NAME (PRINT): ____

Last/Surname First /Given Name

STUDENT #: ______________________ **SIGNATURE:** ________________

UNIVERSITY OF TORONTO MISSISSAUGA
APRIL 2018 FINAL EXAMINATION
STA441H5S
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours
Aids: Statistical Calculators; Formula sheet and printouts supplied

The University of Toronto Mississauga and you, as a student, share a commitment to academic integrity. You are reminded that you may be charged with an academic offence for possessing any unauthorized aids during the writing of an exam. Clear, sealable, plastic bags have been provided for all electronic devices with storage, including but not limited to: cell phones, SMART devices, tablets, laptops, calculators, and MP3 players. Please turn off all devices, seal them in the bag provided, and place the bag under your desk for the duration of the examination. You will not be able to touch the bag or its contents until the exam is over.

If, during an exam, any of these items are found on your person or in the area of your desk other than in the clear, sealable, plastic bag, you may be charged with an academic offence. A typical penalty for an academic offence may cause you to fail the course.

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

Qn. #	Value	Score
1	5	
2	5	
3	8	
4	14	
5	13	
6	14	
7	14	
8	15	
9	12	
Total = 100 Points		

Seat Position

5 points 1. Give an original example of an *experimental* study with two explanatory variables, in which one of the factors is between cases and the other factor is withn-cases.

5 points 2. A psychologist interviewed by the CBC said research shows that more than 5 hours a day of screen time results in anxiety and depression for teenagers. There is a problem with use of the word "results." Why?

Continued on page 3

8 points

3. The Titanic was a passenger ship that hit an iceberg and sank on its very first voyage in 1912. It was the largest passenger ship in the world at the time, and supposedly unsinkable. More than 1,500 of the roughly 2,200 passengers and crew died.

   Passengers were either in 1st class (where there were some lifeboats), 2nd class or 3d class, and they either lived or died. Let c_2 be an indicator dummy variable for 2nd class and c_3 be an indicator dummy variable for 3d class; $y = 1$ means the passenger survived.

   (a) Write a regression equation for the log odds of survival. There is an intercept.

   (b) Compared to the odds of survival for a passenger in 1st class, the odds of survival for a passenger in 3d class are _____ times as great. Write your answer in the space below. Give the answer in terms of the β values from your regression model. You don't have to prove it or show any work.

   (c) In terms of the symbols from your model, what is the probability of survival for a passenger in second class? Again, you need not show any work.

   (d) What null hypothesis would you test to determine whether Class (1st versus 2nd versus 3d) was related to survival? Give the answer in terms of the β values from your regression model. Write your answer in the space below.

   (e) What null hypothesis would you test to determine whether passengers in 2nd class had a better chance of survival than passengers in 3d class? Give the answer in terms of the β values from your regression model. Write your answer in the space below.

Continued on page 4

14 points

4. A restaurant owner decides to experiment with changing the appearance of the menu. Copies of five different menus are prepared, and tables are randomly assigned to a type of menu as the customers enter, using a list of random numbers. Naturally, everyone at a given table receives the same menu. The menu types are

   - Existing menu, which has no pictures, and the food categories (meat, seafood etc.) are in no particular order. Within a food category, the dishes (meaning the food on the plate, not the plate itself) are also in no particular order.
   - With small pictures of the dishes, same order as the existing menu.
   - With large pictures of the dishes, same order as the existing menu.
   - With small pictures of the dishes, most expensive dishes first.
   - With large pictures of the dishes, most expensive dishes first.

   The response variable y is the total amount of the bill for the table before the tip. Number of people in the party is a covariate, denoted by x.

   (a) You will use a regression model with an intercept, indicator dummy variables for menu type, and no interaction terms. Write the regression equation below, including ϵ. You do not have to say how the dummy variables are defined yet. You will do that in the next part. I will start you out.

   $y =$

   (b) Make a table showing how you would set up the dummy variables. Make existing menu the reference category. Add a column to the end of your table, showing the expected amount of the bill in terms of your β parameters and number of people in the party.

Continued on page 5

(c) Controlling for number of people in the party, does menu type affect how much people spend? State the null hypothesis in terms of β parameters.

(d) Controlling for number of people in the party, is there a difference in expected amount of the bill between the existing menu and the menu with large picture that shows the most expensive dishes first? State the null hypothesis in terms of β parameters.

(e) When you average across menus with the two different orders of dishes, is there a difference in expected amount of bill, comparing menus with big pictures to menus with small pictures? This is one test controlling for number of people in the party, and you are excluding the existing menu type. State the null hypothesis in terms of β parameters.

(f) Controlling for number of people in the party, is the average expected amount of bill for the four new menu types different from the expected amount of bill for the existing menu? This is one test. State the null hypothesis in terms of β parameters.

(g) For menus with large pictures, is expected amount of bill affected by the order of the menu items, controlling for number of people in the party? This is one test. State the null hypothesis in terms of β parameters.

Continued on page 6

13 points 5. Graduating High School students indicate their plans by choosing one of these alternatives: (1) University (2) Seek employment, or (3) Other.

Plans after graduation is the response variable, with seeking employment the reference category that goes in the denominator of each generalized logit. The explanatory variables are grade point average (denoted by x_1) and an indicator for being in an academic as opposed to an "applied" stream in High School: $x_2 = 1$ if academic and $x_2 = 0$ if applied.

(a) Write the equations of a generalized logit model for these data. There should be an intercept in each equation, and no interactions.

(b) The reference category for the response variable (corresponding to the denominator of the generalized logits) will be seeking employment, so that relative probability means the probability of a choice divided by the probability of seeking employment. In your model, what do the symbols π_1, π_2 and π_3 represent? Answer in words. *Do not give formulas.*

(c) We want to know whether, controlling for GPA, being in the applied versus academic stream is related to plans after graduation. State the null hypothesis in symbols.

(d) We want to know whether, controlling for being in the applied versus academic stream, students with better GPA are more likely to choose university over seeking employment. State the null hypothesis in symbols.

Continued on page 7

14 points 6. In a study of agricultural productivity, small apple farms are randomly assigned to use one of three Pesticides (Type A, B or C) and one of three Fertilizers (Type 1, 2 or 3). The response variable is total crop yield in kilograms, and there are two covariates: number of trees on the farm, and crop yield last year.

(a) In the table below, fill in the definitions of the dummy variables for Pesticide and Fertilizer Use *effect coding* (the scheme with $0,\ 1,\ -1$). *Do not put any products in the table.*

Pesticide	*Fertilizer*	
A	1	
A	2	
A	3	
B	1	
B	2	
B	3	
C	1	
C	2	
C	3	

(b) Write $E(y|\mathbf{x})$ for a model that includes a possible Pesticide by Fertilizer interaction as well as their main effects. Denote the covariates by x_1 and x_2. Of course the vector $\mathbf{x}$ includes the dummy variables as well as x_1 and x_2. There are no interactions between the two covariates, or between covariates and factors.

(c) Give the null hypothesis you would test to answer each question below. The answers are in terms of the β parameters in your model. Some of the answers are the same. Except for the last one, assume that each question begins with "Controlling for number of trees and crop yield last year ..." .

Question	*Null Hypothesis*
Averaging over fertilizer types, does type of pesticide affect average crop yield?	
Does the effect of fertilizer type on crop yield depend on the type of pesticide used?	
Does the effect of pesticide type on crop yield depend on the type of fertilizer used?	
Averaging over pesticide types, does fertilizer type affect average crop yield?	
Test both covariates simultaneously, controlling for the main effects and the interaction.	

Continued on page 8

14 points 7. Salmon spend part of their lives in fresh water and part in salt water. It is possible to tell from the composition of their scales how much they grew in salt water and how much they grew in fresh water. Canadian and Alaskan salmon (pretend they are random samples) are either female or male. Denote the growth of a fish in the fresh water environment by y_1 and its growth in the salt water environment by y_2.

Think of this as a three-factor design. We will use the "multivariate" approach to repeated measures. In the multivariate approach, there is a regression model with effect coding for the between-cases factors, and the response variables are linear combinations of the observations that are available for each case.

(a) To carry out the analysis, you would use a just one regression model, with different linear combination response variables depending on the hypothesis being tested. Let x_1 be a dummy variable for species and x_2 be a dummy variable for sex. Give $E(L|\mathbf{x})$ below, where L stands for some linear combination of y_1 and y_2. Allow for the possibility of an interaction.

$E(L|\mathbf{x}) =$

(b) For each of the effects in the 3-way design, give the linear combination of the y_j that you would use as the response variable, and the null hypothesis in terms of β values from your regression equation.

Effect	**Linear combination**	**Null hypothesis**
Species		
Sex		
Environment		
Species × Sex		
Species × Environment		
Sex × Environment		
Species × Sex × Environment		

Continued on page 9

15 points 8. Please refer to the printout for the **Birthweight data**.

(a) You wish to test whether, controlling for the other variables, smoking is related to having a low birth weight baby.

i. Fill in the table below.

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

ii. In plain, non-statistical language, what do you conclude? No marks for this without the first part right. Start your answer with "Allowing for other possible risk factors . . . "

iii. Controlling for the other explanatory variables, the estimated odds of a low birth weight baby are ___ times as great for a mother who smokes. Write your answer (a number) in the space below.

(b) You wish to test whether, controlling for the other variables, race is related to having a low birth weight baby.

i. Fill in the table below.

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

ii. Give the *Bonferroni-corrected* p-values for the pairwise comparisons of racial groups.

Black versus White

Black versus Other

White versus Other

iii. Do the overall test and the Bonferroni-corrected pairwise comparisons point to the same conclusion? **Answer Yes or No**. Briefly comment.

Continued on page 10

12 points 9. Please refer to the printout for the **$\mathbf{CO_2}$ data**.

(a) Give the null hypothesis for the "Null Model Likelihood Ratio Test," in symbols.

(b) You want to know whether, averaging over CO_2 concentration and type of plant, chilling the plants had an effect on CO_2 uptake.

i. Fill in the table below.

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

ii. In plain, non-statistical language, what do you conclude? No marks for this without the first part right. Start your answer with "Averaging over . . . "

(c) Look at the plots on the last page of the printout. The plot on the left is Mississippi, and the plot on the left is Quebec. There is no colour on the plots, but the curves for the chilled plants are lower.

It appears that for the Mississippi plants, the shape of the curve relating ambient CO_2 concentration to mean CO_2 uptake (the shape, not just the level) differs depending on whether the plant was chilled. For the Quebec plants, the shapes of the curves are much more similar, though there is still a difference in level. One of the statistical tests reflects this difference between the two plots. Please fill in the table below.

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

Total Marks = 100 points