NAME (PRINT): ____

Last/Surname First /Given Name

STUDENT #: ______________________ **SIGNATURE:** ______________________

UNIVERSITY OF TORONTO MISSISSAUGA
DECEMBER 2016 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, once this exam has begun, you ***CANNOT*** *re-write it.*

Qn. #	Value	Score
1	10	
2	8	
3	8	
4	10	
5	8	
6	12	
7	15	
8	4	
9	10	
10	15	
Total = 100 Points		

Seat Position

The questions on this exam refer to the general linear model on the formula sheet, with X an $n \times (k+1)$ matrix of fixed, observable constants. Unless otherwise indicated, the columns of X matrix are linearly independent, and $n > k+1$. That is, X is *not a square matrix*, and *does not have an inverse.*

10 points

1. In the `marks` data, the independent variables are quiz average, average on the computer assignments, and score on the midterm test. The dependent variable is score on the final examination. We have complete data for the 58 students who took the final exam. The regression model includes an intercept, and there are no interactions. For each of the matrices below, give the number of rows and the number of columns. The answers are numbers.

Matrix	Number of Rows	Number of Columns
$\mathbf{y}$		
X		
$\boldsymbol{\beta}$		
$\boldsymbol{\epsilon}$		
$(X'X)^{-1}$		
$X\boldsymbol{\beta}$		
H		
$\widehat{\mathbf{y}}$		
$\mathbf{b}$		
$(X'X)^{-1}X'\mathbf{y}$		

8 points

2. Prove that if the columns of X are linearly dependent, the least-squares estimate $\mathbf{b}$ does not exist. You have more room than you need.

Continued on page 4

8 points

3. Let the $p \times 1$ random vector $\mathbf{w}$ have expected value $\boldsymbol{\mu}$ and variance-covariance matrix Σ. Let A be an $m \times p$ matrix of constants, and let B be an $n \times p$ matrix of constants. Find a nice simple expression for the $m \times n$ matrix of covariances $cov(A\mathbf{w}, B\mathbf{w})$. Show your work. You have more room than you need. **Circle your final answer**.

Continued on page 5

10 points

4. In the general linear regression model, assume only that $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $cov(\boldsymbol{\epsilon}) = \sigma^2 I_n$, not multivariate normality.

   (a) Prove that $\mathbf{e} = (I - H)\mathbf{y}$. You're proving a fact on the formula sheet, so do not use it directly. Also, don't use $\widehat{\mathbf{y}} = H\mathbf{y}$. You can use the formula for $\mathbf{b}$; that's a good place to start.

   (b) Calculate $E(\mathbf{e})$. Show your work and simplify.

   (c) Calculate $cov(\mathbf{e})$. Show your work and simplify. It's okay to use properties of $(I - H)$ that are not on the formula sheet, if you know them. **Circle your final answer**.

Continued on page 6

8 points

5. Let $\mathbf{w}_1 \sim N_p(\boldsymbol{\mu}_1, \Sigma_1)$ and $\mathbf{w}_2 \sim N_p(\boldsymbol{\mu}_2, \Sigma_2)$ be independent multivariate normal random vectors (*not* scalar random variables). Using without proof the fact that the moment-generating function of a sum of independent random vectors is the product of moment-generating functions, find the distribution of $\mathbf{w} = \mathbf{w}_1 + \mathbf{w}_2$. Show your work. Don't just calculate mean and covariance; use moment-generating functions. **Circle your final answer**.

Continued on page 7

12 points

6. Let $\mathbf{w} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is positive definite. Show that $(\mathbf{w} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{w} - \boldsymbol{\mu}) \sim \chi^2(p)$. You may use anything on the formula sheet except the fact you are proving.

Continued on page 8

15 points

7. Assuming the multivariate normality of $\boldsymbol{\epsilon}$ as on the formula sheet,

   (a) What is the distribution of the scalar (1×1) random variable $\boldsymbol{\ell}'\mathbf{b}$? Show a little work. **Circle your answer**.

   (b) Standardize the random variable from the last part to obtain a standard normal. Write the formula for Z.

   (c) Divide Z by a well-chosen $\sqrt{w/\nu}$, and simplify. The result is a formula for t. **Circle your final answer**.

Continued on page 9

(d) Continuing with Question 7, how do you know numerator and denominator are independent?

(e) What is the distribution of the random variable in Question 7c? Don't forget the degrees of freedom.

(f) Suppose you want to test $H_0 : \boldsymbol{\ell}'\boldsymbol{\beta} = \gamma$. Using your work on this question, give a formula for the test statistic.

4 points 8. If the normal linear model is correct and you look at a scatterplot of the residuals against the fitted $\widehat{y}$ values, you should see a random cloud of points. Why? (This question is *not* asking you to calculate a sample correlation.)

Continued on page 10

10 points 9. In an extended version of the SAT data, the dependent variable is first-year university Grade Point Average (GPA). The independent variables are

x_1 = Verbal SAT score	x_2 = Math SAT score
x_3 = High school Grade Point Average	x_4 = Mother's education, in years
x_5 = Father's education, in years	x_6 = Total family income,

and also Location of the family home: City, Suburbs or Country.

(a) First, write the regression equation. It is up to you which dummy variable variable scheme you use, as long as the regression planes are parallel. You will specify how your dummy variables are defined in the next part.

(b) Make a table with one row for each location of the family home, showing how your dummy variables are defined. Make one more column showing $E(y|\mathbf{x})$ for each location.

Continued on page 11

(c) Continuing Question 9, for each of the following questions give the null hypothesis in the form of a statement about the β values.

i. Correcting for all other variables, is location of the family home related to first-year GPA?

ii. Controlling for all other variables, is either Verbal SAT score or Math SAT score (or both) related to GPA?

iii. When you allow for all the other variables, is family income a useful predictor of GPA?

iv. Controlling for all other variables, does expected GPA change faster as a function of Verbal SAT, or does it change faster as a function of Math SAT?

v. Once you correct for the two SAT scores and High School marks, do any of the family variables (including parents' education) matter?

vi. Correcting for all other variables, does expected GPA change faster as a function of Mother's education, or does it change faster as a function of father's education?

vii. Holding all the other variables constant at fixed values, is Math SAT related to first-year university GPA?

viii. Once you allow for location of the family home, do any of the other predictors matter?

Continued on page 12

15 points

10. This last part of the exam is based on the `hospital` data. The questions are mixed in with my R printout.

```
> hospital = read.table("http://www.utstat.toronto.edu/~brunner/data/legal/openSENIC.data.t
> head(hospital); attach(hospital)
     region mdschl census nbeds nurses lngstay  age xratio culratio infpercent
1 Northeast     No    237   298    115   12.01 52.8   96.9     10.8        4.8
2 Northeast    Yes    144   184    151   10.05 52.0   87.5     36.7        4.5
3 Northeast     No    127   165    158    9.36 54.1   90.6     18.3        4.8
4 Northeast    Yes    240   270    198    9.78 52.3   95.9     17.6        5.0
5      West     No     51    76     79    6.70 48.6   80.8     13.0        4.5
6     South     No     59    95     56    8.93 56.0   72.5      6.2        2.0
> full1 = lm(infpercent ~ region + mdschl + census + nbeds + nurses + lngstay +
+ age + xratio + culratio)
> summary(full1)

Call:
lm(formula = infpercent ~ region + mdschl + census + nbeds +
    nurses + lngstay + age + xratio + culratio)

Residuals:
    Min      1Q  Median      3Q     Max
-1.8685 -0.4972  0.0319  0.4433  1.9928

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.622529   1.239778   0.502  0.61683
regionNortheast -0.400251   0.269460  -1.485  0.14102
regionSouth     -0.266435   0.237064  -1.124  0.26411
regionWest       0.760288   0.297932   2.552  0.01244 *
mdschlYes       -0.772632   0.336651  -2.295  0.02411 *
census           0.007849   0.003611   2.174  0.03241 *
nbeds           -0.007001   0.002901  -2.414  0.01787 *
nurses           0.004579   0.001869   2.450  0.01628 *
lngstay          0.231925   0.071235   3.256  0.00161 **
age             -0.006134   0.022681  -0.270  0.78745
xratio           0.007600   0.005428   1.400  0.16496
culratio         0.053184   0.010607   5.014 2.74e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8873 on 88 degrees of freedom
Multiple R-squared:  0.605,Adjusted R-squared:  0.5556
F-statistic: 12.25 on 11 and 88 DF,  p-value: 1.238e-13

> # Get critical value t_alpha/2
> crit1 = qt(0.975,88); crit1
[1] 1.98729
```

Continued on page 13

(a) Recall that `census` is the average number of patients in the hospital during the study period. Controlling for all other variables, we want to know whether number of patients is related to infection risk.

i.

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

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

(b) Allowing for all other variables, we want to know whether average age of the patients in the hospital is related to infection risk.

i.

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

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

(c) Correcting for all other variables, we want to know whether number of nurses in the hospital is related to infection risk.

i.

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

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

(d) Holding all other characteristics of the hospital to fixed values, calculate a 95% confidence interval for the difference in infection risk between hospitals with and without a medical school affiliation. The answer is a pair of numbers. Show a little work and **circle your answer**.

Continued on page 14

```
> # Some custom tests

> # Test 1
> redmodel1 = lm(infpercent ~ region + mdschl + lngstay + age + xratio + culratio)
> anova(redmodel1,full1)
Analysis of Variance Table

Model 1: infpercent ~ region + mdschl + lngstay + age + xratio + culratio
Model 2: infpercent ~ region + mdschl + census + nbeds + nurses + lngstay +
    age + xratio + culratio
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     91 83.192
2     88 69.277  3    13.915 5.8918 0.001033 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> # Test 2
> redmodel2 = lm(infpercent ~ census + nbeds + nurses)
> anova(redmodel2,full1)
Analysis of Variance Table

Model 1: infpercent ~ census + nbeds + nurses
Model 2: infpercent ~ region + mdschl + census + nbeds + nurses + lngstay +
    age + xratio + culratio
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     96 137.182
2     88  69.277  8    67.905 10.782 1.855e-10 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

> # Test 3
> source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/ftest.txt")
> C3 = rbind(c(0,1,0,0,0,0,0,0,0,0,0,0,0),
+            c(0,0,1,0,0,0,0,0,0,0,0,0,0),
+            c(0,0,0,1,0,0,0,0,0,0,0,0,0))
> ftest(full1,C3)
          F         df1         df2     p-value
 4.75132484  3.00000000 88.00000000  0.00405991

> # Test 4
> C4 = rbind(c(0,0,0,0,1,0,0,0,0,0,0,0,0),
+            c(0,0,0,0,0,1,0,0,0,0,0,0,0),
+            c(0,0,0,0,0,0,1,0,0,0,0,0,0),
+            c(0,0,0,0,0,0,0,1,0,0,0,0,0),
+            c(0,0,0,0,0,0,0,0,1,0,0,0,0),
+            c(0,0,0,0,0,0,0,0,0,1,0,0,0),
+            c(0,0,0,0,0,0,0,0,0,0,1,0,0),
+            c(0,0,0,0,0,0,0,0,0,0,0,0,1))
> ftest(full1,C4)
           F          df1          df2      p-value
1.461432e+01 8.000000e+00 8.800000e+01 2.232659e-13
```

Continued on page 15

(e) Controlling for all other variables, we want to know whether infection risk varies by region of the country.

i.

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

ii. Does infection risk appear to vary by region of the country? Just answer Yes or No (no need for directional conclusions).

(f) One can think of number of patients, number of beds and number of nurses as all basically reflecting the size of the hospital, so it makes sense to test them simultaneously. Controlling for all other variables, we want to know whether size of hospital is related to infection risk.

i.

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

ii. Does size of hospital appear to be related to infection risk? Just answer Yes or No (no need for directional conclusions).

(g) Now we will treat the Studentized deleted as t-statistics to test for outliers.

```
> # Studentized deleted residuals
> n = length(infpercent)
> crit2 = qt(1-0.05/(2*n),88); crit2
[1] 3.614922
> estar1 = rstudent(full1); summary(estar1)
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.678000 -0.623300  0.037230 -0.009751  0.538600  2.430000
```

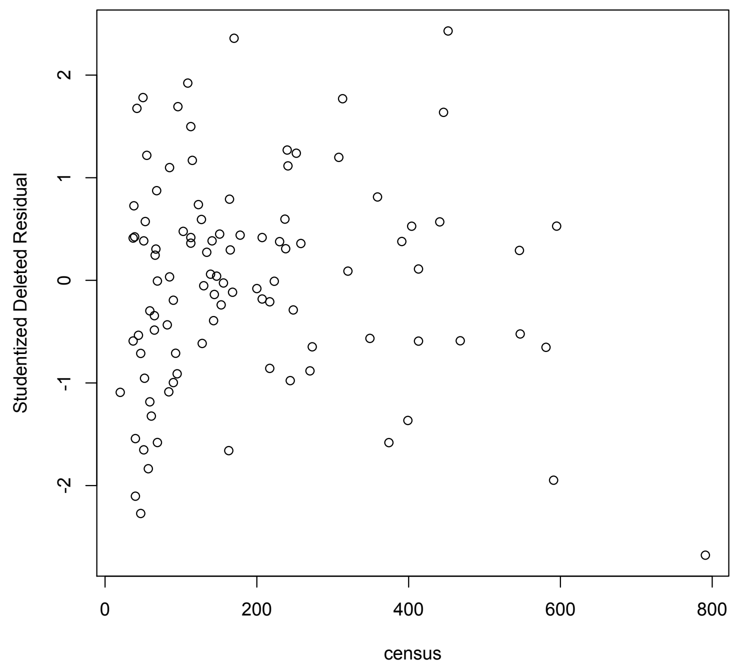
i. Is there evidence of outliers? Answer Yes or No and briefly explain.

ii. Would there have been apparent evidence of outliers without the Bonferroni correction? Answer Yes or No and briefly explain. For full marks, give the critical value you are using.

Continued on page 16

(h) I like to plot the Studentized deleted residuals instead of the raw residuals. There are many potential plots; here is an interesting one.

```
> # Look for decreasing variance
> plot(census,estar1,ylab = 'Studentized Deleted Residual')
```



i. Why was I looking for variance that decreased with the number of patients in the hospital?

ii. If there had been decreasing variance, what might I have done about it?

iii. Instead of non-constant variance, I see a curve. Assuming you can (sort of) see what I am seeing, is this curve concave up or is it concave down?

Continued on page 17

(i) This calls for polynomial regression.

```
> # Add a quadratic term.
> csquared = census^2
> full2 = update(full1, ~ . + csquared) # Quick way to add csquared to the model.
> # Dot means everything that was in there before.
> summary(full2)

Call:
lm(formula = infpercent ~ region + mdschl + census + nbeds +
    nurses + lngstay + age + xratio + culratio + csquared)

Residuals:
     Min       1Q   Median       3Q      Max
-1.73562 -0.53234 -0.00737  0.47875  1.85345

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)     -3.215e-01  1.203e+00  -0.267 0.789846
regionNortheast -4.003e-01  2.544e-01  -1.573 0.119273
regionSouth     -3.138e-01  2.243e-01  -1.399 0.165289
regionWest       9.553e-01  2.870e-01   3.328 0.001283 **
mdschlYes       -6.927e-01  3.187e-01  -2.173 0.032468 *
census           1.601e-02  4.162e-03   3.848 0.000227 ***
nbeds           -7.702e-03  2.747e-03  -2.804 0.006220 **
nurses           2.929e-03  1.830e-03   1.601 0.113074
lngstay          2.085e-01  6.761e-02   3.083 0.002745 **
age              6.115e-03  2.171e-02   0.282 0.778880
xratio           6.996e-03  5.128e-03   1.364 0.176037
culratio         5.544e-02  1.004e-02   5.523 3.41e-07 ***
csquared        -9.804e-06  2.866e-06  -3.421 0.000951 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.8378 on 87 degrees of freedom
Multiple R-squared:  0.6518,Adjusted R-squared:  0.6038
F-statistic: 13.57 on 12 and 87 DF,  p-value: 2.648e-15
```

Is there evidence that the quadratic term is useful?

i.

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

ii. Does the quadratic term help? Answer Yes, No, or No Conclusion.

The analysis continues. Next I looked at influence diagnostics, and indeed that big hospital had high leverage; it was potentially an influential observation. So I re-did the analysis omitting it, and .... But it's time to go. Have a good holiday!

Total Marks = 100 points