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

STUDENT #: ______________________ **SIGNATURE:** ____________________

UNIVERSITY OF TORONTO MISSISSAUGA
DECEMBER 2013 FINAL EXAMINATION
STA302H5F
Regression Analysis
Jerry Brunner
Duration - 3 hours
Aids: Calculator Model(s): Any calculator without wireless connectivity 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, 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	12	
2	8	
3	8	
4	10	
5	15	
6	14	
7	5	
8	5	
9	23	

Total = 100 Points

12 points

1. Let $Y_1, \ldots, Y_n$ be independent *scalar* (not matrix) random variables with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ for $i = 1, \ldots, n$.

   (a) Let $c_1, \ldots, c_n$ be constants and define the linear combination L by $L = \sum_{i=1}^n c_i Y_i$. What condition on the c_i values makes L an unbiased estimator of μ? Show your work.

   (b) What is the variance of the linear combination L? Show a little work.

   (c) Show that if the linear combination L is unbiased for μ, the constants c_i that make variance of L as small as possible are $c_i = \frac{1}{n}$ for $i = 1, \ldots, n$. That is, the sample mean is the Best Linear Unbiased Estimator (BLUE).

continued on page 3

8 points 2. Let the $p \times 1$ random vector $\mathbf{Y}$ have mean $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and let $\mathbf{c}$ be a $p \times 1$ vector of constants. Choose one of the statements below and prove it is true.

$$cov(\mathbf{Y}+\mathbf{c}) = \boldsymbol{\Sigma} \quad cov(\mathbf{Y}+\mathbf{c}) = \boldsymbol{\Sigma}+\mathbf{cc}' \quad cov(\mathbf{Y}+\mathbf{c}) = \mathbf{c}\boldsymbol{\Sigma}\mathbf{c}' \quad cov(\mathbf{Y}+\mathbf{c}) = \mathbf{0}$$

8 points 3. Show that if $\mathbf{Y} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{AY} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$. You are proving something on the formula sheet, so you may use anything on the formula sheet *except* what you are proving.

continued on page 4

10 points

4. For the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, prove that if the columns of $\mathbf{X}$ are linearly dependent, the least squares estimator $\widehat{\boldsymbol{\beta}}$ does not exist. You have more room than you need.

continued on page 5

15 points

5. Show $SST = SSR + SSE$. You are proving something on the formula sheet, so you may use anything on the formula sheet *except* what you are proving. Assume that the regression model has an intercept, so that $\sum_{i=1}^{n} \widehat{Y}_i = \sum_{i=1}^{n} Y_i$.

continued on page 6

continued on page 7

14 points

6. For the linear regression model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ with normal errors,

   (a) What is the distribution of $\mathbf{C}\widehat{\boldsymbol{\beta}}$? Note $\mathbf{C}$ is $q \times (k+1)$. Just write down the answer.

   (b) If $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ is true, what is the distribution of $(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{t})'(\sigma^2\mathbf{C}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{C}')^{-1}(\mathbf{C}\widehat{\boldsymbol{\beta}} - \mathbf{t})$? Just write down the answer.

   (c) Using what you have just shown, complete the proof that the F statistic on the formula sheet does indeed have an F distribution when the null hypothesis is true. You may use anything from the formula sheet except what you are proving.

continued on page 8

continued on page 9

5 points

7. Assume that the independent variables in a regression model are actually random variables rather than fixed constants. In this case, the usual fixed-x regression model is a *conditional* model, in which all the usual results hold conditionally upon $\mathbf{X} = \mathbf{x}$. Using the fact that $E(\widehat{\boldsymbol{\beta}}|\mathbf{X}) = \boldsymbol{\beta}$, show that $E(\widehat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, so that the usual estimator is unbiased even when $\mathbf{X}$ is random.

5 points

8. As in the prededing question, assume that the usual linear regression model is a conditional one. Let $\mathbf{X}_i$ denote the $k \times 1$ random vector of independent variable values for observation i. The conditional model with normal error terms says that the conditional distribution of ϵ_i given $\mathbf{X}_i = \mathbf{x}_i$ is $N(0, \sigma^2)$. Show how this implies that $\mathbf{X}_i$ and ϵ_i are independent. For convenience, you may assume that $\mathbf{X}_i$ has a density.

continued on page 10

23 points

9. This question is based on an analysis of the `birthweight` data with R. Assume the usual $\alpha = 0.05$ significance level. First comes just the input, then the questions, and finally a complete listing of the input and output.

```
library(MASS); attach(birthwt); head(birthwt)
n = length(age); n
mean(age)
# For race, 1=White, 2=Black, 3=Other
r2=numeric(n); r2[race==2]=1
r3=numeric(n); r3[race==3]=1
babysweight = bwt
momsweight = lwt-mean(lwt) # Mom's weight is centered
r2mw = r2*momsweight; r3mw = r3*momsweight

modelA = lm(babysweight ~ momsweight)
modelB = lm(babysweight ~ momsweight + r2 + r3)
modelC = lm(babysweight ~ momsweight + r2 + r3 + r2mw + r3mw)

anova(modelA,modelB)
anova(modelB,modelC)

summary(modelC)

# Now look at Model B
summary(modelB)
# Some additional t-tests on Model B
V = vcov(modelB); betahat=modelB$coefficients
dfe = modelB$df.residual # dfe = n-k-1

a = rbind(1,0,0,0)
T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
p = 2*(1-pt(abs(T),dfe)); T; p

a = rbind(1,0,-1/2,-1/2)
T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
p = 2*(1-pt(abs(T),dfe)); T; p

a = rbind(0,1,-1,0)
T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
p = 2*(1-pt(abs(T),dfe)); T; p

a = rbind(0,1,0,-1)
T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
p = 2*(1-pt(abs(T),dfe)); T; p

a = rbind(0,0,1,-1)
T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
p = 2*(1-pt(abs(T),dfe)); T; p

a = rbind(0,0,1,1)
T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
p = 2*(1-pt(abs(T),dfe)); T; p
```

continued on page 11

(a) Is there evidence that the slope of the regression line relating mother's weight to baby's weight is different for Black mothers and White mothers? Give two numbers and the word "Yes" or "No."

Test Statistic (F or t)	p-value	Answer Yes or No

(b) Is there evidence that race differences in baby's weight depend on the weight of the mother? Give two numbers and the word "Yes" or "No."

Test Statistic (F or t)	p-value	Answer Yes or No

For the rest of the questions, please treat Model B as the full model.

(c) Allowing for mother's race, is there evidence that baby's weight is related to mother's weight? Give two numbers and the word "Yes" or "No."

Test Statistic (F or t)	p-value	Answer Yes or No

(d) If the answer to the last question was "Yes," describe the results in plain, non-statistical language.

(e) Give an estimate of expected baby's weight for White mothers of average (sample mean) weight. The answer is a number.

(f) Give an estimate of expected baby's weight for B;ack mothers of average (sample mean) weight. The answer is a number.

(g) Give an estimate of expected baby's weight for Other mothers of average (sample mean) weight. The answer is a number.

(h) Allowing for mother's weight, is there evidence that baby's weight is related to mother's race? Give two numbers and the word "Yes" or "No."

Test Statistic (F or t)	p-value	Answer Yes or No

continued on page 12

(i) In the table below, fill in the p-values for the pairwise comparisons of racial groups, comparing expected baby's weight controlling for mother's weight.

	White	Black	Other
White	x		
Black	x	x	
Other	x	x	x

(j) Describe the results of the pairwise comparisons in plain, non-statistical language. You may begin your answer with "Allowing for mother's weight, . . . "

That's the end of the exam questions. The rest of the exam paper consists of R input and output.

```
> library(MASS); attach(birthwt); head(birthwt)
   low age lwt race smoke ptl ht ui ftv  bwt
85   0  19 182    2     0   0  0  1   0 2523
86   0  33 155    3     0   0  0  0   3 2551
87   0  20 105    1     1   0  0  0   1 2557
88   0  21 108    1     1   0  0  1   2 2594
89   0  18 107    1     1   0  0  1   0 2600
91   0  21 124    3     0   0  0  0   0 2622
> n = length(age); n
[1] 189
> mean(age)
[1] 23.2381
> # For race, 1=White, 2=Black, 3=Other
> r2=numeric(n); r2[race==2]=1
> r3=numeric(n); r3[race==3]=1
> babysweight = bwt
> momsweight = lwt-mean(lwt) # Mom's weight is centered
> r2mw = r2*momsweight; r3mw = r3*momsweight
>
> modelA = lm(babysweight ~ momsweight)
> modelB = lm(babysweight ~ momsweight + r2 + r3)
> modelC = lm(babysweight ~ momsweight + r2 + r3 + r2mw + r3mw)
>
> anova(modelA,modelB)
Analysis of Variance Table

Model 1: babysweight ~ momsweight
Model 2: babysweight ~ momsweight + r2 + r3
  Res.Df      RSS Df Sum of Sq      F   Pr(>F)
1    187 96521017
2    185 91444408  2   5076610 5.1352 0.006753 **
```

continued on page 13

```
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> anova(modelB,modelC)
Analysis of Variance Table

Model 1: babysweight ~ momsweight + r2 + r3
Model 2: babysweight ~ momsweight + r2 + r3 + r2mw + r3mw
  Res.Df      RSS Df Sum of Sq     F Pr(>F)
1    185 91444408
2    183 91150564  2    293844 0.295 0.7449
>
> summary(modelC)

Call:
lm(formula = babysweight ~ momsweight + r2 + r3 + r2mw + r3mw)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.19  -450.29    55.55   493.54  1932.54

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3091.532     72.246  42.792   <2e-16 ***
momsweight     5.000      2.489   2.009   0.0460 *
r2          -413.101    167.446  -2.467   0.0145 *
r3          -226.272    117.479  -1.926   0.0556 .
r2mw          -2.572      4.344  -0.592   0.5545
r3mw           1.120      4.260   0.263   0.7929
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 705.8 on 183 degrees of freedom
Multiple R-squared: 0.08822,Adjusted R-squared: 0.06331
F-statistic: 3.541 on 5 and 183 DF,  p-value: 0.004429

>
> # Now look at Model B
> summary(modelB)

Call:
lm(formula = babysweight ~ momsweight + r2 + r3)

Residuals:
     Min       1Q   Median       3Q      Max
-2096.21  -419.56    41.39   478.57  1929.49

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 3092.285     71.863  43.031  < 2e-16 ***
momsweight     4.663      1.750   2.665  0.00839 **
r2          -451.838    157.566  -2.868  0.00462 **
r3          -241.301    113.887  -2.119  0.03544 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

continued on page 14

```
Residual standard error: 703.1 on 185 degrees of freedom
Multiple R-squared: 0.08528,Adjusted R-squared: 0.07045
F-statistic: 5.749 on 3 and 185 DF,  p-value: 0.000881

> # Some additional t-tests on Model B
> V = vcov(modelB); betahat=modelB$coefficients
> dfe = modelB$df.residual # dfe = n-k-1
>
> a = rbind(1,0,0,0)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> p = 2*(1-pt(abs(T),dfe)); T; p
[1] 43.03055
[1] 0
>
> a = rbind(1,0,-1/2,-1/2)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> p = 2*(1-pt(abs(T),dfe)); T; p
[1] 20.85298
[1] 0
>
> a = rbind(0,1,-1,0)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> p = 2*(1-pt(abs(T),dfe)); T; p
[1] 2.891767
[1] 0.004289915
>
> a = rbind(0,1,0,-1)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> p = 2*(1-pt(abs(T),dfe)); T; p
[1] 2.165632
[1] 0.03161921
>
> a = rbind(0,0,1,-1)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> p = 2*(1-pt(abs(T),dfe)); T; p
[1] -1.245203
[1] 0.2146321
>
> a = rbind(0,0,1,1)
> T = as.numeric( t(a)%*%betahat/sqrt(t(a)%*%V%*%a) )
> p = 2*(1-pt(abs(T),dfe)); T; p
[1] -3.196971
[1] 0.001633545
>
```

Total Marks = 100 points