STAC67: Assignment 3

Deadline to hand in: **Mar. 15, 2021**
Total Points: 100

Q. 1 (22 points) 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 of severity of illness, and more anxiety. The data file, "PatientSatisfaction.txt" can be found in Quercus. Please provide variable names of the dataset.

i	:	1	2	3	$\cdots$	44	45	46
X_{i1}	:	50	36	40	$\cdots$	45	37	28
X_{i2}	:	51	46	48	$\cdots$	51	53	46
X_{i3}	:	2.3	2.3	2.2	$\cdots$	2.2	2.1	1.8
Y_i	:	48	57	66	$\cdots$	68	59	92

(a) (3 pts) Prepare a histogram for each of the predictor variables. Are any noteworthy features revealed by these plots?

(b) (3 pts) Obtain the scatter plot matrix and the correlation matrix. Interpret these and state your principal findings. Is there any concern about multicollinearity?

(c) (4 pts) Fit regression model for three predictor variables to the data and state the estimated regression function. How is $\hat{\beta}_2$ interpreted here?

(d) (4 pts) Test whether there is a regression relation; use $\alpha = 0.10$. State the alternatives, decision rule, and conclusion. What does your test imply about β_1, β_2, and β_3? What is the P-value of the test?

(e) (4 pts) Calculate the coefficient of determination, and also adjusted coefficient of determination. What does it indicate here?

(f) (4 pts) Obtain a 90 % prediction interval for a new patient's satisfaction when $X_{h1} = 35$, $X_{h2} = 45$, and $X_{h3} = 2.2$. Interpret your prediction interval.

Q. 2 (10 points) A researcher fits a multiple linear regression model, relating yield (Y) of a chemical process to temperature (X_1), and the amounts of 2 additives (X_2 and X_3, respectively). She fits the following model:

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$$

She wishes to test the following three hypotheses simultaneously:

- The mean response when $X_1 = 70,\ X_2 = 10, X_3 = 10$ is 80
- The average yield increases by 4 units when temperature increases by 1, controlling for X_2 and X_3
- The partial effect of increasing each additive is the same (controlling for all other factors)

(a) Specify following matrix and vectors that she is testing (this is her null hypothesis):

$$H_0 : \boldsymbol{K}'\underset{\sim}{\beta} - \underset{\sim}{m} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \Rightarrow$$

(b) She obtains the following results from fitting the regression based on $n = 24$ measurements while conducting the experiment:

$$(\boldsymbol{K}'\widehat{\underset{\sim}{\beta}} - \underset{\sim}{m})'(\boldsymbol{K}'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{K})^{-1}(\boldsymbol{K}'\widehat{\underset{\sim}{\beta}} - \underset{\sim}{m}) = 1800, \quad \underset{\sim}{Y}'(I - H)\underset{\sim}{Y} = 7800$$

Conduct her test at the $\alpha = 0.05$ significance level.

Q. 3 (20 points) Suppose that X is a categorical variable with 3 levels (A, B, C) and we define the indicator variable I_1 and I_2 as:

$$I_1 = \begin{cases} 1, & X = A \\ 0, & \text{otherwise} \end{cases} \qquad I_2 = \begin{cases} 1, & X = B \\ 0, & \text{otherwise} \end{cases}$$

For a continuous response variable Y consider fitting the linear model

$$Y = \beta_0 + \beta_1 I_1 + \beta_2 I_2 + \epsilon.$$

We take a total sample of n individuals. Let n_A, n_B, n_C be the number of individuals in each category of X and let $\bar{y}_A$, $\bar{y}_B$, $\bar{y}_C$ be the sample means of Y for individuals in each category of X

(a) (5 pts) Find $\boldsymbol{X}'\boldsymbol{X}$ and $\boldsymbol{X}'\underset{\sim}{Y}$.

(b) (10 pts) Show that the least squares estimates for this model are

$$\hat{\beta}_0 = \bar{y}_C, \quad \hat{\beta}_1 = \bar{y}_A - \bar{y}_C, \quad \hat{\beta}_2 = \bar{y}_B - \bar{y}_C.$$

using both options (each option is 5 points each)
(option 1) $\hat{\beta} = (X^t X)^{-1} X\mathbf{y}$.
(option 2) For any parameter values $\beta_0, \beta_1, \beta_2$ we therefore need to minimize the sum of squared errors

$$S(\beta_0, \beta_1, \beta_2) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 I_{1i} - \beta_2 I_{2i})^2.$$

(c) (5 pts) Let s_A^2, s_B^2, s_C^2 be the usual sample standard deviations of Y for individuals in each category of X. Show that the error sum of squares can be written as

$$SSE = (n_A - 1)s_A^2 + (n_B - 1)s_B^2 + (n_C - 1)s_C^2$$

Q. 4 (28 points) The data, **egyptcttn.txt** is from the article, Hamza, A.A. and Z.N. Sokkar (1981). "Brightness of Egyptian Cotton Fibers," Textile Research Journal, Vol. 51, pp. 587-590.

Following are the descrption of variables:

- Variety: 5 different cotton varieties.
- Luminance: Luminance scores (response variable)
- lnGrade: log transformed Grade score.

(a)(4 pts) Read the dataset to R and provide the variable names which are the same as above. Write down the full model with the interaction terms. Fit the full model in R.

(b) (5 pts) Test whether the slopes relating lnGrade to Luminance are the same for each cotton variety at the $\alpha = 0.05$ significance level.

(c) (5 pts) If you failed to reject the null hypothesis in (b), test whether the regression lines relating lnGrade to Luminance are the same for each cotton variety at the $\alpha = 0.05$ significance level.

(d) (4 pts) What model would you choose for this data? Justify your answer.

(e) (10 pts) For the model you chose in (d), check and comment on the standard assumptions for this model.

Q. 5 (20 points) The dataset, "StrengthWool.txt" posted with the assignment come from an experiment to understand the strength of wool under load stress as a function of three factors The length of the test specimen, the amplitude of the loading cycle to which the wool was subjected and the amount of the load. Each of these factors were under control of the experimenter and three settings for each factor was used in the experiment.

In this question, we shall not be interested in the actual values used but will assume that each is a categorical (factor) variable with three categories. The response was the number of loading cycles before the specimen failed.

(a) (5 pts) Fit a linear model with just main effects for the three class variables and show that this is not a good fit to the data.

(b) (5 pts) Add in all interactions between pairs of the class variables and refit the model. Summarize the results of this model. Does it fit better than the model in (a)?

(c) (5 pts) Using the model without interactions, look for a transformation of the response which gives a good fit.

(d) (5 pts) Show that, in the transformed scale, the model with the interactions is no better than the model with main effects only.