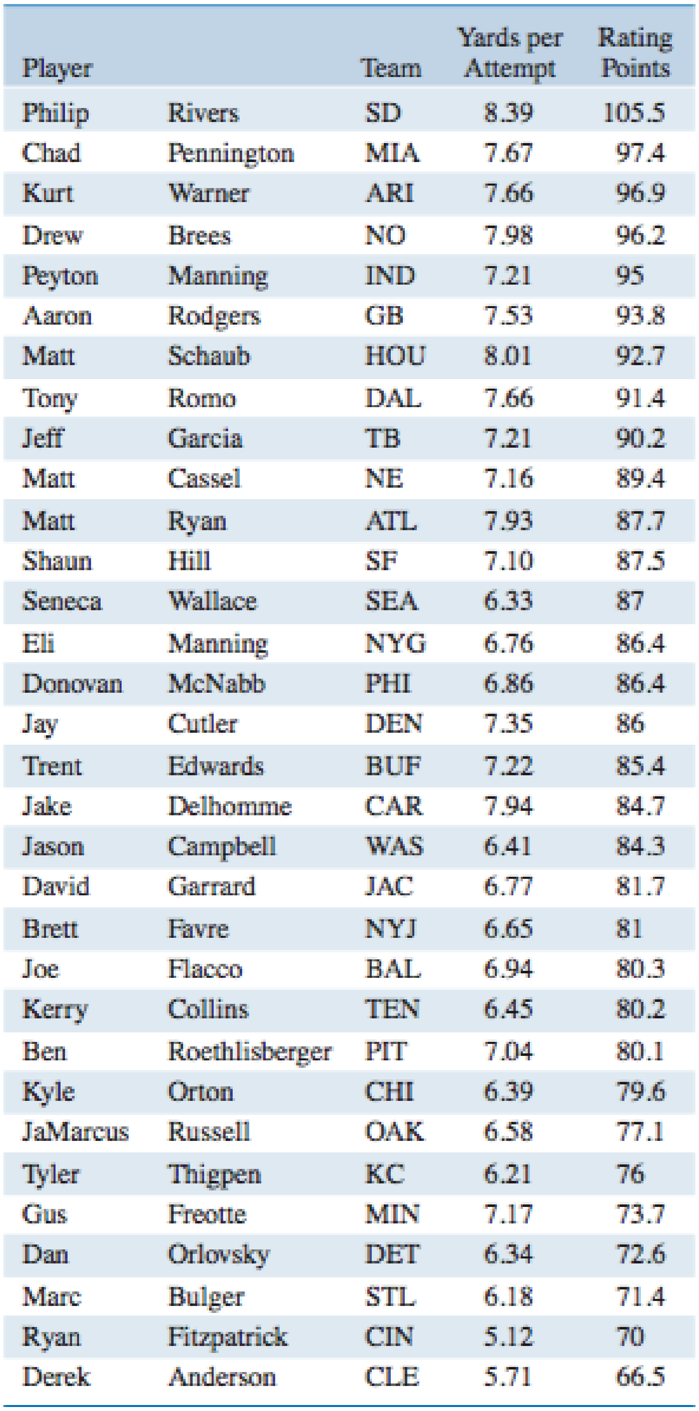
(a) Calculate the least squares estimates of the slope and intercept. What is the estimate of σ^2? Graph the regression model.

(b) Find an estimate of the mean rating if a quarterback averages 7.5 yards per attempt.

(c) What change in the mean rating is associated with a decrease of one yard per attempt?

(d) To increase the mean rating by 10 points, how much increase in the average yards per attempt must be generated?

(e) Given that $x = 7.21$ yards, find the fitted value of y and the corresponding residual.



Player		Team	Yards per Attempt	Rating Points
Philip	Rivers	SD	8.39	105.5
Chad	Pennington	MIA	7.67	97.4
Kurt	Warner	ARI	7.66	96.9
Drew	Brees	NO	7.98	96.2
Peyton	Manning	IND	7.21	95
Aaron	Rodgers	GB	7.53	93.8
Matt	Schaub	HOU	8.01	92.7
Tony	Romo	DAL	7.66	91.4
Jeff	Garcia	TB	7.21	90.2
Matt	Cassel	NE	7.16	89.4
Matt	Ryan	ATL	7.93	87.7
Shaun	Hill	SF	7.10	87.5
Seneca	Wallace	SEA	6.33	87
Eli	Manning	NYG	6.76	86.4
Donovan	McNabb	PHI	6.86	86.4
Jay	Cutler	DEN	7.35	86
Trent	Edwards	BUF	7.22	85.4
Jake	Delhomme	CAR	7.94	84.7
Jason	Campbell	WAS	6.41	84.3
David	Garrard	JAC	6.77	81.7
Brett	Favre	NYJ	6.65	81
Joe	Flacco	BAL	6.94	80.3
Kerry	Collins	TEN	6.45	80.2
Ben	Roethlisberger	PIT	7.04	80.1
Kyle	Orton	CHI	6.39	79.6
JaMarcus	Russell	OAK	6.58	77.1
Tyler	Thigpen	KC	6.21	76
Gus	Freotte	MIN	7.17	73.7
Dan	Orlovsky	DET	6.34	72.6
Marc	Bulger	STL	6.18	71.4
Ryan	Fitzpatrick	CIN	5.12	70
Derek	Anderson	CLE	5.71	66.5

11-25. Consider the National Football League data in Exercise 11-3.

(a) Test for significance of regression using $\alpha = 0.01$. Find the *P*-value for this test. What conclusions can you draw?

(b) Estimate the standard errors of the slope and intercept.

(c) Test H_0: $\beta_1 = 10$ versus H_1: $\beta_1 \neq 10$ with $\alpha = 0.01$. Would you agree with the statement that this is a test of the hypothesis that a one-yard increase in the average yards per attempt results in a mean increase of 10 rating points?

11-41. Refer to the NFL quarterback ratings data in Exercise 11-3. Find a 95% confidence interval on each of the following:

(a) Slope

(b) Intercept

(c) Mean rating when the average yards per attempt is 8.0

(d) Find a 95% prediction interval on the rating when the average yards per attempt is 8.0.

2. **Smoking and Cancer** (Optional.)

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?