STAT 3008: Applied Regression Analysis
2019-20 Term 2
Assignment #2 **(Prob 5(c) revised)**

Due: April 1st, 2020 (Wednesday) at 5:30pm

This assignment covers material from Section 2.4 to 4.2 of the lecture notes.

** Please submit the hardcopy of the R-code and R-outputs for Problem 3.

You need to show your calculation in details order to obtain full scores.

Note that the solutions will be available on April 2nd (Thu) at 5pm, as the Mid-term exam will be on April 7th. No late assignment will be accepted after the solutions are posted.

Problem 1 [25 points]: Suppose simple linear regression is fitted to the data $\{(x_1, y_1), \dots (x_{20}, y_{20})\}$,

with $$E(Y \mid X = x) = \beta_0 + \beta_1 x, \qquad \mathrm{Var}(Y \mid X = x) = \sigma^2$$

The coefficient table and ANOVA table below shows some of the estimated values:

Coefficient Table

Variable	Coefficient	Std. Error	t-statistic	p-value
Constant	-23.4325	12.74	?	0.0824
X	?	0.15280	8.320	?

ANOVA Table

Source	df	SS	MS	F	p-value
Regress	1	1848.76	?	?	?
Residual	?	?	?		
Total	?	?			

(a) [14 points] Replicate the two tables above, and fill in ALL the missing values (in 5 significant figures) from the tables.

(The p-values can be obtained from R command like "> 1-pf(F_0, df1, df2)" for the right-hand tailed probability of $F_{df1,\, df2}$).

(b) [3 points] Based on the results in part (a), what is the sample correlation coefficient between x and y? That is, $r_{xy} = \hat{C}orr(x,y) = \sum(x_i - \bar{x})(y_i - \bar{y}) / \sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}$.

(c) [8 points] Based on the results in part (a), test the hypotheses on whether β_o = -10.0 at α=0.05. You should setup the 4 steps of hypothesis testing as on Ch2 page 64.

Problem 2 [17 points]: Consider the multiple linear regression:

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(p+1)}{\mathbf{X}} \ \underset{(p+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\mathbf{e}}$$, with $E(\mathbf{e}) = \mathbf{0}_{n\times 1}$ and $\mathrm{Var}(\mathbf{e}) = \sigma^2 \mathbf{I_n}$

(a) [10 points] Based on the fact that the OLS estimates $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y}$, show that

$$E(\hat{\mathbf{Y}}'\hat{\mathbf{Y}}) = \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta} + (p+1)\sigma^2$$

(b) [7 points] Based on the fact that $E(RSS) = E(\hat{\mathbf{e}}'\hat{\mathbf{e}}) = \cdots = \sigma^2(n-p-1)$ and the result from (a), show that $\sum_{i=1}^{n} E(y_i^2) = \sum_{i=1}^{n} E(\hat{y}_i^2) + \sum_{i=1}^{n} E(\hat{e}_i^2)$

Problem 3 [27 points]: Let $\mathbf{Y}$ = (21, 25, 21, 24, 9, 36, 36, 24, 10)', $\mathbf{X}_1$= (3, 9, 5, 3, -1, 7, 8, 4, 1)' and $\mathbf{X}_2$=(3, 9, 5, 3, 0, 7, 9, 4, 1)'. Suppose we want to model the response $\mathbf{Y}$ by $\mathbf{X}_1$, $\mathbf{X}_2$ and the intercept using the multiple linear regression.

(a) [12 points] Based on matrix operations in R(i.e. A%*%B, t(A), solve(A) on Ch3 page 30),

(a) show that $\hat{\boldsymbol{\beta}}$ = (11.6819, 0.32316, 2.1527)',

(b) compute the value of $\hat{\mathbf{Y}}$, $\hat{\mathbf{e}}$, SYY, RSS, SSreg, $\hat{\sigma}^2$, $\hat{\mathrm{V}}\mathrm{ar}(\hat{\boldsymbol{\beta}})$ and R^2.

(Note: In R, command like"RSS<-t(y)%%y-t(y)%*%X%*%solve(t(X)%*%X)%*%t(X)%*%y" will assign RSS as a 1x1 matrix object instead of a numeric object. You may want to use the command "as.numeric(RSS)" to bring it back to a scalar quantity.)*

(b) [5 points] Consider a new data point (x_1, x_2) = (-1, 1). What is the best point estimator for the response, and a 95% prediction interval for the response?

(c) [10 points] The ANOVA table below compares Model 1: E(Y|$\mathbf{X}$) = β_o and Model 2: E(Y|$\mathbf{X}$) = $\beta_o + \beta_2 x_2$:

Source	df	SS	MS	F_0	p-value
Regression	1	516.44	516.44	18.035	0.0038
Residual	7	200.45	28.636		
Total	8	716.89			

Suppose we want to test the hypotheses

H_o: E(Y|$\mathbf{X}$) = $\beta_o + \beta_2 x_2$ vs H_1: E(Y|$\mathbf{X}$) = $\beta_o + \beta_1 x_1 + \beta_2 x_2$

Based on the ANOVA table and the results from part (a), construct the appropriate ANOVA table. What decision and conclusion you can make from the table?

(The p-value can be obtained from R command like "> 1-pf(F_0, df1, df2)" for the right-hand tailed probability of $F_{df1,\, df2}$).

Problem 4 [11 points]: Consider data {(u_i, v_i, y_i), i = 1, 2, …, n} with $\bar{u} = \bar{v} = 0$, and $SUV = \sum_{i=1}^{n} u_i v_i = 0$. The data is fitted by a multiple linear regression with mean function

$$E(Y \mid U = u, V = v) = \beta_0 + \beta_1 u + \beta_2 v$$

(a) [6 points] Show that the OLS estimates are $\hat{\beta}_1 = SUY / SUU$, $\hat{\beta}_2 = SVY / SVV$ and $\hat{\beta}_0 = \bar{y}$.

(b) [5 points] Suppose a simple linear regression $E(Y \mid U = u) = \alpha_0 + \alpha_1 u$ is fitted to the data.

Do the OLS estimates $\hat{\alpha}_0$ and $\hat{\alpha}_1$ the same as the corresponding estimates in part (a)?

Problem 5 [20 points]: The kinetic energy of an object (y) is related with its velocity (x) through $y = \beta_0 + \beta_1 x^2 + e, \quad e \sim N(0, \sigma_0^2)$

Suppose we fit the data $\{(x_i, y_i), i = 1, \ldots, n\}$ based on $y = \alpha_0 + \alpha_1 x + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2)$.

(a) [5 points] Show that $E(\hat{\boldsymbol{\alpha}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}_2\boldsymbol{\beta}$, with $\hat{\boldsymbol{\alpha}} = \begin{pmatrix} \hat{\alpha}_0 \\ \hat{\alpha}_1 \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}$ and $\mathbf{X}_2 = \begin{bmatrix} 1 & x_1^2 \\ 1 & x_2^2 \\ \vdots & \vdots \\ 1 & x_n^2 \end{bmatrix}$

(b) [11 points] Based on the result from part (a),

(i) show that $E(\hat{\alpha}_0) = \beta_0 + \dfrac{(\overline{x^2})^2 - \bar{x}(\overline{x^3})}{\overline{x^2} - \bar{x}^2}\beta_1$

(ii) express $E(\hat{\alpha}_1)$ in terms of $\bar{x}$, $\overline{x^2}$, $\overline{x^3}$, n, β_o and β_1.

(c) [4 points] Given that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i = 0$, $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i^2 = \sigma_x^2 > 0$ and $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} x_i^3 = \kappa_x \sigma_x^3$ with $\kappa_x \neq 0$. Express $E(\hat{\alpha}_0)$ and $E(\hat{\alpha}_1)$ in terms of $\beta_0, \beta_1, \sigma_x^2$ and κ_x as $n \to \infty$. Show that (i) ~~$\hat{\alpha}_0$ is NOT a consistent estimator for β_0, and (ii) $\hat{\alpha}_1$ is NOT a consistent estimator for β_1.~~ ~~*(That is,* $\lim_{n\to\infty} \hat{\alpha}_0 \neq \beta_0$ *and* $\lim_{n\to\infty} \hat{\alpha}_1 \neq \beta_1$*)*~~

$\lim_{n\to\infty} E(\hat{\alpha}_0) \neq \beta_0$ and (ii) $\lim_{n\to\infty} E(\hat{\alpha}_1) \neq \beta_1$.

- End of the Assignment -