UNIVERSITY OF TORONTO MISSISSAUGA

APRIL 2012 FINAL EXAMINATION
STA442H5S
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours

Aids: Calculator Model(s): Any calculator is okay

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, including but not limited to any electronic devices with storage, such as cell phones, pagers, personal digital assistants (PDAs), iPods, and MP3 players. Unauthorized calculators and notes are also not permitted. Do not have any of these items in your possession in the area of your desk. Please turn the electronics off and put all unauthorized aids with your belongings at the front of the room before the examination begins. If any of these items are kept with you during the writing of your exam, you may be charged with an academic offence. A typical penalty may cause you to fail the course.

Please note, you **CANNOT** petition to **re-write** an examination once the exam has begun.

Last/Surname (Print): ______________________________

First/Given Name (Print): ______________________________

Student Number: ______________________________

Signature: ______________________________

$$F = \left(\frac{n-p}{s}\right)\left(\frac{a}{1-a}\right)$$

$$a = \frac{sF}{n-p+sF}$$

Qn. #	Value	Score
1	3	
2	7	
3	4	
4	4	
5	16	
6	13	
7	14	
8	11	
9	18	
10	10	

Total = 100 Points

3 points

1. In the table below, write a check mark ($\surd$) in each cell where the appropriate elementary statistical method is actually some kind of multiple regression (including multivariate regression, if need be). You do *not* need to give the name of the elementary method, and no explanation is necessary.

	Dependent Variable		
Independent Variable	Categorical: Two Categories	Categorical: More than Two Categories	Quantitative
Categorical: Two Categories			
Categorical: More than Two Categories			
Quantitative			

7 points

2. Answer each question below True or False. Write "T" or "F" on the line. You will get full marks on this question if you answer at least 10 out of 13 correct. If you miss more than 3, you will get zero marks.

____ We seek to predict the independent variable from the dependent variable.

____ The p-value is the probability that the null hypothesis is true.

____ When a relationship between the independent variable and the dependent variable is statistically significant, we conclude there is evidence that the two variables are actually related.

____ The p-value is the probability of failing to replicate significant results in a second independent random sample of the same size.

____ The greater the p-value, the stronger the evidence that the independent and dependent variable are related.

____ If a subject (case) provides data for more than one value of an independent variable, we call that independent variable a *within-subjects* variable.

____ We observe $r = -0.70$, $p = .009$. We conclude that that high values of X tend to go with low values of Y and low values of X tend to go with high values of Y.

____ If $p < .05$ we say the results are statistically significant at the .05 level, and we conclude that the independent variable and the dependent variable are unrelated.

____ In a study attempting to predict income from education and race, there is a significant interaction between education and race. This means that income and race are related.

____ When you add another independent variable in multiple regression, R^2 cannot go down.

____ We observe $r = 0.50$, $p = .002$. This means that 50% of the variation in the dependent variable is explained by a linear relationship with the independent variable.

____ An experimental study is one in which cases are randomly assigned to the different values of an independent variable.

____ A multivariate analysis is one with multiple dependent variables.

continued on page 3

4 points 3. Make up an *original* example of a study for which an appropriate analysis would be a two-factor ANOVA with one of the factors within-cases and the other factor between-cases. The word "original" means that your example should not be too similar to any of the examples in lecture, homework or the text.

4 points 4. A medical study finds that the more organic food a person eats (as a percent of total calories), the better the person's overall health on average. Briefly discuss at least one confounding variable that could have produced this result.

continued on page 4

16 points 5. 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 dependent 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 (p_1 and p_2), and the dummy variables for Fertilizer (f_1 and f_2). Use *effect coding* (the scheme with 0, 1, -1).

Pesticide	*Fertilizer*	p_1	p_2	f_1	f_2
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 p_1, p_2 and so on 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*
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?	
Does fertilizer type affect average crop yield?	
Is there a main effect for pesticide type?	
Is there a main effect for fertilizer type?	
Is there an interaction between fertilizer type and pesticide type?	
Test both main effects and the interaction, all at the same time.	
Test both covariates simultaneously, controlling for the main effects and the interaction.	

continued on page 5

13 points

6. In a study of the effects of combining two blood pressure drugs, patients with high blood pressure are randomly assigned to one of four treatment conditions. They get either Drug A or a placebo, and they get either Drug B or a placebo. So each patient takes two pills a day for 6 weeks. Their blood pressure after 6 weeks is the dependent variable. Age is a covariate. Use this regression model for the problem:

$$E[Y|\mathbf{X}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5, \text{ where}$$

- $x_1 = 1$ if the patient got both drugs. Otherwise, $x_1 = 0$.
- $x_2 = 1$ if the patient got Drug A but not B. Otherwise, $x_2 = 0$.
- $x_3 = 1$ if the patient got Drug B but not A. Otherwise, $x_3 = 0$.
- $x_4 = 1$ if the patient got neither drug. Otherwise, $x_4 = 0$.
- $x_5 =$ the patient's age.

(a) Write $E[Y|\mathbf{X}]$ for each treatment combination in the table below.

	Drug B	
Drug A	*Yes*	*No*
Yes		
No		

(b) Give the null hypothesis you would test to answer each question below. The answers are in terms of the β parameters of the regression model. Some of the answers are the same. You may assume that each question begins with "Controlling for the patient's age, . . . "

Question	*Null Hypothesis*
Averaging across Drug A versus Placebo, does Drug B affect blood pressure?	
Does the effect of Drug A on blood pressure depend on whether the patient is also taking Drug B?	
Is it better to take both drugs, or neither drug?	
Is it better to take just Drug A, or just Drug B?	
Is Drug A alone better than nothing?	
Is Drug B alone better than nothing?	
Is there a main effect for Drug A?	
Is there a main effect for Drug B?	
Is the average response to just Drug A and just Drug B different from the response to both drugs at once?	
Is there a statistically significant interaction?	
Test both main effects and the interaction, all at the same time.	

continued on page 6

14 points 7. Please refer to the printout for the `furnace` data.

(a) For each type of chimney liner, what is the estimated energy consumption with vent damper active for a house that is average on chimney area and average on energy consumption with vent damper inactive? Write your answers in the boxes below; Two decimal places of accuracy will be enough. Please show a little work below the table. You have more room than you need.

Unlined	*Tile*	*Metal*

(b) Controlling for chimney area and energy consumption with vent damper inactive, is there evidence that type of chimney liner is related to energy consumption with vent damper in?

i. Answer Yes, No or "Information not on printout."

ii. Give the value of the test statistic. The answer is a number. If the answer is not on the printout, write,"Information not on printout."

iii. Give the p-value. The answer is a number. If the answer is not on the printout, write,"Information not on printout."

(c) After allowing for energy consumption with vent damper inactive and chimney area, what proportion of the *remaining* variation in energy consumption with vent damper active is explained by type of chimney liner? The answer is a number between zero and one. Show a little work and **circle your answer**.

(d) In the table below, write Bonferroni-corrected p-values for the pairwise tests comparing chimney liners after controlling for chimney area and energy consumption with vent damper inactive. If you want to show any work (not required), please show it below the table.

	Unlined	*Tile*	*Metal*
Unlined	×		
Tile	×	×	
Metal	×	×	×

continued on page 7

(e) Based on the results of the Bonferroni-corrected pairwise comparisons, state your conclusion in simple, non-technical language. Start with "Allowing for . . . ". My answer is one sentence.

11 points 8. Please refer to the printout for the `salmon` data.

(a) State what the factors are in this study. Classify each one as within-cases or between-cases.

(b) If you look at the results, you can see that basically nothing is going on with gender, so we will concentrate on the two-way table of marginal means for the other two factors. In the table below, fill in the sample cell means, which are marginal means averaging over gender.

	Environment	
Country	*Freshwater*	*Seawater*
Alaskan		
Canadian		

(c) Averaging across gender, does the average difference between growth in freshwater and seawater environments depend on country?

i. What is the value of the test statistic? The answer is a single number.

ii. What is the p-value? The answer is a number or possibly a range of numbers.

iii. Are the results statistically significant at the 0.05 level? Answer Yes or No.

iv. Describe the results in plain, non-statistical language. One sentence is enough. You have more room than you need.

continued on page 8

18 points

9. Please refer to the printout for the `poverty` data. For each dependent variable, the model implies six regression lines relating GNP (Gross National Product per capita: Roughly, wealth) to the expected value of the dependent variable.

   (a) There is a test for parallel sets of regression lines for all the dependent variables simultaneously. Give the following:

   i. F statistic

   ii. p value

   iii. Is there evidence that the regression lines are non-parallel for at least one dependent variable? Answer Yes or No.

   iv. With a Bonferroni correction, which dependent variables show evidence of non-parallel regression lines? Just list the variables. No explanation is needed. "All" is an acceptable answer if it is justified by the printout. So is "None."

   (b) Focus just on the results for *infant mortality rate.*

   i. Write the estimated regression equation for industrialized nations:

   $\widehat{Y}\ =$

   ii. Is there evidence that GNP is related to infant mortality rate among the industrialized nations? Write Yes or No and give the p-value.

   iii. Write the estimated regression equation for African nations:

   $\widehat{Y}\ =$

   iv. On the printout, there is a test for difference between African nations and industrialized nations in the slopes of the lines relating GNP to expected infant mortality rate. Just give the p-value.

   v. Why does it make sense that the slopes of all these regression lines are negative?

continued on page 9

10 points 10. For the `mantids` data, there are two statistically significant main effects. In plain, non-statistical language, say what each one means. Here, you are being asked for *two separate answers.* Still, you have a lot more room than you need.

Total Marks = 100 points