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

STUDENT #: _____ **SIGNATURE:** _____

UNIVERSITY OF TORONTO MISSISSAUGA
APRIL 2016 FINAL EXAMINATION
STA441H5S
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours
Aids: Calculator Model(s): Any calculator is okay.
Formula sheet and printouts will be 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	4	
3	12	
4	8	
5	13	
6	14	
7	14	
8	15	
9	15	
Total = 100 Points		

5 points

1. A High School principal is thinking of cancelling the Music programs to save money. The head of the Music department compares the overall grade point averages of students who participate in Band or Orchestra to the GPA of students do not. She finds that students in Band and Orchestra get better marks. She argues that musical training helps students do better in all their subjects. While it's true that music is important, the argument is flawed. Briefly explain why.

4 points

2. There are no elementary tests on this final exam, because all the elementary tests are special cases of more advanced methods we have covered. In each cell of the table below, write the *letter* of the most appropriate statistical method. The same letter can appear in more than one cell. This is very easy, so you will get full marks if all the answers are correct, and half marks if just one is wrong. Two or more wrong gets a zero.

	Reponse Variable		
Explanatory Variable	Categorical: Two Categories	Categorical: More than Two Categories	Quantitative
Categorical: Two Categories			
Categorical: More than Two Categories			
Quantitative			

(a) Multiple regression with normal errors, possibly with dummy variables.

(b) Logistic regression, possibly with dummy variables.

(c) Multinomial logistic regression (logistic regression with more than 2 outcomes), possibly with dummy variables.

continued on page 3

12 points

3. In a study of happiness, health and friendship, the response variable Y is reported happiness. For one analysis, the explanatory variables were reported number of hours per week spent socializing with friends (X_1), and a quantitative index of health status based on a complete physical exam, in which 100 indicates perfect health, and zero would indicate death (X_2). The investigators employ a regression model with

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2.$$

Estimating the β_j quantities, with b_j, they find that $b_1 > 0$ and $b_2 > 0$. The t-tests for these regression coefficients have $p = 0.002$ for b_1 and $p = 0.147$ for b_2. Also, $b_0 < 0$ and the p-value for b_0 equals 0.006.

(a) For any fixed number of hours per week spent socializing with friends, an increase of one point on the health status scale implies a true average increase of _____ units on the happiness measure.

(b) For any fixed value of health status, an increase of one hour per week spent socializing with friends implies a true increase of _____ units on the happiness measure.

(c) Consider the test of hours socializing controlling for health status.

i. What is the null hypothesis? Give the answer in symbols.

ii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

iii. Is the relationship with reported happiness statistically significant? Answer Yes or No.

iv. **Circle the answer**: Allowing for health status, people who spend more time socializing with friends tend to report that they are
- Happier.
- Less happy.
- No conclusion is justified.

(d) Consider the test of health status controlling for hours socializing.

i. What is the null hypothesis? Give the answer in symbols.

ii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

iii. Is the relationship with reported happiness statistically significant? Answer Yes or No.

iv. **Circle the answer**: Allowing for time spent socializing with friends, beople in better health tend to report that they are
- Happier.
- Less happy.
- No conclusion is justified.

(e) In simple, non-statistical language, what do you conclude from the test of $H_0 : \beta_0 = 0$?

continued on page 4

8 points

4. A researcher in Exercise Science suspects that the relationship of percent body fat to cardiovascular fitness to might be different for men and women. Fitness is measured in terms of the body's ability to use oxygen during exercise. Let Y = oxygen consumption (VO_2) as determined by a treadmill test, X_1 = percentage of body fat, and X_2 = Gender, coded 0 = Male and 1 = Female.

   (a) Write a regression equation in which the linear relationship of percent body fat to fitness to might depend on gender. Complete the equation below.

   $E(Y|\mathbf{X}) =$

   (b) Using your notation from Question 4a above, write the equations for the two regression lines (one for males and one for females) in the table below. Note: If the symbol X_2 appears in the table, the answer is wrong.

Female		
Male		

   (c) To see if the slopes are different for men and women, what null hypothesis would you test? Answer in terms of the β coefficients of your regression model.

   (d) Suppose that the population mean percent body fat for females is known to be μ_1, and the population mean percent body fat for males is known to be μ_2. You want to compare the average cardiovascular fitness for males of average percent body fat (for males) to the average cardiovascular fitness for females of average percent body fat (for females). Using the notation of your regression model from Question 4a, what is the null hypothesis? **Circle your final answer**.

continued on page 5

13 points

5. For a particular form of cancer that is not very aggressive, the standard treatment is a combination of chemotherapy and radiation therapy. Both chemotherapy and radiation have serious side effects. Some patients may be so weakened by the treatment that they die from other things (such as infections) that are apparently unrelated to the cancer.

   Volunteer patents who were considering no treatment at all were randomly assigned to one of three experimental conditions. They received either Chemotherapy only, Radiation only, or Both treatments. The response variable is five-year survival. Age is an important predictor of survival, and is used as a covariate.

   (a) Write the logistic regression equation, denoting the probability of survival by π. There should be *no interactions* in the model. You do not need to say how your dummy variables are defined. You will do that in the next part. Complete the equation below.

   $\ln\left(\frac{\pi}{1-\pi}\right) =$

   (b) In the table below, make columns showing how your dummy variables are defined. In the last column, write the *odds* of five-year survival, using the notation of your logistic regression model from Question 5a above. If *symbols* for your dummy variables appear in the last column, the answer is wrong.

		Odds
Chemotherapy		
Radiation		
Both		

   (c) In the notation of your model, what are the five-year survival odds for a 25-year-old patient receiving both radiation and chemotherapy?

   (d) In the notation of your model, what is the five-year survival *probability* for a 25-year-old patient receiving both radiation and chemotherapy?

   (e) For a 60-year-old patient receiving radiation only, the odds of surviving five years are _____ times as great as the odds for a 60-year-old receiving both radiation and chemotherapy. Answer in terms of the Greek letters from your model.

   (f) You want to know whether it is better for patients to get just radiation or just chemotherapy. What is the null hypothesis? Answer in terms of the Greek letters from your model.

continued on page 6

14 points

6. Please refer to the TV Data part of the computer printouts. Recall that Stevens County is divided into 75 districts including rural, small-town and urban areas. For each of 500 households interviewed, the data file contains district number, household number within district, assessed value of home in US dollars (an indirect measure of income, which was not asked), and answers to 9 questions related to the respondents' interest in getting cable TV.

   (a) Give $E(Y|\mathbf{X})$ corresponding to the `model` statement in `proc reg`. The explanatory variables must be in the right order.

   $E(Y|\mathbf{X}) =$

   (b) Controlling for number of TV sets, hours of TV watched last week and assessed value of home, is Location (Rural, Small town, City) related to price willing to pay for cable TV?

      i. In Greek letters from your answer to Question 6a, what is the null hypothesis?

      ii. Give the value of the test statistic. The answer is a number from the printout.

      iii. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

      iv. Are the results statistically significant? Answer Yes or No.

      v. Guided strictly by the $\alpha = 0.05$ significance level but without mentioning it directly, what do you conclude? Use plain, non-statistical language.

   (c) Controlling for other explanatory variables, what proportion of the remaining variation in price willing to pay for cable is explained by assessed value of the home? The answer is a number. Show a little work. **Circle your answer**.

   (d) For a small-town household that is average on number of TV sets, hours of TV watched last week and assessed value of home, what is the predicted amount they would be willing to pay for cable TV? The answer is a number **in dollars**. Show a little work. **Circle your answer**.

continued on page 7

14 points

7. Please refer to the Donner Party Data part of the computer printouts. The Donner party were a group of American pioneers who got lost in the mountains. Only some of them came back. The variables are Age, Sex and Survival.

   (a) Write a linear model for the log odds, corresponding to the `model` statement in `proc logistic`. The explanatory variables must be in the right order.

   $\ln\left(\frac{\pi}{1-\pi}\right) =$

   (b) What is $\widehat{\beta}_2$? The answer is a number from the printout.

   (c) Controlling for age, were the chances of survival different for men and women?

   i. In Greek letters from your answer to Question 7a, what is the null hypothesis?

   ii. Give the value of the test statistic. The answer is a number from the printout.

   iii. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

   iv. Are the results statistically significant? Answer Yes or No.

   v. Guided strictly by the $\alpha = 0.05$ significance level but without mentioning it directly, what do you conclude? Use plain, non-statistical language.

   (d) Controlling for age, the estimated odds of survival were _____ times as great for women. The answer is a number from the printout.

   (e) What is the estimated *probability* of survival for a 20-year old man? The answer is a number. Show a little work. **Circle your answer**.

continued on page 8

15 points

8. Please refer to the Math Data part of the computer printouts. First-year university students signed up for either a Mainstream calculus course, an Elite Course or a Catch-up course. Their High School Grade Point Average is available as a predictor, and also their marks in High School Calculus and High School English.

   (a) In the space below, write a collection of regression-like equations corresponding to the `model` statement in `proc logistic`. First take a quick look at the results file. Let the *numbers* in π_1, π_2, π_3 correspond to the "Ordered Values" in the "Response Profile." Also, your explanatory variables must be in the same order as ones in the `model` statement.

   (b) What is $\widehat{\beta}_{2,1}$? The answer is a number from your printout.

   (c) Look at the table entitled `Testing Global Null Hypothesis: BETA=0`. In Greek letters from your answer to Question 8a, what is the null hypothesis?

   (d) In the table entitled `Type 3 Analysis of Effects`, what is the null hypothesis for the test of `hscalc`? Give the answer in terms of Greek letters from your answer to Question 8a.

   (e) Look at the tests for `hscalc` in the table entitled `Analysis of Maximum Likelihood Estimates`. Guided strictly by the $\alpha = 0.05$ significance level but without mentioning it directly, what do you conclude? Use plain, non-statistical language.

continued on page 9

15 points

9. The Beat the Blues data come from a longitudinal clinical trial of an interactive, multimedia program designed to deliver cognitive behavioural therapy to depressed patients via a computer terminal. Patients with depression recruited in primary care were randomised to either the Beating the Blues treatment, or to "Treatment as Usual" (TAU). Variables include treatment condition, whether they are taking anti-depressant drugs, how long their current episode of depression has lasted, and score on the Beck Depression Inventory before treatment began and after 2, 4, 6 and 8 months.

   (a) In the `model` statement of `proc mixed`, what are the factors?

   (b) Classify each factor as within or between cases.

   (c) Look at the table entitled `Null Model Likelihood Ratio Test`. In Greek letters from the formula sheet, what is the null hypothesis?

   (d) Look at the F-tests for the three-factor ANOVA. Sticking strictly to the 0.05 level, only two of them are significant. In plain, non-statistical language, what do you conclude from these two tests? Remember, depression measurements were subtracted from the pre-test, so that the response variable Change represents *improvement* (decrease in reported depression).