STA 248 - Final Problem Set

Due: April 13, 2020 @ 12 PM EDT - NO EXCEPTIONS

Submit through Crowdmark (emailed link)

The goal of this evaluation is to assess your proficiency and ability to apply **concepts taught in the course**. While there may be different paths to the solution, you must demonstrate in your solutions that you have, without a doubt, learned the course material. This is an **individual** assignment. **Read the academic integrity statement below and fill in your student information. This is mandatory as part of your submission.**

Problems labeled with **[R script]** indicate problems that require the use of R and your (well organized/labeled) R-script will need to be submitted separately. If you are using R-Studio, you can easily export your script as a pdf file. Your written responses should **not** be included in the script.

Academic Integrity Statement

Academic integrity is a fundamental value of learning and scholarship at the UofT. Participating honestly, respectfully, responsibly, and fairly in this academic community ensures that your UofT degree is valued and respected as a true signifier of your individual academic achievement.

Prior to beginning this problem set, you must attest that you will follow the Code of Behaviour on Academic Matters and will not commit academic misconduct in the completion of this assessment. Affirm your agreement to this by completing the following Statement:

*By signing this Statement, I, _____ **(name)**, _____ **(student number)**, agree to fully abide to the Code of Behaviour on Academic Matters. I will not commit academic misconduct, and am aware of the penalties that may be imposed if I commit an academic offence.*

The University of Toronto's Code of Behaviour on Academic MattersLinks to an external site. outlines the behaviours that constitute academic misconduct, the processes for addressing academic offences, and the penalties that may be imposed. You are expected to be familiar with the contents of this document.

Potential offences include, but are not limited to:

- *Using someone else's ideas or work as part of your solutions.*
- *Obtaining or providing unauthorized assistance on any assignment (this includes working in groups, or using solutions from external sources).*
- *Looking at someone else's answers, letting someone look at (and potentially use) your answers, or working together to answer questions.*

Problem 1 [25 points, R - 9]. In this problem you will be examining the COVID-19 cases in Canada. Our data set comes from the Johns Hopkins University Center for Systems Science and Engineering. See https://github.com/CSSEGISandData/COVID-19 for details. Our data set consists of two columns of interest: number of confirmed cases (called 'cases') and the number of days passed since the 10th case (called 'days_since_10th_case'). You will be exploring the relationship between the explanatory variable 'days_since_10th_case' and the response variable 'cases'.

To import the data, you may use the following command (ensure that the 'covid19_canada.csv' file is in your working directory): `df = read.csv('covid19_canada.csv')` or the usual `df=read.csv(file.choose(), header=T)`.

Your submission for this problem should consist of one (typed) document with all your short responses with their supporting plots, labeled by question, and one pdf document of your R script. Your R-script submission should only contain all the (ordered and labeled) scripts that you used to create plots, transform data, fit models, etc. Be sure that you include proper axis labels for all of your graphs, rather than the default labeling.

a) (2 points, R - 1 point) **[R script]** Create a scatterplot of the data with the response variable on the y-axis. Based on your scatterplot, decide whether to use SLR to explore the relationship between the number of cases and days since 10th case data. Justify your response.

b) (1 point, R - 2 points)**[R script]** Fit an SLR model to the dataset, and call it 'model1'. Plot the fitted line on top of the scatterplot from (b) using the command `abline(model1)`. Comment on the fit of your model to the data.

c) (1 point, R - 1 point) **[R script]** Make a QQ-plot of the residuals to check the normality assumption. Comment on your findings.

d) (3 points, R - 1 point) **[R script]** While not ideal, there are ways to work with data sets such as these. Note the shape of your data. Apply a log-transformation (base 'e' is fine, but if you prefer to use base '10' (or any other base), please indicate it in your script) on the *response* variable and call this new variable 'log.y'. Make a new scatterplot of your transformed data.

e) (1 point, R - 1 point) **[R script]** Fit an SLR model to your transformed dataset and call it 'model2'. Plot the fitted line on top of your scatterplot in (e) and comment on the fit.

f) (6 points, R - 3 points)**[R script]** Perform a complete residual analysis on model 2. Include your residual plots in your analysis. Is SLR appropriate for this transformed dataset?

g) (5 points) State the estimated model (be careful of notation and pay attention to the variables in your model!). What does the estimated model say about the potential association between the **number of cases** and the **number of days since the 10th case**? i.e. Interpret the coefficients of model 2.

h) (6 points) As you may have noticed, the data only goes as far as March 21, 2020 which is roughly one week since the beginning of social distancing. For all intents and purposes, this data shows the progression of the number of cases without any intervention. Use your model to predict the expected number of cases, nationally, for Sunday April 12, 2020 and compare it with the most recently available numbers (either Saturday April 11 or Sunday April 12). Based on your model, does it appear that social distancing is having an effect in curbing the number of cases in Canada? Comment on the validity of your prediction.

BONUS 4 pts i) Using what you have explored in this dataset, describe how you would be able to recognize whether a dataset should be log-transformed before seeing if a SLR model would be appropriate. Describe how you would decide whether to transform the response variable or the explanatory variable. *(This is a thinking problem. You are expected to carefully consider why a log-transformation was useful in this problem so you can apply this to future scenarios.)*

Problem 2 [16 points]. Twenty grade 10 students were randomly separated into four equal groups, and each group was taught a mathematical concept using a different teaching method. At the end of the teaching period, progress was measured by a unit test. The scores are shown below. One student in group 3 was absent on the day that the test was administered:

Group			
1	2	3	4
112	111	140	101
92	129	121	116
124	202	130	105
89	136	106	126
97	99		119

a) (1 point) What type of design has been used in this experiment?

b) (2 points) If ANOVA were applied to this data set, you would be testing for association between which two variables?

c) (3 points) With such a small data set, what assumptions must be made if you were to apply ANOVA?

d) (10 points) Conduct **by hand** a hypothesis test for difference in mean scores using ANOVA. Be sure to state your hypotheses and use R to compute the exact p-value of your hypothesis test.

Problem 3 [14 points]. The weights in grams of 15 male and 19 female juvenile ring-necked pheasants are given below:

Males		Females	
1384	1672	1073	1058
1286	1370	1053	1123
1503	1659	1038	1089
1627	1725	1018	1034
1450	1394	1146	1281
1642	1751	1168	1242
1527	1250	1067	1171
1466		1085	1030
		1218	1184
		1070	

Analyze **by hand (you can use R for longer computations, such as s^2)** the data to determine whether male and female rink-necked pheasants differ significantly in weight, on average. Use R to find exact p-values. Be sure to interpret your results and p-values in the context of this problem.

Problem 4 [22 points, R - 4 points]. You are trying to determine if the median time until failure (in hours) of some component after it's been exposed to moisture is shorter than advertised. The time until failure is believed to be Weibull distributed with shape parameter $k = 0.25$.
Recall from assignment 2 that a Weibull distribution with shape parameter k and scale parameter λ has the following probability density function:

$$f(x; k, \lambda) = \begin{cases} \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

Initially, the company producing this component believes that the scale parameter λ is 10. Due to difficulties in observing and testing this component, you are only able to collect a sample of 10 times until failure of this component:

111.7217	0.2531	0.3014	0.2494	0.0025
0.1975	2351.5894	22.6345	0.1363	0.9781

a) (1 point) Briefly describe why a χ^2 Goodness of Fit test would not be applicable here.

b) (2 points) Briefly describe why we cannot conduct our usual hypothesis tests (using either Z or T distributions) for this data set.

c) (4 points) The cumulative distribution function of a $Weibull(k, \lambda)$ distribution is

$$F(x) = \begin{cases} 1 - e^{-(x/\lambda)^k}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

If the company's belief is correct, what should the median time until failure be, in theory? State your null and alternative hypotheses.

d) (2 points) What is the sample statistic for the hypothesis test? What would the p-value of your hypothesis test represent?

e) (4 points) Describe the steps you would take to use R to simulate the p-value for the hypothesis test that you would use to test the company's claim.

f) (R - 4 points) **[R script]** In R, using `set.seed(2020248)`, conduct a hypothesis test using 100,000 simulations to estimate the p-value.

g) (3 points) Report your p-value. Are your results surprising, compared with your response to (c) and (d)? Why or why not?

h) (6 points) Write the likelihood function for 10 random $Weibull(0.25, \lambda)$ observations and derive the maximum likelihood estimator for λ.

Problem 5 [9 points]. Is the price range of IKEA furniture dependent on furniture type? State your hypothesis using appropriate notation, and use the following subset of IKEA pricing data to answer this question. Provide a complete solution and show all steps. Use R to compute your exact p-value.

	Price Range ($)				
Furniture Type	[1, 100)	[100, 200)	[200, 300)	[300, 400)	[400, ∞)
Beds	36	66	68	107	267
Bookcases & Shelving	473	229	183	107	267
Display Cabinets	248	109	170	98	211
Sofas & Armchairs	251	183	106	63	508

Problem 6 [14 points]. Data on blood pressure (systolic blood pressure) and body mass index (BMI) of 40 randomly selected persons was collected. After performing residual analysis, it was determined that a simple linear regression model can be fitted to the data. The R output of the linear model is provided below, with some missing values.

```
Call:
lm(formula = systolic$systolic ~ systolic$bmi)

Residuals:
      Min         1Q     Median         3Q        Max
-52.329471 -15.493438  -5.661066  16.654884  58.758381

Coefficients:
               Estimate  Std. Error t value  Pr(>|t|)
(Intercept) [redacted]   13.8909973 [redacted] 1.911e-09 ***
systolic$bmi   1.5445413 [redacted] 0.0045919 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.87061 on [redacted] degrees of freedom
Multiple R-squared:  0.1927789, Adjusted R-squared:  0.1715362
F-statistic: 9.075082 on [redacted] and [redacted] DF,  p-value: 0.004591914
```

$\sum x_i = 1031.79145 \quad \sum x_i^2 = 29361.51006 \quad \sum y_i = 5944 \quad \sum y_i^2 = 917268 \quad \sum x_i \cdot y_i = 157566.5548$

a) (5 points) Find all missing values from the R output. Show your work.

b) (3 points) Does there appear to be compelling evidence that systolic blood pressure and body mass index are associated? State the null and alternative hypotheses, report and interpret the p-value.

c) (3 points) The dataset ranged from 12 to 49 for BMI. Knowing this, estimate the average systolic blood pressure of people with BMI of 25, and provide a 95% confidence interval for your estimate.

d) (3 points) The next person that is surveyed at random has a BMI of 32. Estimate their systolic blood pressure to 95% confidence.

e) (3 points) How would you use your model to predict an individuals BMI based on their systolic blood pressure? Explain.

Academic Integrity Statement

All suspected cases of academic dishonesty will be investigated following the procedures outlined in the Code of Behaviour on Academic Matters. Please sign the Statement below to complete your assessment.

By signing this Statement, I am attesting to the fact that I,
_____ ***(name)***, _____ ***(student number)***, *have abided fully to the Code of Behaviour on Academic Matters. I have not committed academic misconduct, and am aware of the penalties that may be imposed if I have committed an academic offence.*