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

STUDENT #: _____ **SIGNATURE:** _____

UNIVERSITY OF TORONTO MISSISSAUGA
DECEMBER 2017 FINAL EXAMINATION
STA302H5F
Regression Analysis
Jerry Brunner
Duration - 3 hours
Aids: Calculator Model(s): Any calculator is okay ; Formula sheet provided

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	8	
2	16	
3	12	
4	20	
5	16	
6	28	
Total = 100 Points		

Seat Position

8 points

1. In the linear model $\mathbf{y} = X\boldsymbol{\beta}+\boldsymbol{\epsilon}$, suppose that the columns of X are linearly dependent. Prove that $(X'X)^{-1}$ does not exist (so we can't do most of what we've done in this course). I will start the proof for you.

   "Columns of X linearly dependent means there is a *non-zero* vector $\mathbf{v} \in \mathbb{R}^{k+1}$ with $X\mathbf{v} = \mathbf{0}$. To produce a contradiction, suppose that $(X'X)^{-1}$ exists."

   Now you complete the proof. You have more room than you need.

Continued on page 3

16 points

2. Let $\mathbf{y} = [y_j]$ be a $p \times 1$ random vector with expected value $\boldsymbol{\mu}$ and covariance matrix Σ.

   (a) Show that Σ is non-negative definite.

   (b) Show that $\sum_{j=1}^{p} Var(y_j)$ equals the sum of the eigenvalues of Σ.

Continued on page 4

12 points

3. Let $\mathbf{y} \in \mathbb{R}^p$ be a multivariate normal random vector with expected value $\boldsymbol{\mu}$ and positive definite covariance matrix Σ.

   (a) Let $\mathbf{z} = \Sigma^{-1/2}(\mathbf{y} - \boldsymbol{\mu})$. Find the moment-generating function of $\mathbf{z}$. Show your work. Simplify.

   (b) Based on the moment-generating function of $\mathbf{z}$, state its distribution.

Continued on page 5

20 points

4. For the general linear regression model, suppose we want to estimate $\boldsymbol{\ell}'\boldsymbol{\beta}$ based on sample data, where $\boldsymbol{\ell}$ is a $(k+1) \times 1$ vector of constants. The Gauss-Markov Theorem tells us that the most natural choice is also the best choice in a large class of possible estimators. You will supply the core of the proof, following an approach that is *different* from the one taken in lecture and the text.

   As preparation, note that $\boldsymbol{\ell}'\mathbf{b}$ is an unbiased estimator of $\boldsymbol{\ell}'\boldsymbol{\beta}$. It is termed a *linear* unbiased estimator because it is a linear combination of the y values: $\boldsymbol{\ell}'\mathbf{b} = \mathbf{c}_0'\mathbf{y}$, where $\mathbf{c}_0' = \boldsymbol{\ell}'(X'X)^{-1}X'$ (and $\mathbf{c}_0 = X(X'X)^{-1}\boldsymbol{\ell}$).

   (a) What is $Var(\mathbf{c}'\mathbf{y})$? Show the calculation in matrix notation.

   (b) Unbiased means $E(\mathbf{c}'\mathbf{y}) = \boldsymbol{\ell}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$. We have seen that this implies $\boldsymbol{\ell} = \mathbf{X}'\mathbf{c}$. The proof of this part is rather involved and you can skip it. Do the following instead.

   Bearing in mind that the hat matrix H is a projection operator, verify that the projection of $\mathbf{c}$ onto the space spanned by the columns of the X matrix is none other than $\mathbf{c}_0$.

Continued on page 6

(c) Based on your answer to Question 4a, you need to show that for a general $n \times 1$ vector $\mathbf{c}$ satisfying $\boldsymbol{\ell} = \mathbf{X}'\mathbf{c}$, $\mathbf{c}'\mathbf{c}$ is minimized when $\mathbf{c} = \mathbf{c}_0$. Start by subtracting off the projection and adding it back on, like this:

$\mathbf{c}'\mathbf{c} = (\mathbf{c}-\mathbf{c}_0+\mathbf{c}_0)'(\mathbf{c}-\mathbf{c}_0+\mathbf{c}_0) = \ldots$

Continue the calculation and complete the proof, keeping $\mathbf{c} - \mathbf{c}_0$ together.

Continued on page 7

16 points

5. The prediction interval for a new observation y_0 is based on the t distribution. This question guides you through the derivation of the t statistic on the formula sheet. At several points you will use independence, but to it make easier *you do not need to mention independence* in order to get full marks.

   Recall that $\mathbf{x}_0$ is the vector of independent variable values for the new observation, possibly including a one in the first position to pick up the intercept.

   (a) What is the distribution of y_0? Just write down the answer.

   (b) What is the distribution of $\mathbf{x}_0'\mathbf{b}$?

   (c) What is the distribution of $y_0 - \mathbf{x}_0'\mathbf{b}$?

   (d) Now standardize to obtain z.

Continued on page 8

(e) Divide z from the last part by the square root of a well-chosen chi-squared random variable, divided by its degrees of freedom. Simplify, obtaining the expression on the formula sheet. You have more room than you need.

Continued on page 9

28 points

6. This last part of the exam is based on the census tract data. The questions come after the R printout.

```
> rm(list=ls()); # options(scipen=999) # No scientific notation
> census =
+ read.table("http://www.utstat.toronto.edu/~brunner/data/illegal/CensusTract.data.txt")
> attach(census)
> # Calculate rates: Divide by population size, yielding number per 1,000 people
> doc_rate = docs/pop
> bed_rate = beds/pop
> labor_rate = labor/pop
> ave_income = income/pop # In thousands of dollars per person
> crime_rate = crimes/pop
> density = pop/area # Thousands of people per square mile
> # Make region a factor
> region = factor(region,labels=c("Northeast","NorthCentral","South","West"))
> contrasts(region) # Note 1 = Northeast is still alphabetically first
             NorthCentral South West
Northeast               0     0    0
NorthCentral            1     0    0
South                   0     1    0
West                    0     0    1
> source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/ftest.txt")
>
> # Based on stepwise, we will use these variables, except centered
> examdata = data.frame(region, ave_income, density, hs, crime_rate)
> summary(examdata)
          region     ave_income       density              hs          crime_rate
 Northeast   :27   Min.   :3.315   Min.   : 0.0361   Min.   :30.30   Min.   :15.69
 NorthCentral:35   1st Qu.:5.853   1st Qu.: 0.1793   1st Qu.:50.00   1st Qu.:45.97
 South       :51   Median :6.357   Median : 0.3026   Median :53.90   Median :56.55
 West        :28   Mean   :6.380   Mean   : 0.5776   Mean   :54.54   Mean   :55.87
                   3rd Qu.:6.883   3rd Qu.: 0.4877   3rd Qu.:59.90   3rd Qu.:64.38
                   Max.   :8.461   Max.   :12.0000   Max.   :72.80   Max.   :93.58
>
> # Create centered quantitative variables
> ave_incomeCen = ave_income - mean(ave_income)
> densityCen = density - mean(density)
> hsCen = hs - mean(hs)
>
> full = lm(crime_rate ~ region + ave_incomeCen + densityCen + hsCen)
> summary(full)

Call:
lm(formula = crime_rate ~ region + ave_incomeCen + densityCen +
    hsCen)

Residuals:
    Min      1Q  Median      3Q     Max
-24.014  -8.338  -1.512   7.237  25.313
```

Continued on page 10

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)         40.2259     2.1762  18.485  < 2e-16 ***
regionNorthCentral  11.1311     2.8472   3.909 0.000146 ***
regionSouth         20.7775     2.7465   7.565 5.55e-12 ***
regionWest          27.0058     3.3713   8.010 4.91e-13 ***
ave_incomeCen        3.7583     1.5028   2.501 0.013591 *
densityCen           2.1941     0.8583   2.556 0.011693 *
hsCen                0.3116     0.1586   1.965 0.051536 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 10.55 on 134 degrees of freedom
Multiple R-squared:  0.489,Adjusted R-squared:  0.4661
F-statistic: 21.37 on 6 and 134 DF,  p-value: < 2.2e-16

>
> # Now some reduced models and hypothesis matrices
> # for F-tests of H0: C beta = gamma
>
> red1 = lm(crime_rate ~ region)
> red2 = lm(crime_rate ~ ave_incomeCen + densityCen + hsCen)
>
> C1 = rbind(c(0, 1, 0, 0, 0, 0, 0),
+            c(0, 0, 1, 0, 0, 0, 0),
+            c(0, 0, 0, 1, 0, 0, 0) )
>
> C2 =   cbind(0, 1,-1, 0, 0, 0, 0)
> C3 =   cbind(0, 1, 0,-1, 0, 0, 0)
> C4 =   cbind(0, 0, 1,-1, 0, 0, 0)
>
> C5 = rbind(c(0, 0, 0, 0, 1, 0, 0),
+            c(0, 0, 0, 0, 0, 1, 0),
+            c(0, 0, 0, 0, 0, 0, 1) )
>
> # Now the F-tests
>
> anova(red1,full)
Analysis of Variance Table

Model 1: crime_rate ~ region
Model 2: crime_rate ~ region + ave_incomeCen + densityCen + hsCen
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    137 18010
2    134 14903  3    3106.5 9.3105 1.235e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> anova(red2,full)
Analysis of Variance Table

Model 1: crime_rate ~ ave_incomeCen + densityCen + hsCen
Model 2: crime_rate ~ region + ave_incomeCen + densityCen + hsCen
  Res.Df   RSS Df Sum of Sq      F    Pr(>F)
1    137 24789
2    134 14903  3    9885.6 29.628 9.282e-15 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Continued on page 11

```
> # Round ftest output to 5 decimal places -- easier to read
> round( ftest(full,C1), 4)
       F      df1      df2  p-value
 29.6284   3.0000 134.0000   0.0000

> round( ftest(full,C2), 4)
       F      df1      df2  p-value
 14.0522   1.0000 134.0000   0.0003

> round( ftest(full,C3), 4)
       F      df1      df2  p-value
 29.6661   1.0000 134.0000   0.0000

> round( ftest(full,C4), 4)
       F      df1      df2  p-value
  3.9949   1.0000 134.0000   0.0477

> round( ftest(full,C5), 4)
       F      df1      df2  p-value
  9.3105   3.0000 134.0000   0.0000
```

(a) In the following table, write estimates of expected crime rate when average income, population density and percent high school graduates are set to their sample mean values. These numbers could be described to a journalist as "Adjusted crime rate, in serious crimes per thousand." Fill in the table. Two decimal places of accuracy are sufficient.

Region	**Adjusted Crime Rate**
Northeast	
North Central	
South	
West	

(b) We want to know whether, controlling for region, the other independent variables taken together are related to crime rate. This is one test.

i. State the null hypothesis in terms of β values from R's model.

ii. Fill in the table.

Test Statistic (t or F) Value: A number	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iii. Controlling for region, are any of the other variables (one or more) related to crime rate? Answer Yes or No.

Continued on page 12

(c) Allowing for region, average income and population density, is percent of high school graduates related to crime rate?

i. State the null hypothesis in terms of β values from R's model.

ii. Fill in the table.

Test Statistic (t or F) Value: A number	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iii. In plain, non-statistical language, what do you conclude?

(d) Correcting for region, average income and percent of high school graduates, is population density related to crime rate?

i. State the null hypothesis in terms of β values from R's model.

ii. Fill in the table.

Test Statistic (t or F) Value: A number	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iii. In plain, non-statistical language, what do you conclude?

(e) Taking region, population density and percent of high school graduates into account, is average income related to crime rate?

i. State the null hypothesis in terms of β values from R's model.

ii. Fill in the table.

Test Statistic (t or F) Value: A number	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iii. In plain, non-statistical language, what do you conclude?

Continued on page 13

(f) Controlling for average income, population density and percent of high school graduates, are there differences between regions in average crime rate? This is one test.

i. State the null hypothesis in terms of β values from R's model.

ii. Fill in the table.

Test Statistic (t or F) Value: A number	Reject H_0 at $\alpha = 0.05$? (Yes or No)

iii. Controlling for the other variables, is crime rate related to region? Answer Yes or No.

(g) We are sure that the regions do not all have the same crime rate once we control for the other variables. But which ones are different from which other ones? We need to test pairs of mean crime rates agains each other: North east versus North Central, Northeast versus South, and so on. There are six of these *pairwise comparisons*, and since we are carrying out a well-defined set of multiple tests, we decide to protect them jointly against Type I error with a Bonferroni correction.

i. With the Bonferroni correction, we will compare the p-values on the printout not to 0.05, but to another number. Write the number in the space below.

ii. When I do this, I see only one difference that is *not* significant with the Bonferroni correction. To express this in plain language, we might put a footnote to the table of adjusted means that says "_____ and _____ are in a statistical tie." Fill in the blanks.