STAT3008: Applied Regression Analysis
2019/20 Term 2
Mid-Term Examination

Date: 7th April 2020 (Tuesday)

Time: 9:30am – 12:15pm (165 minutes)

Total Score: 100 points

- Please present your answers in 4 significant figures.
- Submission Requirement: (1) **Name and SID on the 1st page** of your work,
  (2) Only a **single file in .pdf or .doc* format** **(size < 10MB)** will be accepted
  (3) **Filename** in the format of "**LAST NAME First Name – SID.pdf/doc***"
- **How to submit your exam work?** A dropbox button is now available on Blackboard.

Problem 1 [27 points]: Suppose the following regression model is fitted to a data set with observations $\{(x_{i1}, x_{i2}, y_i), i = 1, 2, \ldots, n\}$:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + e_i, \quad e_i \overset{iid}{\sim} N(0, \sigma^2)$$

Assume that $\boxed{\sum_{i=1}^{n} x_{i1} x_{i2} = 0}$.

(a) [8 points] Derive the OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$.

(b) [6 points] Setup the log-likelihood function $l(\beta_1, \beta_2, \sigma^2)$.

(c) [4 points] Do you expect the MLE $\tilde{\beta}_1$ and $\tilde{\beta}_2$ to be the same as their corresponding OLS estimates $\hat{\beta}_1$ and $\hat{\beta}_2$ in part (a)? Explain. *(No computation required)*

(d) [5 points] Is $\hat{\beta}_1$ an unbiased estimator for β_1? Verify.

(e) [4 points] Does the point $(x_1, x_2, y) = \left(\overline{x_1^2},\ \overline{x_2^2},\ \overline{x_1 y} + \overline{x_2 y}\right) = \left(\frac{1}{n}\sum x_{i1}^2,\ \frac{1}{n}\sum x_{i2}^2,\ \frac{1}{n}\sum x_{i1} y_i + \frac{1}{n}\sum x_{i2} y_i\right)$ pass through the regression line based on the OLS estimates? Verify.

Problem 2 [16 points]: Consider multiple linear regression $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} + \mathbf{e}_{n\times 1}$ with $E(\mathbf{e}) = \mathbf{0}_{n\times 1}$ and $\mathbf{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}_n$. Let $\mathbf{A} = \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$ and $\mathbf{B} = \mathbf{I}_n - \mathbf{X}(\mathbf{X'X})^{-1}\mathbf{X'}$.

(a) [4 points] Prove or disprove the following: $\mathbf{ABA} = \mathbf{A}$.

(b) [4 points] Prove or disprove the following: $\mathbf{A}^5 = \mathbf{I}_n - \mathbf{B}^7$.

(c) [8 points] Simplify the following in terms of σ^2, n and p: $E\left[\mathbf{e'X}(\mathbf{X'X})^{-1}\mathbf{X'Y}\right]$.

Problem 3 [24 points]: A simple linear regression is fitted to the data $\{(x_1, y_1), \dots (x_{48}, y_{48})\}$, 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 regression results:

Coefficient Table

Variable	Coefficient	Std. Error	t-stat	p-value
Constant	?	5.3871	-1.8392	?
X	0.6579	?	?	?

ANOVA Table

Source	df	SS	MS	F-stat	p-value
Regression	?	?	?	?	?
Residuals	?	850.00	?		
Total	?	?			

<u>**It's known that R^2 = 15%.**</u>

(a) [16 points] Replicate the two tables above and fill in ALL the missing values (in 4 significant figures).

(b) [8 points] Based on the results in part (a), test the hypotheses on whether β_0 <u>is greater than -2.0</u> at α=0.05. You should setup the 4 steps of hypothesis testing as on Ch2 page 64.

Note: R functions like "pf", "pt", "qf" and "qt" could be useful in this problem.

Problem 4 [19 points]: Consider multiple linear regression with 3 explanatory variables (EVs) x_1, x_2 and x_3. Two hypothesis testing was performed on models with selected EVs, and the results were summarized by the two ANOVA tables below:

H_0: $E(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0$

vs H_1: $E(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$

ANOVA Table

	df	SS	MS	F-stat	p-value
Regression	2	1,132.6	566.323	2204.3	<2E-16
Residuals	51	13.10	0.25692		
Total	53	1,145.7			

H_0: $E(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$

vs H_1: $E(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

ANOVA Table

	df	SS	MS	F-stat	p-value
Regression	1	1.374	1.37427	18.027	9.44E-05
Residuals	50	3.812	0.07624		
Total	51	5.186			

It's known that the sample correlation between y and each of the x_i are 91.118%, -44.260% and 99.556% respectively. That is, $\hat{\rho}(y, x_1) = 91.118\%$, $\hat{\rho}(y, x_2) = -44.260\%$ and $\hat{\rho}(y, x_3) = 99.556\%$

(a) [11 points] Replicate the table below, and fill in ALL the missing values (in 4 significant figures). (*df and RSS of Model 7: $E(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$ have already been included in the table*)

Model	Explanatory Variable(s)	df	RSS
1	Null (No EV, constant only)	?	?
2	x1	?	?
3	x2	?	?
4	x3	?	?
5	x1, x2	?	?
6	x1, x3	?	?
7	x2, x3	51	7.3141
8	x1, x2, x3	?	?

(b) [4 points] Do you think multicollinearity exists in Model 8: $E(Y \mid \mathbf{X} = \mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$? Explain.

(c) [4 points] Do you think the sample correlation between x_1 and x_2 (i.e. $\hat{\rho}(x_1, x_2)$) is close to 0? Explain.

Problem 5 [14 points]: Suppose we are interested in explaining the sale price of a house by 4 variables relating to its size and age (grey columns below). The table below shows the data of the first 6 houses in the data set:

	Year of Construction	Area of the 1st Floor (in sq. ft)	Area of the Basement (in sq. ft)	Total Area in sq. ft (Area of 1st Floor +2nd Floor +Basement)	Sale Price: Price (in USD) of the house sold in 2010
House	Year	FirstFloor	Basement	Total	SalePrice
#1	2003	856	856	2566	208500
#2	1976	1262	1262	2524	181500
#3	2001	920	920	2706	223500
#4	1915	961	756	2473	140000
#5	2000	1145	1145	3343	250000
#6	1993	796	796	2158	143000

A multiple linear regression was fitted into $\boxed{y = \ln(\text{SalePrice})}$ based on the 4 EVs. The table below shows the parameter estimates:

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	8.947e-01	4.175e-01	2.143	0.0323 *
Year	5.231e-03	2.151e-04	24.323	< 2e-16 ***
FirstFloor	3.378e-05	3.102e-05	1.089	0.2764
Basement	-2.274e-04	2.948e-05	-7.714	2.6e-14 ***
Total	3.954e-04	1.446e-05	27.335	< 2e-16 ***

Residual standard error: 0.212 on 1169 degrees of freedom
Multiple R-squared: 0.7353, Adjusted R-squared: 0.7344
F-statistic: 811.8 on 4 and 1169 DF, p-value: < 2.2e-16

Note that most of the parameter estimates are intuitive. For example, $\hat{\beta}_{\text{Year}} = 0.005231 > 0$ is consistent with the fact that a <u>newer house (larger Year) is supposed to be sold at a higher price</u>.

(a) [12 points] Based on the parameter estimates above, comment on whether each of the following are consistent with your intuition:

(I) $\hat{\beta}_{\text{Basement}} = -0.0002274 < 0$

(II) $\hat{\beta}_{\text{Total}} = 0.0003954 > \hat{\beta}_{\text{FirstFloor}} = 0.00003378 > 0$

(b) [2 points] What is the sample size n of the data set?

- End of the Exam -