University of Toronto

Faculty of Arts and Science

2024 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 errors. Consider for $i = 1, \ldots, n$,

$$Y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e_i,$$
$$E(e_i) = 0, \qquad \text{Var}(e_i) = 4\sigma^2 \qquad e_i \text{ are independent but not normal.}$$

However, despite misspecification 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 $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$ 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:

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

Rubric: No credit for no work shown. Avoid penalizing more than once for an error carried through. If they make a mistake in an earlier step and carry this into the subsequent steps, award them points based on whether the remaining work is correct after original mistake occurs.

- 1 point for applying expectation to LS estimator
- 1 point for removing constants from expectation
- 1 point for using linearity of expectation (-0.5 if they did not note this use as instructed in question)
- 1 point for simplifying to final answer.
- Note: Due to a typo regarding x_1, x_2 not being $x_{i,1}, x_{i,2}$, a student could also answer mentioning that $(X^TX)^{-1}$ not being invertible will be an issue. However, likely this won't be an issue.

(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:

$$\begin{aligned} Cov(\hat{\boldsymbol{\beta}}|\boldsymbol{X}) &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{Y}|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} && \text{(properties of covariance)} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T Cov(\boldsymbol{e})\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} && \text{(population/regression model)} \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T 4\sigma^2\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}. && \text{(population covariance assumption)} \\ &= 4\sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned}$$

Rubric: No credit for no work shown. Avoid penalizing more than once for an error carried through. If they make a mistake in an earlier step and carry this into the subsequent steps, award them points based on whether the remaining work is correct after original mistake occurs.

- 1 point for applying covariance to estimator
- 1 point for correctly removing constants from covariance (this includes the constant from $Cov(e)$
- 1 point for correctly replacing Y with population relationship and simplifying
- 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 the same mean but 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 have the same mean

- 1 point for noting the misspecified model will have different variance (-0.5 if only notes they are different).
- 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 257 sets comprising three themes: NINJAGO sets, Star Wars sets and Friends sets. The number of pieces and the number of minifigures in each set were also collected. A model was fit to these data, with the original purchase price of a set as the response, and with predictors pieces, minifigures and theme, and an interaction term between pieces and theme. Below is the model output:

```
Call:
lm(formula = price ~ pieces + minifigures + theme + pieces:theme,
    data = d1)

Residuals:
     Min       1Q   Median       3Q      Max
-120.861   -3.418   -0.586    3.969   94.420

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)             2.681256   1.982549   1.352  0.17746
pieces                  0.094997   0.005483  17.325  < 2e-16 ***
minifigures             0.652170   0.528529   1.234  0.21839
themeNINJAGO®           4.137466   2.682981   1.542  0.12431
themeStar Wars™        -8.231855   2.731335  -3.014  0.00284 **
pieces:themeNINJAGO®   -0.027146   0.005844  -4.645  5.5e-06 ***
pieces:themeStar Wars™  0.032473   0.005655   5.742  2.7e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.43 on 250 degrees of freedom
Multiple R-squared:  0.9574,        Adjusted R-squared:  0.9564
F-statistic: 935.9 on 6 and 250 DF,  p-value: < 2.2e-16
```

(a) (3 points) What role does the interaction term between pieces and Star Wars theme play in this model? Write your answer using the context of the problem.

Solution: 0.03274 is the additional increase to the average change in price for a Star Wars set compared to a Friends set for an additional piece when the number of minifigures remains fixed.

OR

For Star Wars, the slope term for pieces is

$$(0.094997 + 0.032473) = 0.12747$$

so the interaction term increases the slope term of pieces (i.e. the change in mean Price for a unit

increase in pieces) from the Friends Lego to the Star Wars Lego by 0.032473.

Rubric: No partial marks beyond those listed below:

- 1 point for mention of change/difference between this theme and the reference level
- 1 point for referencing average response in context
- 1 point for noting the 1-unit increase in pieces while other predictor is fixed. Referencing the change in the slope term for pieces is acceptable.

(b) (1 point) What value represents the (estimated) average price for a NINJAGO set with no pieces or minifigures?

Solution:

$$\begin{aligned}\hat{E}[Y \mid \mathbf{X}] &= \hat{\beta}_0 + \hat{\beta}_1(0) + \hat{\beta}_2(0) + \hat{\beta}_3(1) + \hat{\beta}_4(0) + \hat{\beta}_5(0) + \hat{\beta}_6(0) + \hat{\beta}_7(0) \\ &= 2.681256 + 4.137466 = 6.818722\end{aligned}$$

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

(c) (1 point) The 95% confidence interval for the coefficient of the NINJAGO theme indicator variable is $(-1.147, 9.422)$ which includes 0. How would the interpretation of the intercept change if this indicator variable were not included when fitting this model?

Solution: The intercept would now represent the average price for a non-Star Wars (NINJAGO or Friends) set when there are no minifigures and no pieces.

Rubric: No partial marks. Answer must include the new reference level and that it still represents the average response with other predictors fixed.

(d) (6 points) On page 8, you will find the residual plots from two models: the original model whose output is presented earlier, and a transformed model with the same predictors as the original but price and pieces have been transformed with natural logarithms. Based on these plots, answer the following:

ii. (3 points) Why were these transformations applied to our model? Be sure to reference specific plots and assumptions in your answer.

Solution: The natural logarithm applied to the response price could be an attempt to correct the slight deviations present in the Normal plot and fix the residuals. It could alternatively be an attempt to correct possible non-constant variance issue as seen by different spread in the residuals vs fitted and residuals vs theme plot. The natural logarithm applied to pieces could be an attempt to correct for a decreasing pattern in the residuals vs fitted plot, which possibly indicates an issue of non-linearity.
The response versus fitted could also be used to check for issues with constant variance and non-linearity. Some students may answer this and it would be accepted.

Rubric: points can only be awarded if answer references a specific plot and the pattern visible there.

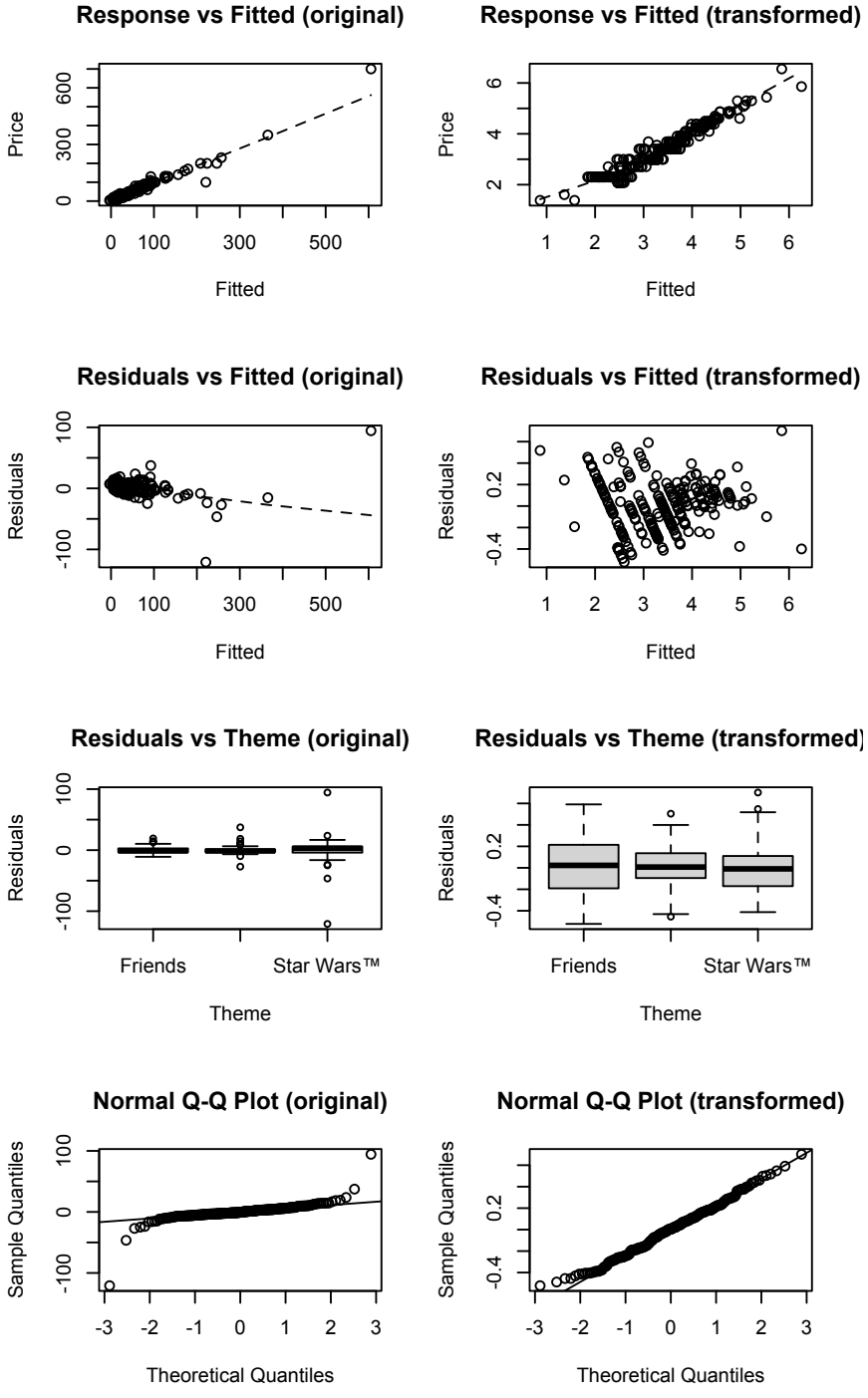
- 1 point for mentioning normality and how it connects to the transformation on the response
- 1 point for mentioning constant variance and how it connects to the transformation on the response
- 1 point for mentioning linearity and how that connects to transformation on the predictor.

iii. (3 points) Which model appears to have satisfied the assumptions better? Justify your answer, making sure to reference specific plots and assumptions in your answer.

Solution: It appears that the transformed model has better model assumptions. The residuals vs fitted values plot demonstrates a more uniform distribution, while the residuals vs theme plot shows relatively constant spread throughout and a narrower spread of residuals overall. The Normal QQ plot shows a stronger one-to-one relationship with the normal quantiles.

Rubric:

- 1 point for describing that Normality improved in transformed model with mention of plot
- 1 point for describing that constant variance is improved in transformed model with mention of plot
- 1 point for describing that linearity is improved in transformed model with mention of plot


Response vs Fitted (original)
Price
Fitted
Response vs Fitted (transformed)
Price
Fitted
Residuals vs Fitted (original)
Residuals
Fitted
Residuals vs Fitted (transformed)
Residuals
Fitted
Residuals vs Theme (original)
Residuals
Friends
Star Wars™
Theme
Residuals vs Theme (transformed)
Residuals
Friends
Star Wars™
Theme
Normal Q-Q Plot (original)
Sample Quantiles
Theoretical Quantiles
Normal Q-Q Plot (transformed)
Sample Quantiles
Theoretical Quantiles

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 (in thousands) and the time (in hours) to produce them.

   A polynomial regression model

   $$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + e_i$$

   for $i = 1, \ldots, 50$ with independent errors $e_i \sim N(0, 1)$ are used. The summary output from R with the **lm** function is provided below.
   Note that **vcov(fit)** is the estimated covariance matrix for the estimated coefficients in this model and $I(parts^2)$ 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
```

```
> vcov(fit)
            (Intercept)  parts I(parts^2)
(Intercept)       0.037 -0.157      0.140
parts            -0.157  0.869     -0.865
I(parts^2)        0.140 -0.865      0.920
```

(a) (2 points) What are the challenges with interpreting the estimated coefficients in this model?

Solution: Since x features in both "predictors" in the model, it is not possible to "hold the other predictor fixed" in order to interpret a single coefficient.

Rubric:

- 1 point for noting x in both variables
- 1 point for noting inability to hold other fixed.

(b) (5 points) Compute a 95% prediction interval with 3 thousand phone parts. (Round your answer to 3 decimal places)

Solution: We first compute

$$\hat{Y} = 0.552 + 3 * 8.054 + 3^2 * 0.879 = 32.625.$$

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

$$\hat{Var}(\hat{Y}) = \begin{pmatrix} 1 \\ 3 \\ 3^2 \end{pmatrix}^T \hat{\sigma}^2 X^T X^{-1} \begin{pmatrix} 1 \\ 3 \\ 3^2 \end{pmatrix} = 37.246.$$

So then the estimated standard error for the prediction interval is

$$\hat{se}(Y - \hat{Y}) = \sqrt{0.435^2 + 37.246} = 6.118433.$$

The prediction interval at $x = 3$ is

$$[32.625 - 2.011741 * 6.118433, 32.625 + 2.011741 * 6.118433] = [20.3163, 44.9337]$$

Rubric: An **incorrect** confidence interval would be $[20.34745, 44.90255]$. 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 $\hat{Var}(\hat{y})$
- 1 point for finding correct value of $s^2 = \hat{\sigma}^2$

- 1 point for correct computation of either standard error or estimated variance of prediction error $Y - \hat{Y}$
- 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

$$\hat{Var}(\hat{Y}) = \begin{pmatrix} 1 \\ 3000 \\ 3000^2 \end{pmatrix}^T \hat{\sigma}^2 X^T X^{-1} \begin{pmatrix} 1 \\ 3000 \\ 3000^2 \end{pmatrix} = 7.44733 \times 10^{13}.$$

So then the estimated standard error for the prediction interval is

$$\hat{se}(Y - \hat{Y}) = \sqrt{0.435^2 + 7.44733 \times 10^{13}} = 8629791.$$

The prediction interval at $x = 3$ is

$$[7935163 - 2.011741 * 8629791, 7935163 + 2.011741 * 8629791] = [-9425741, 25296067]$$