University of Toronto

Faculty of Arts and Science

2024 Makeup Mid-term Exam

STA302 Methods of Data Analysis 1

Duration: 1 hour 40 minutes

Aids Allowed: Basic calculator and provided formula sheet

Instructions:

1. Correctly fill out the front page.
2. The exam is printed on both sides of the pages. There are a total of 12 exam pages.
3. Answer questions in the spaces provided on the question sheet. There is a blank page for scratch work. Answers written on the blank page will not be graded.
4. **Do not remove any pages from this booklet.**
5. You must show all of your work to receive full credit. Your grade is influenced on how clear you express your ideas and organization of your solutions.

Marking scheme

Question	Points	Score
1	12	
2	11	
3	7	
Total:	30	

Problem 1

1. We will look at a misspecified linear regression model with non-normal and correlated errors. Consider, for $i = 1, \dots, n$, the population model $Y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$, where

$$E(\boldsymbol{\epsilon}) = \boldsymbol{X}c, \qquad \text{Cov}(\boldsymbol{\epsilon}) = \sigma^2 C \qquad \epsilon_i \text{ are not normal.}$$

where

$$c = (1, \dots, 1)^T \qquad C = \begin{pmatrix} 1 & 1/2 & 0 & \cdots & \cdots & 0 \\ 1/2 & 1 & 1/2 & \cdots & \cdots & 0 \\ 0 & 1/2 & 1 & 1/2 & \cdots & 0 \\ 0 & \vdots & \vdots & \vdots & \vdots & 1/2 \\ 0 & 0 & \cdots & \cdots & 1/2 & 1 \end{pmatrix}.$$

Here $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ is the vector containing errors. However, despite mispecification of the population, the standard least squares estimator is used:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

Here $\boldsymbol{Y} = (Y_1, \dots, Y_n)^T$ is from the misspecified population.

(a) (4 points) Determine if the unbiased property of the least squares estimator $\hat{\boldsymbol{\beta}}$ changes due to the misspecified regression model. You must derive the answer showing/explaining all of your work and must state precisely where you use any misspecified regression modeling assumption.

Solution:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= E\left((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}|\boldsymbol{X}\right) && \text{(definition)} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E\left(\boldsymbol{Y}|\boldsymbol{X}\right) && \text{(linearity of the expectation)} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\left(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}c\right) && \text{(population/regression model)} \\ &= \boldsymbol{\beta} + c. \end{aligned}$$

Rubric:

- No credit for no work shown. Partial credit for showing work even if there are some errors.
- 1 point for applying expectation to LS estimator

- 1 point for removing constants from expectation
- 1 point for using population model (-0.5 if they did not note this use as instructed in question)
- 1 point for simplifying to final answer.

(b) (5 points) Determine if the covariance of the least squares estimator $\hat{\boldsymbol{\beta}}$ changes due to the misspecified regression model. You must derive the answer showing/explaining all of your work and must state precisely where you use any misspecified regression modeling assumption.

Solution:

$$\begin{aligned}
Cov(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} && \text{(properties of covariance)}\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{e})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} && \text{(population/regression model)}\\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \sigma^2 C\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}. && \text{(population covariance assumption).}
\end{aligned}$$

Rubric:

- No credit for no work shown. Partial credit for showing work even if there are some errors.
- 1 point for applying covariance to estimator
- 1 point for correctly removing constants from covariance
- 1 point for correctly replacing Y with the correct population/regression model result
- 1 point for utilizing population assumption (-0.5 if they did not note this as instructed in question)
- 1 point for simplifying and noting that this is not equal to covariance of usual situation.

(c) (3 points) Based on the results of part (a) and (b) as well as the information in the question, how does the sampling distribution of $\hat{\boldsymbol{\beta}}$ in this misspecified model compare to the sampling distribution we would expect under a correctly specified model?

Solution: The sampling distribution of $\hat{\boldsymbol{\beta}}$ in the misspecified model will have a different mean and a different covariance than a correctly specified model. In addition, it will not be Normally distributed.

Rubric: If errors occurred in parts (a) or (b) and they do not obtain the correct answers, award 1 point for the mean and 1 point for the variance as long as the answer is correct based on what they derived in (a) and (b).

- 1 point for noting they would not have the same mean
- 1 point for noting the misspecified model will a different covariance.
- 1 point for noting it won't be Normal

Problem 2

2. Data was collected on LEGO sets for sale between January 1, 2018 and September 11, 2020. The data was restricted to 189 sets comprising three themes: City sets, Marvel sets and Disney sets. The number of pages in the manual and the year of sale of each set were also collected. A model was fit to these data, with the number of pieces in a set as the response, and with predictors pages, year of sale and theme, and an interaction between year and theme. Below is the model output:

```
Call:
lm(formula = pieces ~ pages + theme + year + year:theme,
    data = d2)

Residuals:
    Min      1Q  Median      3Q     Max
-291.10  -67.02  -14.74   50.58  965.19

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)            -13.7342    28.4612  -0.483  0.63000
pages                    3.0304     0.1268  23.891  < 2e-16 ***
themeDisney™            32.2360    49.5671   0.650  0.51630
themeMarvel            127.1881    43.1506   2.948  0.00363 **
year2019               -60.9644    34.5360  -1.765  0.07923 .
year2020              -107.9736    36.0664  -2.994  0.00315 **
themeDisney™:year2019   17.0090    63.7667   0.267  0.78998
themeMarvel:year2019    54.4428    61.3764   0.887  0.37625
themeDisney™:year2020  148.0980    67.3647   2.198  0.02920 *
themeMarvel:year2020    60.7238    59.9685   1.013  0.31262
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 141.2 on 179 degrees of freedom
Multiple R-squared:  0.7785, Adjusted R-squared:  0.7673
F-statistic: 69.89 on 9 and 179 DF,  p-value: < 2.2e-16
```

(a) (1 point) What value represents the (estimated) average number of pieces of a Disney set sold in 2018 with 10 pages in the manual?

Solution:

$$\begin{aligned} E[Y \mid \mathbf{X}] &= \hat{\beta}_0 + \hat{\beta}_1(10) + \hat{\beta}_2(1) + \hat{\beta}_3(0) + \hat{\beta}_4(0) + \hat{\beta}_5(0) + \hat{\beta}_6(0) + \hat{\beta}_7(0) + \hat{\beta}_8(0) + \hat{\beta}_9(0) + \hat{\beta}_{10}(0) \\ &= -13.7342 + 3.0304(10) + 32.2360 = 48.8058 \end{aligned}$$

Rubric: No partial marks. Must show all work to receive the point.

(b) (1 point) The 95% confidence interval for the coefficient of the interaction between the Marvel set and the year of sale at 2019 is $(-66.67, 175.56)$. What does this interval tell us about how these two factors/predictors affect the average number of pieces?

Solution: Because 0 is contained in the interval, it means that there is statistical evidence that there is no difference between the average number of pieces for a 2019 Marvel set compared to a 2018 City set with the same number of pages.

This would also be acceptable: Because 0 is contained in the interval, it means that there is statistical evidence (or it is plausible) that the coefficient for this interaction is 0 and so there is no difference between the intercept term for the average number of pieces for a 2019 Marvel set compared to a 2018 City set.

Rubric: Must mention this notion of no difference in the means of these two subgroups to receive the point.

(c) (6 points) On the next page are residual plots from the above model. Based on these plots, answer the following:

i. (3 points) What assumptions appear to be violated (if any) for this model? Be sure to reference specific plots in your answer.

Solution: From both the residuals vs fitted plot and the residuals vs pages plot, we see a possible curve in the residuals indicating an issue with linearity, as well as a possible fanning pattern indicating issues with constant variance. Additionally the Normal QQ plot shows deviations from the Normal quantiles, indicating an issue with Normality.

Rubric:

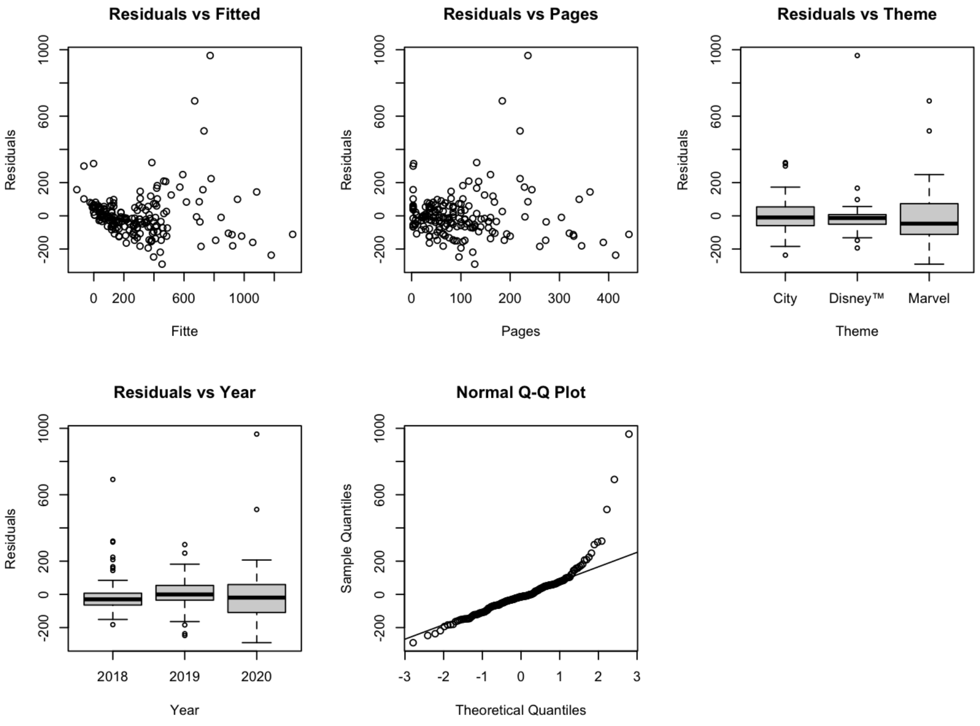
- 1 point for noting an issue with linearity with appropriate justification and reference to plot
- 1 point for noting an issue with constant variance with appropriate justification and reference to plot
- 1 point for noting an issue with Normality with appropriate justification and reference to plot

ii. (3 points) What techniques would you use to correct the violated assumptions you noted in (i)? Be specific about the terminology and what variables these methods are applied to.

Solution: For non-constant variance, we would attempt a variance stabilizing transformation or other transformation selected by trial and error on the response variable pieces. For normality, we would use Box-Cox to determine a reasonable transformation on the response variable. For linearity, we could use Box-Cox to try to find a reasonable transformation on the predictor pages or use a trial and error approach.

Rubric:

- 1 point for mentioning variance stabilizing transformation or trial and error for non-constant variance
- 1 point for mentioning box-cox for normality and for linearity (or can also say trial and error for linearity).
- 1 point for explicitly noting correct variance and Normality with transformations on Y but linearity could be Y or X.



(d) (3 points) Suppose a new model is fit to the same data but no interaction terms are used (see the R output below). What is the interpretation of the coefficient for the main effect of 2020 being the year of sale in this fitted regression?

```
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)         -33.0018    25.5395  -1.292  0.19792    
pages                 3.0341     0.1277  23.765  < 2e-16 ***
themeDisney™         84.7376    27.7645   3.052  0.00261 ** 
themeMarvel         163.2173    25.4800   6.406 1.23e-09 ***
year2019            -47.9803    25.3362  -1.894  0.05984 .  
year2020            -59.4474    25.9149  -2.294  0.02293 *  
---
```

Solution: -59.4474 is the change/difference in the estimated average pieces for a City set sold in 2020 compared to a City set sold in 2018 when the number of pages remains fixed.

Rubric: No partial marks beyond those listed below:

- 1 point for mention of change/difference between this year and the reference level
- 1 point for referencing average response in context
- 1 point for noting the reference level of theme while other predictor is fixed.

Problem 3

3. A large phone part manufacturing company has exceeded new orders and is interested in predicting their production time. The company wishes to know the time in hours it will take to produce 3 thousand new parts in an outgoing shipment. They randomly select 50 of their current orders recording both the number of parts produced (X, in thousands) and the time (Y, in hours) to produce them.

   A polynomial regression model
   $$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$
   for $i = 1, \ldots, 50$ with independent errors $\epsilon_i \sim N(0, 1)$ is used. The summary output from R with the **lm** function is provided below.

   Note that $I(parts^2)$ means we are fitting an x^2 term in the model and forcing R to perform the transformation in the model.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.552     0.1915   2.883  0.00593 **
parts         8.054     0.9323   8.640 2.88e-11 ***
I(parts^2)    0.879     0.9591   0.916  0.36417
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.435 on 47 degrees of freedom
Multiple R-squared:  0.9683,	Adjusted R-squared:  0.967
F-statistic: 718.4 on 2 and 47 DF,  p-value: < 2.2e-16
```

```
> qt(.95, 47)
[1] 1.677927

> qt(.95, 48)
[1] 1.677224

> qt(.975, 47)
[1] 2.011741

> qt(.975, 48)
[1] 2.010635
```

```
> predict(fit, data.frame(parts = 3), interval = "prediction")
     fit    lwr    upr
1 32.625 20.321 44.928
```

(a) (5 points) Compute a 95% confidence interval for the average time with 3 thousand phone parts. (Round your answer to 3 decimal places)

Solution: From the summary output $\hat{\sigma}^2 = 0.435^2$. First

$$\hat{Y} = 32.625$$

Then we compute

$$\hat{sd}(Y - \hat{Y}) = \sqrt{\hat{\sigma}^2 + \begin{pmatrix} 1 \\ 3 \\ 3^2 \end{pmatrix}^T \hat{\sigma}^2 X^T X^{-1} \begin{pmatrix} 1 \\ 3 \\ 3^2 \end{pmatrix}} = (44.928 - 32.625)/2.011741 = 6.115598.$$

So then

$$\hat{sd}(\hat{Y}) = \sqrt{6.115598^2 - 0.435^2} = 6.100108.$$

The confidence interval at $x = 3$ is The prediction interval at $x = 3$ is

$$[32.625 - 2.011741 * 6.100108, 32.625 + 2.011741 * 6.100108] = [20.35316, 44.89684]$$

Rubric: If a calculation mistake occurs early in the question, all subsequent steps should be marked as correct, as long as no additional calculation or conceptual error has been made (i.e. do not double penalize for a mistake that has been carried over). Steps without work shown should not receive the associated mark.

- 1 point for finding the correct fitted value
- 1 point for correct setup and value of $\hat{Var}(Y - \hat{Y})$
- 1 point for finding correct value of s^2
- 1 point for correct computation of either standard error or variance of $\hat{y}$
- 1 point for correct final interval calculation

Students may have not understood the units of the variable parts (i.e. in thousands) and may have used $x = 3000$ in their calculation. In this situation, **deduct 1 mark for "finding the correct fitted value"** and as long as remaining work is correct using those same values, they can receive the remaining 4 marks in full. See below for the solution under this scenario:

$$\hat{Y} = 0.552 + 3000 * 8.054 + 3000^2 * 0.879 = 7935163.$$

From the summary output $s^2 = \hat{\sigma}^2 = 0.435^2$. Then we compute

$$\hat{sd}(Y - \hat{Y}) = \sqrt{\hat{\sigma}^2 + \begin{pmatrix} 1 \\ 3000 \\ 3000^2 \end{pmatrix}^T \hat{\sigma}^2 X^T X^{-1} \begin{pmatrix} 1 \\ 3000 \\ 3000^2 \end{pmatrix}} = (44.928 - 32.625)/2.011741 = 6.115598.$$

So then the estimated standard error for the prediction interval is

$$\hat{se}(\hat{Y}) = \sqrt{6.115598^2 - 0.435^2} = 6.100108.$$

The confidence interval at $x = 3000$ is

$$[7935163 - 2.011741 * 6.100108, 7935163 + 2.011741 * 6.100108] = [7935151, 7935175]$$

(b) (2 points) Suppose you were to check for a linear relationship using a scatterplot between x and x^2. What would you expect to see in this plot and why? Why would this model still be considered linear?

Solution: We would expect to see a prominent curved or quadratic relationship between x and x^2 because the two predictors in question are functions of one another by design.

We could still consider this model linear because "linear" in regression models refers to the coefficients not the form of the predictors.

Rubric:

- 1 point for noting a likely curved or quadratic pattern due to forced relationship between predictors (if no justification given, no point awarded).
- 1 point for mentioning that "linear" refers to the coefficients not the functional form of the predictors.