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

STUDENT #: _____ **SIGNATURE:** _____

UNIVERSITY OF TORONTO MISSISSAUGA
DECEMBER 2015 FINAL EXAMINATION
STA302H5F
Regression Analysis
Jerry Brunner
Duration - 3 hours
Aids: Calculator Model(s): Any calculator is okay. Formula sheet 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, you **CANNOT** petition to **re-write** an examination once the exam has begun.*

Qn. #	Value	Score
1	8	
2	10	
3	12	
4	12	
5	12	
6	12	
7	10	
8	24	
Total = 100 Points		

Seat Position

Unless otherwise indicated, the questions on this exam refer to the general linear model on the formula sheet, with $\mathbf{X}$ an $n \times (k+1)$ matrix of fixed, observable constants. If the question does not specifically mention linear independence, assume that the columns of the $\mathbf{X}$ matrix are linearly independent. On this exam (and in real applications), $n > k+1$, so that $\mathbf{X}$ is *not a square matrix*, and *does not have an inverse*.

8 points

1. Suppose you want to predict score on the final exam from $k = 5$ quiz marks, using a multiple regression model with an intercept. But everyone got 10 out of 10 on Quiz 1.

   (a) Using the definition on the formula sheet, either (1) Prove that the columns of the $\mathbf{X}$ matrix are linearly independent, or (2) Prove that the columns of the $\mathbf{X}$ matrix are linearly dependent. *You really do have to use the definition on the formula sheet* in order to get any marks.

   (b) Is it possible to compute $\widehat{\boldsymbol{\beta}}$? Answer Yes or No. If the answer is Yes, give the formula you would use. If the answer is No, *very* briefly explain.

Continued on page 3

10 points

2. Let $\mathbf{y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{w} = \mathbf{A}\mathbf{y}$, where $\mathbf{A}$ is an $r \times p$ matrix of constants. Use moment-generating functions to obtain the distribution of $\mathbf{w}$. Show your work. You may use anything from the formula sheet except the specific result you are proving. You have more room than you need.

Continued on page 4

12 points

3. In a market research study, volunteer consumers are randomly assigned to view one of three advertisements, and then they rate their interest in purchasing the product. The regression model uses *cell means coding*, which is the model with no intercept and a zero-one indicator dummy variable for each category of the independent variable.

   To introduce some additional notation, the numbers of consumers viewing each advertisement are n_1, n_2 and n_3 (so that $n = n_1 + n_2 + n_3$), and the corresponding sample mean interest ratings are $\overline{y}_1$, $\overline{y}_2$ and $\overline{y}_3$.

   (a) What are the dimensions of the $\mathbf{X}'\mathbf{X}$ matrix? Give the number of rows and the number of columns.

   (b) What is the $\mathbf{X}'\mathbf{X}$ matrix? It has a very simple form.

   (c) What is $(\mathbf{X}'\mathbf{X})^{-1}$? You can just write it down.

Continued on page 5

(d) What are the dimensions of the $\mathbf{X}'\mathbf{y}$ matrix? Give the number of rows and the number of columns.

(e) What is the $\mathbf{X}'\mathbf{y}$ matrix? Give your answer in terms of n_1, n_2, n_3, $\overline{y}_1$, $\overline{y}_2$ and $\overline{y}_3$.

(f) What are the dimensions of the $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ matrix? Give the number of rows and the number of columns.

(g) What is $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$? The answer is simple, and it is very reasonable if you think about it.

Continued on page 6

12 points

4. For the general linear regression model, suppose we want to estimate the single linear combination $\mathbf{a}'\boldsymbol{\beta}$ based on sample data. The Gauss-Markov Theorem tells us that the most natural choice is also (in a sense) the best choice. You will supply the core of the proof. You may use the following preparation *without proof or additional discussion.*

   - Note that $\mathbf{a}'\widehat{\boldsymbol{\beta}}$ is an unbiased estimator of $\mathbf{a}'\boldsymbol{\beta}$. It is termed a *linear* unbiased estimator because it is a linear combination of the y values: $\mathbf{a}'\widehat{\boldsymbol{\beta}} = \mathbf{c}_0'\mathbf{y}$, where $\mathbf{c}_0' = \mathbf{a}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
   - We will compare the variance of $\mathbf{c}_0'\mathbf{y}$ to the variance of other linear unbiased estimators of the form $\mathbf{c}'\mathbf{y}$.
   - Unbiased means $E(\mathbf{c}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, so we have $\mathbf{c}'\mathbf{X} = \mathbf{a}' \Leftrightarrow \mathbf{a} = \mathbf{X}'\mathbf{c}$.
   - $Var(\mathbf{c}'\mathbf{y}) = \sigma^2\mathbf{c}'\mathbf{c}$, and $Var(\mathbf{c}_0'\mathbf{y}) = \sigma^2\mathbf{c}_0'\mathbf{c}_0$.
   - Thus, all we have to show is that $\mathbf{c}'\mathbf{c} > \mathbf{c}_0'\mathbf{c}_0$ for any $n \times 1$ vector $\mathbf{c} \neq \mathbf{c}_0$ satisfying $\mathbf{X}'\mathbf{c} = \mathbf{a}$.

   Continue the proof from here.

Continued on page 7

Continued on page 8

12 points

5. Often there is more than one equivalent way to state a null hypothesis, and we take it on faith that how we express the null hypothesis has no effect on the test. Is this faith justified?

   Suppose the null hypothesis for the general linear test is $H_0 : \mathbf{C}_1\boldsymbol{\beta} = \mathbf{t}$, where $\mathbf{C}_1$ is a q by $k+1$ matrix with linearly independent rows. Now let $\mathbf{C}_2 = \mathbf{A}\mathbf{C}_1$, where $\mathbf{A}$ is a $q \times q$ matrix with an inverse.

   (a) Show that $\mathbf{C}_1\boldsymbol{\beta} = \mathbf{t}$ implies $\mathbf{C}_2\boldsymbol{\beta} = \mathbf{A}\mathbf{t}$. This is very simple.

   (b) Show that $\mathbf{C}_2\boldsymbol{\beta} = \mathbf{A}\mathbf{t}$ implies $\mathbf{C}_1\boldsymbol{\beta} = \mathbf{t}$. This is simple too.

   (c) Calculate the F statistic for testing $H_0 : \mathbf{C}_2\boldsymbol{\beta} = \mathbf{A}\mathbf{t}$, and simplify. Is it the same as the F statistic for testing $H_0 : \mathbf{C}_1\boldsymbol{\beta} = \mathbf{t}$? Actually, all you need to calculate is the numerator. Finish your answer with the word "Same" or "Different."

Continued on page 9

Continued on page 10

12 points

6. Suppose you have a random sample from a univariate normal distribution. If someone randomly selected another observation from this distribution and asked you to guess what it was, your answer would surely be the mean of your sample. But what if you were asked for a prediction *interval*?

   Accordingly, let $y_1, \ldots, y_n \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$, and let y_0 denote another observation independently sampled from this distribution. You already know that $\overline{y}$ is normally distributed, $\frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \overline{y})^2 = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$, and that $\overline{y}$ and S^2 are independent.

   A random sample of size 10 yields $\overline{y} = 51.86$ and $S^2 = 117.74$. The critical value is $t_{0.025} = 2.26$. Give a 95% prediction interval for y_0. Show the work required to get the answer. You don't need to give every detail; for example, you need not mention independence. The final answer is a pair of numbers, a lower prediction limit and an upper prediction limit. **Circle the numbers**.

Continued on page 11

Continued on page 12

10 points

7. The following is a critical ingredient of all the tests and confidence intervals used in this course. Find the distribution of the random variable $Z = \frac{1}{\sigma^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$. Prove your answer, citing facts from the formula sheet as you use them. Don't forget the degrees of freedom. You have more room than you need.

Continued on page 13

8. (*24 points*) Pigs are routinely given large doses of antibiotics even when they show no signs of illness, to protect their health under unsanitary conditions. Pigs were randomly assigned to one of three antibiotic drugs. Dressed weight (weight of the pig after slaughter and removal of head, intestines and skin) was the dependent variable. Independent variables are Drug type, Mother's live adult weight and Father's live adult weight. Please answer the questions below based on the R printout.

   (a) Write the regression equation for the full model, including ϵ_i. This is the regression equation *used in the R printout.*

   (b) Make a table with one row for every drug, with columns showing how the dummy variables were defined. Make another column giving $E(y|\mathbf{x})$ for each drug.

   (c) Predict the dressed weight of a pig getting Drug 2, whose mother weighed 140 pounds, and whose father weighed 185 pounds. Your answer is a single number. **Circle the number**.

Continued on page 14

(d) This parallel planes regression model specifies that the differences in expected weight for the different drug treatments are the same for every possible combination of mother's weight and father's weight. Give a 95% confidence interval for the difference in expected weight between drug treatments 2 and 3. Show some calculations. The final answer is a pair of numbers, a lower confidence limit and an upper confidence limit. **Circle the numbers**. You have more room than you need.

Continued on page 15

(e) In symbols, give the null hypotheses you would test to answer the following questions. Your answers are statements involving the β values from your regression equation.

i. Controlling for mother's weight and father's weight, does type of drug have an effect on the expected weight of a pig?

ii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 2?

iii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 3?

iv. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 2 or Drug 3?

(f) For each of the questions below, give the value of the t or F statistic (a number from the printout), and indicate whether or not you reject the null hypothesis.

i. Controlling for mother's weight and father's weight, does type of drug have an effect on the expected weight of a pig?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

ii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 2?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iii. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 1 or Drug 3?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iv. Controlling for mother's weight and father's weight, which drug helps the average pig gain more weight, Drug 2 or Drug 3?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

Continued on page 16

v. Allowing for which drug they were given, does expected weight of a pig increase faster as a function of the mother's weight, or does it increase faster as a function of the father's weight?

Test Statistic Value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

(g) We can assume that farmers want their pigs to weigh a lot. In plain, non-statistical language, can you offer some advice to a farmer based on these data? Remember, the farmer must be able to understand your answer or it is worthless.