NAME (PRINT): _____
Last/Surname First /Given Name

STUDENT #: _____ **SIGNATURE:** _____

UNIVERSITY OF TORONTO MISSISSAUGA
APRIL 2020 FINAL EXAMINATION
STA441H5S
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours
Aids: This exam is open book and open notes. All materials on the course website are allowed. Use of SAS is mandatory. Statistical Tables and Formula sheet are available on the course website:

http://www.utstat.toronto.edu/~brunner/441s20

The University of Toronto Mississauga and you, as a student, share a commitment to academic integrity. During the exam, you are requested not to consult with anyone, and not to access any website other than the course website and the SAS OnDemand website.

*Please note, once this exam has begun, you **CANNOT** re-write it.*

Qn. #	Value	Score		Qn. #	Value	Score
1	10			6	18	
2	10			7	11	
3	5			8	15	
4	14			9	6	
5	11					
Total = 100 Points						

10 points

1. Please give the required sample size for each problem below. Do it in **one SAS program,** and append your **log file only** to the end of your exam.

   (a) In a 3×4 analysis of covariance, factor A will have three levels, factor B will have four levels, and there will be one covariate. For the test of the interaction between A and B, we want to have an 80% chance of significance provided that the interaction explains 5% of the remaining population variation after allowing for the covariate and main effects. What sample size is required? Write the number in the space below. There is no requirement of equal sample sizes.

   (b) For a regression model with nine explanatory variables, you want the test of $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ to be statistically significant at the $\alpha = 0.05$ level provided that the variables x_1 through x_4 explain at least 3% of the remaining sample variation. What sample size is required?

continued on page 3

10 points

2. In a study of the links between TV violence and aggression, parents of children in a daycare filled out a questionnaire about the TV programs their children watched. Daycare workers recorded the number of attacks and other violent incidents (taking a toy from another child, etc.) for each child. Can this study provide good evidence that violent TV can contribute to violent behaviour? Clearly answer Yes or No, and briefly discuss in terms of concepts from this course.

continued on page 4

5 points

3. Make up an original study for which a two-factor multivariate analysis of covariance would be the appropriate tool. Both factors should be between-cases. After briefly describing the study,

   - List the explanatory variable or variables. For each one, say whether it is quantitative or categorical.
   - List the response variable or variables. For each one, say whether it is quantitative or categorical.

continued on page 5

14 points

4. Steel is made by heating iron and adding some carbon. A steel company conducted an experiment in which knife blades were manufactured using two different amounts of carbon (Low and High), and three different temperatures (Low, Medium and High). Of course even the Low temperature was very hot. A sample of knife blades was manufactured at each combination of carbon and temperature levels, and then the breaking strength of each blade was measured by a specially designed machine. The response variable is breaking strength. The table below shows population treatment means.

	Temperature		
	Low	Medium	High
Low Carbon	μ_{11}	μ_{12}	μ_{13}
High Carbon	μ_{21}	μ_{22}	μ_{23}

In terms of the μ_{ij} values in the table, give the null hypothesis you would test to answer each question.

(a) Is there a main effect of Temperature?

(b) Averaging across Temperature level, is there difference between High Carbon and Low carbon in the mean breaking strength of the knives?

(c) Is there a Temperature by Carbon level interaction?

(d) Averaging across Carbon level, does Temperature affect the mean breaking strength of the knives?

(e) Does the effect of Temperature depend on Carbon level?

(f) Is there an effect of Temperature when the Carbon level is low?

(g) Is there a difference beween the marginal means for low and high Carbon level?

continued on page 6

11 points

5. This question uses the knife manufacturing example of Question 4.

(a) In the table below, make columns showing how you would set up dummy variables for both explanatory variables, using *effect coding* (that's the scheme with 0, 1 and -1). *Write the name of each dummy variable at the top of its column.*

Low Carbon, Low Temperature	
Low Carbon, Medium Temperature	
Low Carbon, High Temperature	
High Carbon, Low Temperature	
High Carbon, Medium Temperature	
High Carbon, High Temperature	

(b) Write $E(Y|\mathbf{X} = \mathbf{x})$ for the regression model, using the variable names from your table above. Include the interactions!

(c) Using the β values from your answer to the preceding question, state the null hypothesis you'd test to answer the following questions.

i. Is there a main effect of Temperature?

ii. Averaging across Temperature level, is there difference between High Carbon and Low carbon in the mean breaking strength of the knives?

iii. Is there a Temperature by Carbon level interaction?

iv. Averaging across Carbon level, does Temperature affect the mean breaking strength of the knives?

v. Does the effect of Temperature depend on Carbon level?

vi. Is there a difference beween the marginal means for low and high Carbon level?

continued on page 7

18 points

6. In lecture, the math data were used to illustrate the multinomial logit model. The response variable `outcome` had three categories: Pass, Disappear and Fail. For this question,

$$\begin{aligned} \pi_1 &= P(\text{Pass}|\mathbf{x}) \\ \pi_2 &= P(\text{Disappear}|\mathbf{x}) \\ \pi_3 &= P(\text{Fail}|\mathbf{x}), \end{aligned}$$

and $P(\text{Fail}|\mathbf{x})$ is the reference category, so that π_3 is in the denominator of all the generalized logits. This corresponds directly to the formula sheet.

Suppose the only explanatory variable in the model is a dummy variable gender, with $x = 1$ for females and $x = 0$ for males.

(a) Write the model equations, starting with

$$\ln\left(\frac{\pi_1}{\pi_3}\right) =$$

(b) In terms of the symbols from your model equations, what is the probability that a male student will fail the course? You don't need to show any work; just write down the answer.

(c) Give the null hypothesis that you would test to answer each question below. Use symbols from your model equations.

i. Is outcome related to gender?

ii. For males, is the probability of passing the same as the probability of disappearing?

iii. For females, is the probability of failing the same as the probability of disappearing?

(d) If you rejected $H_0 : \beta_{0,1} = \beta_{0,2} = 0$, what would you conclude? Use plain, non-statistical language (but you are allowed to use the word "probability.") A bit of High School algebra is optional.

continued on page 8

11 points

7. Please refer to the printout for the **Air Quality Study**.

(a) It makes sense that there should be sequential dependence in air pollution data. The worse it is today, the worse it tends to be tomorrow. To assess this directly, I fit a model with just the intercept – no explanatory variables.

i. What do you conclude? There is no need for plain language.

ii. Write the p-value that supports your conclusion.

(b) Then I fit another model with some explanatory variables this time, using ordinary least squares.

i. Does it appear that we need a time series model? Answer Yes or No.

ii. Write the p-value that supports your conclusion.

iii. Controlling for solar radiation level, wind speed and a trend for time, is ozone level related to air temperature? Fill out the table below.

Test Statistic (a number)	p-value (a number)	Reject Null Hypothesis? (Yes or No)	Statistically Significant? (Yes or No)

iv. In plain, non-statistical language, what do you conclude from the table above?

continued on page 9

15 points

8. Please refer to the printout for the **Self-Esteem Study**.

(a) In the table below, write the p-values for the tests of these standard hypotheses.

Effect	p-value
Main Effect of Diet	
Main Effect of Time	
Diet by TimeInteraction	

(b) In your opinion, should we be interpreting the main effects here? Clearly answer Yes or No, and briefly explain.

(c) The printout has a test of whether Diet affects self esteem at any time period. That is, the null hypothesis is that Diet has no effect at any time period. Give the p-value.

(d) The follow-up tests are also given. Based on these, what do you conclude? Use plain, non-statistical language. Don't bother with a Bonferroni correction.

continued on page 10

6 points

9. Please refer to the printout for the **UCLA Graduate Admissions Data**.

   (a) Controlling for grade point average and score on the Graduate Record Examination, the estimated odds of acceptance for a student from a second-ranked university are _____ times as great as the odds of acceptance for a student from a top-ranked university. Write the number in the space below.

   (b) What is the estimated probability of acceptance for a student from a top-ranked university, who has a GPA of 3.15 and a 600 on the Graduate Record Examination? The answer is a number. Show a little work. **Circle the number**.

Don't forget to append your log file from Question 1.

Total Marks = 100 points