STA302 - Methods of Data Analysis 1
LEC0201 Midterm - October 26, 2022

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 5 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.**

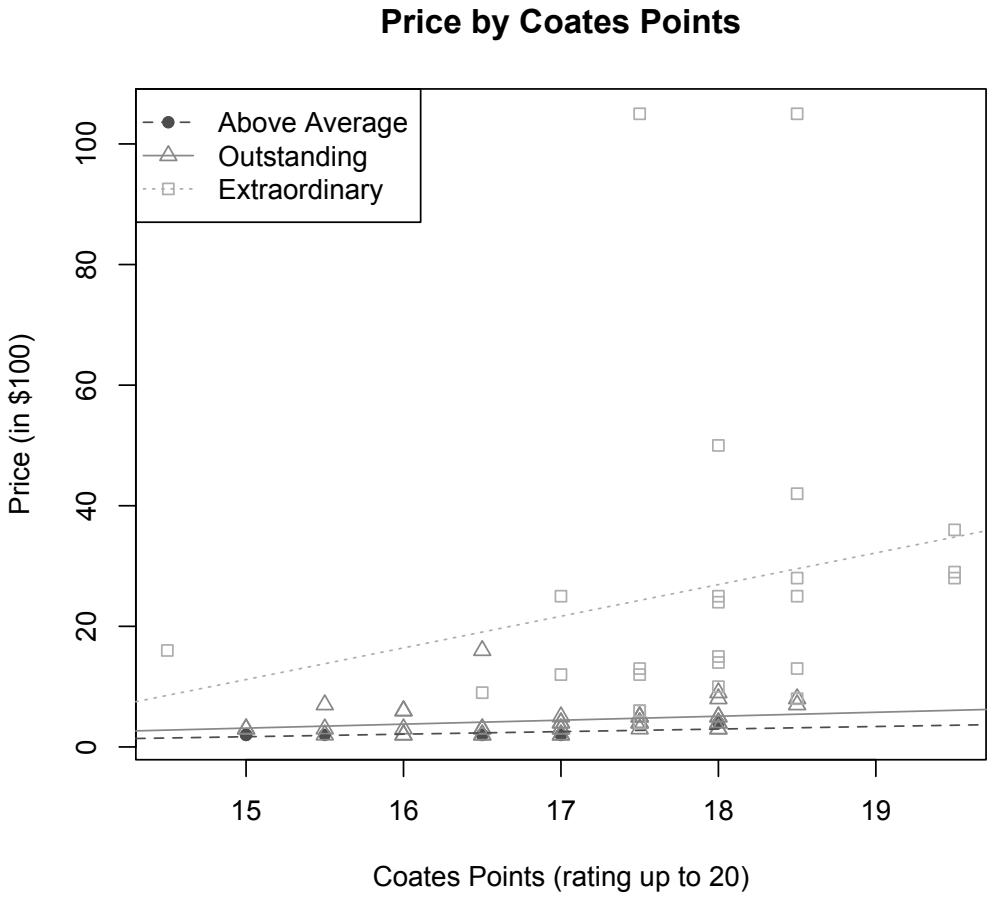
Marking Scheme:

Question	Out of	Grade
1	9	
2	8	
3	13	
Total	30	

1. (9 pts) Data was collected on 72 wines from the Bordeaux region of France. Of interest is understanding how various ratings of wine quality can affect the Price (Y) of a bottle of wine.

Suppose we first want to study the relationship between Price (in \$100) and the Coates Points, accounting for a classification of the wine based on Parker Points, where

- Coates Points = a 20-point rating system where a higher score is better
- Parker Points = a rating from 0 to 100 where we classify wines as:
  - Above Average = a rating between 80 and 89
  - Outstanding = a rating between 90 and 95
  - Extraordinary = a rating between 95 and 100



(a) (2 pts) Given the graph above, write out a single linear regression relationship that represents the trend we see between Price and the Coates Points rating for the different classifications of wine based on Parker Points. Note that there are many points that are overlapping in the bottom of the graph.

We could write the relationship presented in the graph in two ways, depending on whether we use Parker Points as a categorical variable (the way we would if we were using R) or split it up into dummy/indicator variables:

$$y_i = \beta_0 + \beta_1 Coates + \beta_2 Coates * Parker + \beta_3 Parker + \epsilon_i$$

or

$$\begin{aligned} y_i = \beta_0 + \beta_1 Coates + \beta_2 Coates * \mathbb{1}\{Parker = Extraordinary\} \\ + \beta_3 \mathbb{1}\{Parker = Extraordinary\} + \beta_4 \mathbb{1}\{Parker = Outstanding\} + \epsilon_i \end{aligned}$$

1 point for writing the equation such that we see different intercepts for all Parker classification, and 1 point for writing the equation so that at minimum Extraordinary has a different slope from the other 2 groups. Students may have chosen a different baseline level for the second expression so assess based on what they have written whether this would yield different estimates of slopes and intercepts that match the graph.

For use in parts b-d below:

Consider now the fitted linear regression model presented below which looks at the relationship between Price and two new indicator variables: CultWine which takes value 1 if the wine is limited availability and 0 if not, and FirstGrowth where a value of 1 indicates the wine achieves the highest classification for wines, while a value of 0 means is has not achieved this classification.

```
Call:
lm(formula = Price ~ as.factor(CultWine) +
as.factor(FirstGrowth):as.factor(CultWine), data = red)

Residuals:
    Min      1Q  Median      3Q     Max
-32.750  -2.612  -1.362   1.493  56.250

Coefficients:
                        Estimate Std. Error t value    Pr(>|t|)
(Intercept)               5.362      1.296   4.136   9.93e-05 ***
CultWine1                 43.388     5.104   8.501   2.70e-12 ***
CultWine0:FirstGrowth1    21.193     3.537   5.992   8.79e-08 ***
CultWine1:FirstGrowth1    56.250     11.038  5.096   2.97e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.873 on 68 degrees of freedom
Multiple R-squared:  0.7363,Adjusted R-squared:  0.7247
F-statistic:  63.3 on 3 and 68 DF,  p-value: < 2.2e-16
>
>
> X <- model.matrix(mod)
> round(solve(t(X) \%*\% X), 3)
(Intercept) CultWine1 CultWine0:FirstGrowth1   CultWine1:FirstGrowth1
 0.017         -0.017            -0.017                      0.00
 -0.017         0.267             0.017                     -0.25
 -0.017         0.017             0.128                      0.00
 0.000         -0.250             0.000                      1.25
```

(b) (2 pts) Based on this model, what is the interpretation of the slope on the interaction term between when CultWine = 1 and First Growth = 1?

$\hat{\beta}_3 = 56.25$ is the mean Price for a wine that is a limited edition (Cultwine = yes) and also a First Growth wine.

To obtain the full 2 marks, the interpretation must include:

- *average Price*
- *Cultwine = yes*
- *First Growth = yes*

If at least 1 is missing from the interpretation, they can only receive 1 mark maximum.

(c) (5 pts) Find a 95% prediction interval for a wine that is classified as a Cult Wine but not a First Growth wine.

We are predicting an actual response when $\mathbf{x}_0' = (1, 1, 0, 0)$ because when considering a wine that is not First Growth, any term involving it will be 0. Then

$$\hat{y}_0 = 5.362 + 43.388(1) + 0 + 0 = 48.75$$

$$s^2 = 9.873^2 = 97.48$$

To find $var(\hat{y}_0) = s^2(\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0')$ so

$$\mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0' = \begin{pmatrix} 1 & 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} 0.017 & -0.017 & -0.017 & 0 \\ -0.017 & 0.267 & 0.017 & -0.25 \\ -0.017 & 0.017 & 0.128 & 0 \\ 0 & -0.25 & 0 & 1.25 \end{pmatrix} \begin{pmatrix} 1 \\ 1 \\ 0 \\ 0 \end{pmatrix} = 0.25$$

So

$$var(\hat{y}_0) = s^2(1 + \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0') = 97.48(1.25) = 121.85$$

Then the 95% prediction interval is

$$48.75 \pm 2\sqrt{121.85} = [26.67, 70.83]$$

- *1 point for correctly finding* $\mathbf{x}_0'$
- *1 point for predicted response*
- *1 point for estimated error variance*
- *1 point for finding variance of prediction*
- *1 point for correctly finding prediction interval.*

If a calculation error has occurred in a previous step, do not penalize subsequent answers if they used the incorrect value. Instead, check that no new additional calculation error has occurred and only in this case, deduct another point.

2. (8 pts) Suppose we wish to estimate the following linear regression relationship in the population:

$$y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i, \quad i = 1, \ldots, n,$$

where x_i^2 is the predictor X squared. You may assume that the assumptions of linear regression hold and that this is the true relationship in the population.

(a) (4 pts) Derive the least squares estimators for the coefficients β_0 and β_1 in the above relationship.

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

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

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

$$(1) \Rightarrow \sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i^2 \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i^2}{n}$$

$$(2) \Rightarrow \sum_{i=1}^{n} x_i^2 y_i = \beta_0 \sum_{i=1}^{n} x_i^2 + \beta_1 \sum_{i=1}^{n} x_i^4$$

$$= \sum_{i=1}^{n} x_i^2 (\bar{y} - \beta_1 \frac{\sum_{i=1}^{n} x_i^2}{n}) + \beta_1 \sum_{i=1}^{n} x_i^4$$

$$= \bar{y} \sum_{i=1}^{n} x_i^2 - \beta_1 \frac{(\sum_{i=1}^{n} x_i^2)^2}{n} + \beta_1 \sum_{i=1}^{n} x_i^4$$

$$\Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i^2 y_i - \bar{y} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n} x_i^4 - \frac{(\sum_{i=1}^{n} x_i^2)^2}{n}}$$

- *1 point for correct partial derivative for intercept*
- *1 point for correct partial derivative for slope*
- *1 point for finding correct, simplified estimate of intercept*
- *1 point for finding correct, simplified estimate of slope*

If a calculation/derivation error has occurred in a previous step, do not penalize subsequent answers if they used the incorrect value. Instead, check that no new additional calculation error has occurred and only in this case, deduct another point.

(b) (2 pts) Check whether the estimator of β_1 in the above relationship is unbiased.

$$
\begin{aligned}
E(\hat{\beta}_1 \mid X) &= \frac{1}{\sum_{i=1}^n x_i^4 - \frac{(\sum_{i=1}^n x_i^2)^2}{n}} \left[\sum_{i=1}^n x_i^2 E(y_i \mid X) - \sum_{i=1}^n x_i^2 \frac{1}{n} \sum_{i=1}^n E(y_i \mid X) \right] \\
&= \frac{1}{\sum_{i=1}^n x_i^4 - \frac{(\sum_{i=1}^n x_i^2)^2}{n}} \left[\sum_{i=1}^n x_i^2 (\beta_0 + \beta_1 x_i^2) - \sum_{i=1}^n x_i^2 \frac{1}{n} \sum_{i=1}^n (\beta_0 + \beta_1 x_i^2) \right] \\
&= \frac{1}{\sum_{i=1}^n x_i^4 - \frac{(\sum_{i=1}^n x_i^2)^2}{n}} \beta_1 \left[\sum_{i=1}^n x_i^4 - \frac{(\sum_{i=1}^n x_i^2)^2}{n} \right] = \beta_1
\end{aligned}
$$

1 point for correctly using population relationship, and 1 point for arriving at a simplified answer. If a calculation/derivation error has occurred in part a) or in a previous step, do not penalize subsequent answers if they used the incorrect value/expression. Instead, check that no new additional calculation error has occurred and only in this case, deduct another point.

(c) (2 pts) Below is an explanation of the concept of unbiasedness:

An estimator is unbiased when the expected value of the estimator equals the parameter being estimated..

There are two parts in this sentence that would be difficult for a layman (i.e. an audience with no knowledge of statistics) to understand. What are the two parts of this sentence that are overly technical? Rewrite these two parts in a less technical way.

- *expected value of the estimator* = the average value of our best guess across many many samples of data
- *parameter being estimated* = the true value of the unknown quantity

1 point for each rephrasing using non-technical language. As long as there is a reasonable attempt to transition statistical words into a more conceptual explanation then they can receive the mark.

3. (13 pts) A study was conducted to understand the relationship between the Amazon Price (Y) of Lego sets between 2018 and 2020 and two characteristics of these sets: the number of mini figures in each set (X_1), and the weight in kilograms of each set (X_2). A sample of 36 sets has been collected.

The linear regression model they wish to estimate is

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \epsilon_i, \quad i = 1, \ldots, 36.$$

The researchers have produced the following summaries of their data:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 0.084 & -0.010 & -0.037 \\ -0.010 & 0.004 & -0.006 \\ -0.037 & -0.006 & 0.080 \end{pmatrix}, \quad \mathbf{X}'\mathbf{Y} = \begin{pmatrix} 1821 \\ 7468 \\ 1984 \end{pmatrix},$$

as well as the following plots:

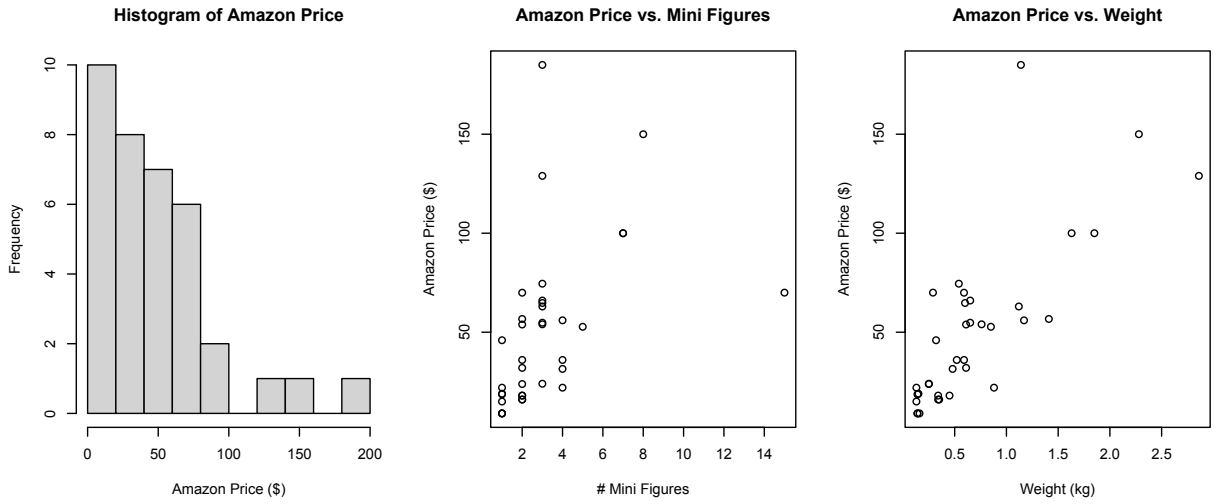


Figure 1: The Amazon price of the Lego sets in the sample of 36 has a mean of 50.58 and a sample standard deviation of 40.78. The average number of Mini Figures in the sample is 3.14 with a standard deviation of 2.67, while the mean and standard deviation of weight is 0.70 and 0.63 kilograms respectively.

(a) (3 pts) Compute the least squares estimates of the coefficients in the above model.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} = \begin{pmatrix} 0.084 & -0.010 & -0.037 \\ -0.010 & 0.004 & -0.006 \\ -0.037 & -0.006 & 0.080 \end{pmatrix} \begin{pmatrix} 1821 \\ 7468 \\ 1984 \end{pmatrix} = \begin{pmatrix} 4.876 \\ -0.242 \\ 46.535 \end{pmatrix}$$

1 point for transcribing the matrices correctly, 2 points for correctly estimated coefficients. If a transcription error occurred but calculation is correct otherwise, they should receive their 2 points for a correct answer but lose the transcription point. If a calculation error has occurred resulting in at most 1 coefficient being incorrectly estimated, then they should still receive 1 point for the two correct estimates.

(b) (3 pts) Based on the exploratory data analysis presented, would we suspect that any of the assumptions of this linear regression model could possibly be violated? Why? *(hint: be sure to comment on linearity, Normality and constant variance.)*

- Linearity: based on the two scatterplots, we don't see any alarming non-linear relationships so we are not concerned about this assumptions
- Normality: could be suspected to be violated because the histogram of Price is very skewed
- Constant variance: may be violated slightly because Price vs minifigures exhibits different spreads as the predictor changes value.

1 point for each assumption with a reasonable assessment with a specific reference to a plot. As long as conclusion for linearity and constant variance is reasonable based on their justification, we can accept a conclusion of satisfied or not, but justification must be sound.

(c) (6 pts) Complete the ANOVA table below, showing all your work.

Source	DF	Sum Squares	Mean Squares	F value
Regression	$p = 2$	37682.86	18841.43	30.30
Residual	$n-p-1 = 33$	20522.43	621.89	
Total	$n-1 = 35$	58205.29		

Given in question:

$$n = 36, \quad s_y = 40.78, \quad p = 2$$

The work that results in the numbers in the table:

$$\begin{aligned} SST &= (n-1)s_y^2 = (35)(40.78)^2 = 58205.29 \\ SSreg &= 2 \times MSreg = 2(18841.43) = 37682.86 \\ RSS &= SST - SSreg = 58205.29 - 37682.86 = 20522.43 \\ MSR &= \frac{RSS}{n-p-1} = \frac{20522.43}{33} = 621.89 \\ F &= \frac{MSreg}{MSR} = \frac{18841.43}{621.89} = 30.30 \end{aligned}$$

- *1 point for all 3 correct DF (0 if at least one is incorrect)*
- *1 point for correct regression sum of squares calculation*
- *1 point for correct sum of squares total calculation*
- *1 point for correct residual sum of squares calculation*
- *1 point for correct mean squares residual calculation*
- *1 point for correct F statistic calculation*

If a calculation error has occurred in a previous step, do not penalize subsequent answers if they used the incorrect value. Instead, check that no new additional calculation error has occurred and only in this case, deduct another point.

(d) (1 pts) A partial F test is conducted to determine whether it is possible to remove the predictor Weight from the model. A test statistic of 40.854 is found. What is the conclusion of this test at a significance level of 0.05?

$$F = 40.854, \quad H_0 : \beta_2 = 0, H_a : \beta_2 \neq 0$$

We find from the F table that $F_{0.95}(1, 33) \approx 4.17$ and since $F > 4.17$, we would reject the null and conclude that it would not be possible to drop weight from the model.

To obtain the point for this question, students must refer to the correct critical value and correctly reject the null and say what that means.