STA302/1001 - Methods of Data Analysis 1
LEC0201 Midterm - February 27, 2020

UToronto Email: ____________________

Last Name ____________________

First Name ____________________

Student ID ____________________

Instructions

1. Write your UToronto email, name and ID number at the top of this page. Make sure that these match the information in Quercus.
2. Questions are on both sides of the page. There should be a total of 9 pages.
3. Answer the questions in the spaces provided. You should not need any extra pages.
4. **Your grade will be influenced by how clearly you express your ideas, and how well you organize your solutions. You must show all your work to get full credit.**

Marking Scheme:

Question	**Out of**	**Grade**
1	8	
2	16	
3	17	
4	18	
MC	5	
Total	64	

1. **(8 pts)** An experimenter wishes to study the rate of change in a response Y when values of a predictor X are changed. The experimenter believes the relationship between the response and predictor is linear and needs help deciding which values of the predictor he should use in his experiment. The region of interest is from $X = 2$ to $X = 15$. He has enough resources to obtain 14 observations.

   (a) **(2 pts)** What values of X should the experimenter use to ensure that his estimate of the average change in the response for a unit increase in the predictor will have smallest variance? Justify your choice.

   Because the standard error of the slope is given by

$$SE(\hat{\beta}_1) = \frac{s_e}{\sqrt{SXX}} = \frac{s_e}{\sqrt{\sum_{i=1}^{10}(x_i - \bar{x})^2}}$$

   we require that the sum of squared predictors (SXX) to be large (1 point). *We can achieve this by selecting predictor values at the end of the range of possible values. Thus we would want to choose 7 observations at $x = 2$ and 7 observations at $x = 15$* (1 point)

   (b) **(2 pts)** The experimenter takes your advice and runs his experiment using your chosen predictor values. When he fits the linear model to his data, he is disappointed that the variance of the average rate of change is still quite large. Explain to him why this might be happening.

   This can be happening for two different reasons. The sample size is quite small and since

$$SE(\hat{\beta}_1) = \frac{s_e}{\sqrt{SXX}} = \frac{\sqrt{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}}{\sqrt{n-2}\sqrt{SXX}}$$

   we could be getting a large standard error because the sample size is too small (1 point) , *or we could have that the deviations of the observed responses from the regression line are too large (so the response values are too variable)*(1 point).

(c) **(2 pts)** The experimenter decides to build a 95% confidence interval for the rate of change in the response. The confidence interval he calculates contains zero so he claims this means we know for certain that there is no linear relationship between the predictor and response. Explain what is wrong with his conclusion.

The researcher has incorrectly interpreted the confidence interval. He can only say that his sample of data yielded a confidence interval that contained 0. He should have said that we are 95% confident that this sample and thus this interval is one of the 95% of all intervals that captured the true slope. (1 point) *If this is true, then we know the true slope may be 0, but it is not possible to verify for certain.* (1 point)

(d) **(2 pts)** Upon inspection of the Normal QQ plot for his data, the experimenter notices that Normality seems to be violated. He wonders how this will impact any conclusions he makes with his confidence interval. Explain how non-Normality will affect his confidence interval.

Non-normality has the potential to reduce the coverage probability of the confidence interval, as seen in assignment 2. (1 point) *So if we were to make conclusions based on this interval, such as to conclude a test, our confidence that we have captured the true value in our interval (i.e. that our interval is one of the 95% that does capture the true slope) should be lower than the 95% we want. This means we have a false sense of confidence in our sample.* (1 point)

2. (16 pts) Nutritional information from 77 different breakfast cereals have been collected by a nutritionist. The nutritionist is interested in the relationship between Calories per serving and Potassium content (in grams). The following summary statistics have already been found:

$$\sum_{i=1}^{77} x_i = 7398, \quad \sum_{i=1}^{77} y_i = 8230, \quad \sum_{i=1}^{77} x_i^2 = 1097002, \quad \sum_{i=1}^{77} x_i y_i = 783690, \quad s_e = 19.57$$

(a) **(3 pts)** What is the fitted regression line for the relationship between Calories and Potassium?

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\\ &= \frac{783690 - 77(7398/77)(8230/77)}{1097002 - 77(7398/77)^2}\\ &= -0.018 \quad \text{(1 point)}\end{aligned}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \frac{8230}{77} + 0.018\left(\frac{7398}{77}\right) = 108.61 \quad \text{(1 point)}$$

So the fitted regression line is $\hat{y} = 108.61 - 0.018x$ (1 point)

(b) **(5 pts)** Determine if the average calories per serving significantly increases when the potassium content increases by 1 gram. Use a formal hypothesis test, including appropriate hypotheses, test statistic, p-value and conclusion in the context of the data.

The hypotheses are

$H_0 : \beta_1 = 0 \quad \text{vs} \quad H_a : \beta_1 > 0$ (1 point, 0 if betas have hats or if alternative is wrong)

To build the test statistic, we need the standard error of the slope:

$$SE(\hat{\beta}_1) = \frac{s_e}{\sqrt{SXX}} = \frac{19.57}{\sqrt{386217.5}} = 0.031 \quad \text{(1 point)}$$

so the test statistic is

$$T = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)} = \frac{-0.018}{0.031} = -0.58 \quad \text{(1 point)}$$

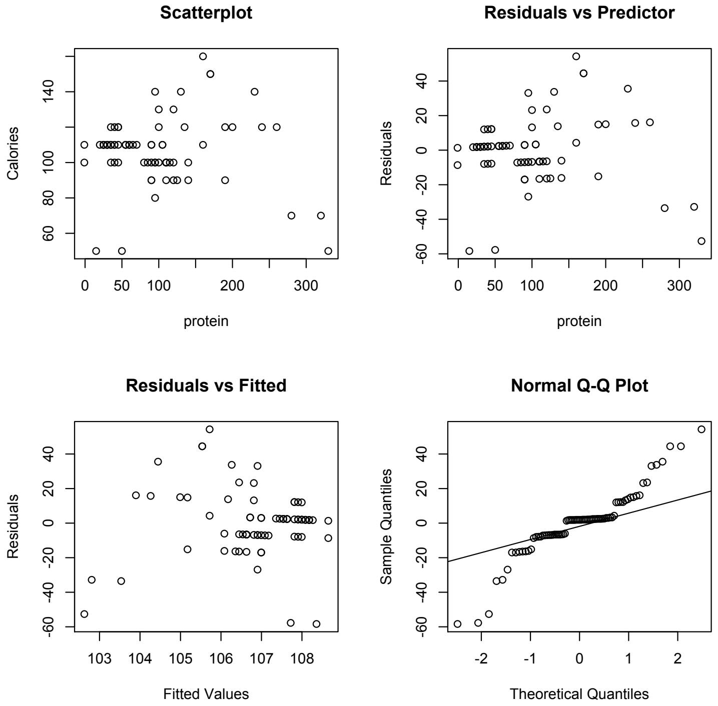
The p-value will require comparing to a T distribution with 75 degrees of freedom (this is not in your table, so you will need to approximate its location)

$$p - value = P(T_{75} > -0.58) \in [0.7, 0.8] \quad \text{(1 point)}$$

because 0.58 is located in between 0.52 and 0.85, which for a one-tailed test means the p-value has to be between 0.7 and 0.8.

Thus we fail to reject the null hypothesis and conclude that mean calories does not significantly increase as potassium increases by 1 gram. (1 point only if they use the context of the data)

(c) **(3 pts)** Based on the plots below, discuss whether each regression modelling assumption is satisfied and if there are any problematic observations.



*** if justifications seem reasonable and aren't completely contradictory to the plots, they can receive full marks.

- *Linearity: Based on the scatterplot, we seem to say linearity should hold, although there seems to be a group of 3 points that may make it appear as if there is a slight curvature* (0.5 point)
- *Independence: There do appear to be some evidence of grouping in both residual plots, with a large cluster in the top left area and a smaller cluster with large negative residuals, so independence may be violated* (0.5 point)
- *Constant variance: The three points in the lower right corner give the impression of a curvature problem. Without these, there still appears to be a linear pattern so it appears constant variance is violated.* (0.5 point)
- *Normality: The lifting in the tails suggests that normality is violated.* (0.5 point)
- *problematic observations: there seems to be a group of 3 observations that have very high predictor values but very small calories that may be influential* (1 point)

(d) **(2 pts)** Provide two possible methods that can be used to correct non-constant variance..

One can transform the response variable and refit the model. (1 point) *Alternatively, one may use weighted least squares regression.* (1 point)

(e) **(3 pts)** Find a variance stabilizing transformation when the response follows a distribution with a mean of θ and a variance of θ^3. Show your work.

Use the Delta Method: $Var(f(E[Y])) = [f'(E[Y])]^2 Var(Y)$
Here the mean of Y *is* θ *and the variance is* θ^3*, so we can write*

$$\begin{aligned} Var(f(\theta)) &= [f'(\theta)]^2\theta^3 = c \quad \text{(1 point)} \\ \Rightarrow [f'(\theta)]^2 &= \frac{c}{\theta^3} \\ \Rightarrow f'(\theta) &= \frac{\sqrt{c}}{\theta^{3/2}} \quad \text{(1 point)} \\ \Rightarrow f(\theta) &= \int \frac{\sqrt{c}}{\theta^{3/2}} d\theta = -2\frac{\sqrt{c}}{\sqrt{\theta}} + c^* \quad \text{(1 point)} \end{aligned}$$

which tells us that an inverse square root transformation will result in constant variance.

3. **(17 pts)** Suppose we have collected a random sample of n pairs (x_i, y_i) from a population where the true relationship between the response and predictor is

$$Y = \beta_0 + \beta_1 X + \epsilon, \quad \text{where } \epsilon \mid X = x \sim N(x, \sigma^2)$$

(a) **(2 pts)** What is the true population average response when $X = x$?

$E(Y \mid X) = E(\beta_0 + \beta_1 X + \epsilon)$(1 point) $= \beta_0 + \beta_1 X + E(\epsilon) = \beta_0 + \beta_1 X + x$ (1 point)

(b) **(5 pts)** Derive the least squares estimators of the intercept and slope for this population regression line.

The least squares estimating equation we use is

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

Take the first derivative with respect to both parameters:

$$\frac{\partial}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0 \quad \text{(1 point)}$$

$$\frac{\partial}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0 \quad \text{(1 point)}$$

Solve for each of the parameters:

$$0 = \sum_{i=1}^{n} x_i y_i - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2$$

$$\Rightarrow \sum_{i=1}^{n} x_i y_i = (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^{n} x_i + \beta_1 \sum_{i=1}^{n} x_i^2 = n\bar{y}\bar{x} - \beta_1 \left[\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 \right] \quad \text{(1 point)}$$

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

$$\Rightarrow \hat{\beta}_0 = \sum_{i=1}^{n} \frac{y_i}{n} - \hat{\beta}_1 \sum_{i=1}^{n} \frac{x_i}{n} = \bar{y} - \hat{\beta}_1 \bar{x} \quad \text{(1 point)}$$

(c) **(6 pts)** Determine whether the least squares estimator of the slope from part (b) is unbiased. If not, is it possible to collect data in such a way as to get an unbiased estimate of the intercept? Justify your answer.

To show unbiasedness, we must take the expectation of the estimator:

$$
\begin{aligned}
E[\hat{\beta}_1] &= E\left[\frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\right] \quad \text{(1 point)} \\
&= \frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\left\{\sum_{i=1}^n x_i E[y_i] - n\bar{x}E[\bar{y}]\right\} \quad \text{(1 point)} \\
&= \frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\left\{\sum_{i=1}^n x_i(\beta_0 + \beta_1 x_i + E[\epsilon]) - n\bar{x}(\beta_0 + \beta_1\bar{x} + E[\bar{\epsilon}])\right\} \quad \text{(1 point)} \\
&= \frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\left\{\beta_1\sum_{i=1}^n x_i^2 - n\bar{x}^2 + \sum_{i=1}^n x_i(x_i) - n\bar{x}(\bar{x})\right\} \quad \text{(1 point)} \\
&= \frac{1}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}\left\{\left[\sum_{i=1}^n x_i^2 - n\bar{x}^2\right](\beta_1 + 1)\right\} \\
&= (\beta_1 + 1)\frac{\sum_{i=1}^n x_i^2 - n\bar{x}^2}{\sum_{i=1}^n x_i^2 - n\bar{x}^2} = \beta_1 + 1 \quad \text{(1 point)}
\end{aligned}
$$

So the estimate of the slope is biased. Since it is biased by a constant term, there is nothing we can do in our data collection that will remove this bias. (1 point)

(d) **(4 pts)** Determine whether the least squares estimator of the intercept from part (b) is unbiased.

Again to show unbiasedness, we take the expectation of the estimator:

$$\begin{aligned} E[\hat{\beta}_0] &= E[\bar{y} - \hat{\beta}_1 \bar{x}] \quad \text{(1 point)} \\ &= E[(\beta_0 + \beta_1 \bar{x} + \bar{\epsilon})] - \bar{x} E[\hat{\beta}_1] \quad \text{(1 point)} \\ &= \beta_0 + \beta_1 \bar{x} + E[\epsilon_i] - \bar{x}(\beta_1 + 1) \quad \text{(1 point)} \\ &= \beta_0 + \bar{x}(\beta_1 + 1) - \bar{x}(\beta_1 + 1) = \beta_0 \quad \text{(1 point)} \end{aligned}$$

4. **(18 pts)** A company that sells photocopiers to businesses also provides maintenance and repairs on them when needed. The company keeps records of the maintenance calls. For each call, they record the years of experience of the serviceperson (X) and the total number of minutes spent working on the machine by a serviceperson (Y). The average years of experience of the servicepeople is 3.98 and the average number of minutes spent working on a call is 76.27. The company models the relationship between number of copiers serviced and repair time with a simple linear regression, with the results seen below.

```
Call:
lm(formula = Y ˜ X, data = copier)

Residuals:
    Min      1Q  Median      3Q     Max
-79.827 -34.873  -4.611  35.650  79.650

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   61.395     20.256   3.031  0.00412 **
X             3.739      4.831   0.774  0.44328
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 42.94 on 43 degrees of freedom
Multiple R-squared:  0.01373,Adjusted R-squared:  -0.009203
F-statistic: 0.5988 on 1 and 43 DF,  p-value: 0.4433
```

Using this R output, answer the following questions:

(a) **(1 pts)** Interpret the coefficient of determination in the context of the data..

1.37% of the variation in total number of minutes spent working on the photocopier is explained by the years of experience of the serviceperson (or alternatively by the model). (1 point, but 0 points if they don't interpret in context of data.)

(b) **(3 pts)** Complete the ANOVA table below. Show all your work.

Source	DF	Sum Squares	Mean Squares	F value
Regression	1	1104.09	1104.09	0.5988
Residual	43	79285.27	1843.84	
Total	44	80389.36		

Degrees of freedom: 1 for Regression (because always 1 in simple linear regression), 43 for Residuals (from DF of F test in output), 44 for Total (by decomposition) (0.25 point if all 3 are correct, 0 otherwise)

F value is easily found in R output under F statistic: 0.5988 (0.25 points)

Residual sum of squares in R output:
$s_e = 42.94 = \sqrt{\frac{RSS}{n-2}} \Rightarrow RSS = (n-2)s_e^2 = 43(42.94)^2 = 79285.27$ (0.5 points)

Mean squares residual: $s_e^2 = \frac{RSS}{n-2} = \frac{79285.27}{43} = 1843.84$ (0.5 point)

Mean squares regression:
$F = \frac{MSreg}{MSres} \Rightarrow MSreg = F(MSres) = 0.5988(1843.84) = 1104.09$ (0.5 point)

Sum of Squares Regression $=$ *Mean squares regression* $= 1104.09$ (0.5 point)

Total sum of squares: $SST = SSreg + RSS = 1104.09 + 79285.27 = 80389.36$ (0.5 point)

(c) **(5 pts)** How long would we expect the actual repair/maintenance time (in total number of minutes) to be for a serviceperson with 6 years of experience? Compute an appropriate 95% interval using the R output above.

$\hat{y}^* = 61.395 + 3.739(6) = 83.829$ (1 point)

In order to compute the standard error of the prediction, we need to find SXX, which we can do using the standard error of the estimate of the slope:

$$SE(\hat{\beta}_1) = \frac{s_e}{\sqrt{SXX}} \Rightarrow SXX = \frac{s_e^2}{SE(\hat{\beta}_1)^2} = \frac{42.94^2}{4.831^2} = 79.004 \quad \text{(1 point)}$$

So the standard error of the prediction of an actual value is

$$\begin{aligned} SE(\hat{y}^*) &= s_e\sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{SXX}} \\ &= 42.94\sqrt{1 + \frac{1}{45} + \frac{(6 - 3.98)^2}{79.004}} \\ &= 44.498 \quad \text{(1 point)} \end{aligned}$$

The closest t-value in the table would be approximately $t_{0.975,43} = 2.021$ (1 point) *so the 95% prediction interval is*

$$\hat{y}^* \pm t_{0.975,43} SE(\hat{y}^*) = 83.829 \pm 2.021(44.497) = [-6.101, 173.76] \quad \text{(1 point)}$$

(d) **(2 pts)** Is there a strong linear relationship between years of experience of the serviceperson and the total number of minutes spent working on the photocopier by a serviceperson? Justify your conclusion by using two pieces of information from the R output (do not use the same number more than once).

No there is not a strong linear relationship. The coefficient of determination tells us that only 1.3% of the variation in total number of minutes spent by a serviceperson is explained by the model (1 point). Further the p-value in the overall test of significance is large, which indicates a non-significant linear relationship exists between total number of minutes and years of experience of the serviceperson (could also use the conclusion of the t-test) (1 point).

(e) **(7 pts)** Consider a maintenance call where the serviceperson had 6 years of experience and only spent a total of 4 minutes on the call. Determine whether this call is a leverage point, an outlier, an influential point or some combination of these. Justify your answer with appropriate numerical summaries.

The predicted value for this service call is $\hat{y} = 61.395 + 3.739(6) = 83.829$ (1 point)

The estimated residual for this observation is $\hat{e}_i = y_i - \hat{y}_i = 4 - 83.829 = -79.83$ (1 point)

The leverage value for this observation is

$$h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{SXX} = \frac{1}{45} + \frac{(6 - 3.89)^2}{79.004} = 0.079 \quad \text{(1 point)}$$

The cutoff for a leverage point is $4/n = 4/45 = 0.089$ (0.5 point for cutoff)*so this observation is not a leverage point.*

The standardized residual is

$$r_i = \frac{\hat{e}_i}{s_e\sqrt{1 - h_{ii}}} = \frac{-79.83}{42.94\sqrt{1 - 0.079}} = -1.93 \quad \text{(1 point)}$$

Since this value is within [-2,2], this observation is also not an outlier.

The Cook's Distance is

$$D_i = \frac{r_i^2}{2}\frac{h_{ii}}{1 - h_{ii}} = \frac{(-1.93)^2}{2}\frac{0.079}{1 - 0.079} = 0.16 \quad \text{(1 point)}$$

Since this value is large (i.e. larger than $4/(n-2) = 0.09$ *(0.5 point for cutoff)), this observation is an influential point.* (1 point)

(5 pts) Multiple Choice Section: Please answer the multiple choice questions on the attached bubble sheet, corresponding to questions 1-5. A correct response is one point while an incorrect or no response is worth 0 points. If you do not answer the questions on the bubble sheet, your answers will not be graded.

1. TRUE or FALSE: an influential observation may not always be a leverage point.

   (a) **True**

   (b) False

2. If the correlation between the predictor and response is 0, the predicted response from the least squares regression will be equal to...

   (a) zero

   (b) the average difference between the response and predictor.

   (c) sample mean of the predictor

   (d) **sample mean of the response**

3. Which of the following statements is incorrect?

   (a) In small samples, residuals may appear to be Normal, even when the population errors are not.

   (b) **If a bad leverage point is present, it is always a good idea to discard it from the data.**

   (c) Violations of model assumptions are more likely if high leverage points are present

   (d) All of the above

   (e) None of the above

4. Which term corresponds to the following definition?
   The variation in the response variable explained by the model.

   (a) **Regression Sum of Squares**

   (b) Coefficient of Determination

   (c) Residual Mean Squares

   (d) Sample Variance

5. In what situation would we not trust the coefficient of determination to properly assess the goodness of the simple linear model?

   (a) When Normality of errors appears to be violated

   (b) When we have an indicator predictor variable.

   (c) **When a curved pattern is visible in a scatterplot of the data**

   (d) All of the above