Stat 6021: Homework Set 3

1. (R required) We will use the dataset "Copier.txt" for this question. The Tri-City Office Equipment Corporation sells an imported copier on a franchise basis and performs preventive maintenance and repair service on this copier. The data have been collected from 45 recent calls on users to perform routine preventive maintenance service; for each call, *Serviced* is the number of copiers serviced and *Minutes* is the total number of minutes spent by the service person.

   (a) What is the response variable in this analysis? What is predictor in this analysis?

   (b) Produce a scatterplot of the two variables. How would you describe the relationship between the number of copiers serviced and the time spent by the service person?

   (c) Use the `lm()` function to fit a linear regression for the two variables. Where are the values of $\hat{\beta}_1$, $\hat{\beta}_0$, R^2, and $\hat{\sigma}^2$ for this linear regression?

   (d) Interpret the values of $\hat{\beta}_1$, $\hat{\beta}_0$ contextually. Does the value of $\hat{\beta}_0$ make sense in this context?

   (e) Use the `anova()` function to produce the ANOVA table for this linear regression. What is the value of the ANOVA F statistic? What null and alternative hypotheses are being tested here? What is a relevant conclusion based on this ANOVA F statistic?

2. (You may only use R as a simple calculator or to find p-values or critical values) Suppose that for $n = 6$ students, we want to predict their scores on the second quiz using scores from the first quiz. The estimated regression line is

$$\hat{y} = 20 + 0.8x.$$

   (a) For each individual observation, calculate its predicted score on the second quiz $\hat{y}_i$ and the residual e_i. You may show your results in the table below.

x_i	70	75	80	80	85	90
y_i	75	82	80	86	90	91
$\hat{y}_i$						
e_i						

(b) Complete the ANOVA table for this dataset below. **Note:** Cells with *** in them are typically left blank.

	DF	SS	MS	F-stat	p-value
Regression					0.0099
Residual				***	***
Total			***	***	***

(c) Calculate the sample estimate of the variance σ^2 for the regression model.

(d) What is the value of R^2 here?

(e) Carry out the ANOVA F test. What is an appropriate conclusion?

3. (No R required) The least squares estimators of the simple linear regression model are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{1}$$

and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}. \tag{2}$$

These are found by minimizing the sum of squared errors, i.e., minimize

$$SS_{res} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \tag{3}$$

Recall that fitted values and residuals from the fitted regression line are defined as

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i \tag{4}$$

and

$$e_i = y_i - \hat{y}_i. \tag{5}$$

Using equations (1) to (5), show that the following equalities, (6) to (9), hold:

$$\sum_{i=1}^{n} e_i = 0 \tag{6}$$

$$\sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \hat{y}_i \tag{7}$$

$$\sum_{i=1}^{n} x_i e_i = 0 \tag{8}$$

$$\sum_{i=1}^{n} \hat{y}_i e_i = 0. \tag{9}$$

Hint: Deriving the partial derivatives of the SS_{res}, (3), with respect to $\hat{\beta}_1$ and $\hat{\beta}_0$ will be useful.

Also, give a one-sentence interpretation of what the equalities (6) to (9) mean.