**NAME (PRINT):** ______________________________________________
Last/Surname First /Given Name

**STUDENT #:** ____________________ **SIGNATURE:** ____________________

**UNIVERSITY OF TORONTO MISSISSAUGA**
**DECEMBER 2014 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, 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 | 5 | |
| 2 | 5 | |
| 3 | 6 | |
| 4 | 3 | |
| 5 | 8 | |
| 6 | 15 | |
| 7 | 3 | |
| 8 | 10 | |
| 9 | 17 | |
| 10 | 13 | |
| 11 | 15 | |

Total = 100 Points

Unless otherwise indicated, the questions in 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. The columns of the $\mathbf{X}$ matrix are linearly independent.

*5 points*

1. Show that if the columns of $\mathbf{X}$ are linearly independent, the matrix $\mathbf{X}'\mathbf{X}$ is positive definite. It may help to notice that in the definitions on the formula sheet, $\mathbf{v}$ is $(k+1) \times 1$, so that $\mathbf{Z} = \mathbf{X}\mathbf{v}$ is $n \times 1$.

*continued on page 3*

*5 points* 2. Let the random vector $\mathbf{Y}$ have expected value $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$, and let $\mathbf{A}$ and $\mathbf{B}$ be matrices of constants. Prove that $C(\mathbf{AY}, \mathbf{BY}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{B}'$.

*6 points* 3. Now for linear regression, locate the matrices $\mathbf{A}$ and $\mathbf{B}$ on the formula sheet and use the last result to show $C(\widehat{\boldsymbol{\epsilon}}, \widehat{\boldsymbol{\beta}}) = \mathbf{0}$.

*3 points* 4. Why does $C(\widehat{\boldsymbol{\epsilon}}, \widehat{\boldsymbol{\beta}}) = \mathbf{0}$ establish that $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\epsilon}}$ are independent?

*continued on page 4*

*8 points*

5. For the usual linear regression model,

   (a) Calculate $\mathbf{X}'\widehat{\boldsymbol{\epsilon}}$ and simplify.

   (b) Without going all the way back to the normal equations, why does the last result show that if a regression model has an intercept, then the residuals must add up to zero?

   (c) Your answer to Question 5a also shows that geometrically, the vector of residuals is at right angles to every column vector of the $\mathbf{X}$ matrix. So, it should be at right angles to any linear combination of those column vectors. A general way to represent such a linear combination is $\mathbf{Xb}$, where $\mathbf{b}$ is a $(k+1) \times 1$ vector that could be random or constant. Show $(\mathbf{Xb})'\widehat{\boldsymbol{\epsilon}} = 0$.

Don't forget this result. It will make your job easier on the next question.

*continued on page 5*

*15 points* 6. Show that $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \widehat{\boldsymbol{\epsilon}}\,'\,\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$.

*3 points* 7. Why does the result above tell you that $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$ is minimal when $\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}$?

*continued on page 6*

*10 points*

8. In the general linear regression model, let $cov(\boldsymbol{\epsilon}) = \sigma^2\mathbf{V}$, where $\mathbf{V}$ is a *known* symmetric and positive definite matrix. As usual, $\sigma^2$ is an unknown constant. Suppose we try to estimate $\boldsymbol{\beta}$ with $\widehat{\boldsymbol{\beta}}_w = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}$.

   (a) Is $\widehat{\boldsymbol{\beta}}_w$ an unbiased estimator of $\boldsymbol{\beta}$? Answer Yes or No and prove your answer.

   (b) If $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{V})$, what is the distribution of $\widehat{\boldsymbol{\beta}}_w$? Use the formula sheet, show your work and simplify.

*continued on page 7*

*17 points*

9. Question 6 established that $(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) = \widehat{\boldsymbol{\epsilon}}'\,\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$.

   (a) What is the distribution of $\frac{1}{\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})$? Prove it, citing any facts from the formula sheet as you use them. Include the degrees of freedom in your answer.

   (b) What is the distribution of $\frac{1}{\sigma^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$? Prove it, citing any facts from the formula sheet as you use them. Include the degrees of freedom in your answer.

*continued on page 8*

(c) How do you know $\frac{1}{\sigma^2}\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}$ and $\frac{1}{\sigma^2}(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ are independent?

(d) Finally (this was the goal), show $\frac{1}{\sigma^2}\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} \sim \chi^2(n-k-1)$.

*continued on page 9*

*13 points* 10. In a study comparing the effectiveness of different weight loss diets, volunteers were randomly assigned to one of two diets ($A$ or $B$) or put on a waiting list and advised to lose weight on their own. Participants were weighed before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable is weight *loss.* The explanatory variables are age and treatment group.

(a) Write the regression equation. Your model should have *parallel regression lines* for the different treatment groups. Please use $x$ for age. You don't have to say how your dummy variables are defined. You'll do that in the next part.

(b) Make a table with three rows, showing how you would set up indicator dummy variables for treatment group. Give $E(Y|x)$ in the last column.

(c) In terms of $\beta$ values, what null hypothesis would you test to find out whether, allowing for age, the three diets (including Wait List) differ in their effectiveness?

(d) In terms of $\beta$ values, what null hypothesis would you test to find out whether, allowing for age, diets $A$ and $B$ differ in their effectiveness?

(e) In terms of $\beta$ values, what null hypothesis would you test to find out whether 25 year old participants who spend six months on the Wait list "diet" have any average tendency to gain or lose weight? Remember, $Y$ is weight loss.

*continued on page 10*

11. (*15 points*) This question is based on a study of low birth weight in babies. The dependent variable is `low`, which equals 1 if the baby is less than 2.5 kg (low birth weight), and zero otherwise. The independent variables are mother’s weight (`lwt`), an indicator variable for smoking (`smoke`), and mother’s race.

```
> # Birth weight: MASS package must be loaded
> attach(birthwt)
> # low is low birth weight, lwt is mother’s weight, smoke is indicator for smoking
> race = factor(race, labels = c("White", "Black", "Other"))
> contrasts(race)
      Black Other
White     0     0
Black     1     0
Other     0     1
> baby = glm(low ~ lwt+smoke+race, family=binomial); summary(baby)

Call:
glm(formula = low ~ lwt + smoke + race, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5278  -0.9053  -0.5863   1.2878   2.0364

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.10922    0.88211  -0.124  0.90146
lwt         -0.01326    0.00631  -2.101  0.03562 *
smoke        1.06001    0.37832   2.802  0.00508 **
raceBlack    1.29009    0.51087   2.525  0.01156 *
raceOther    0.97052    0.41224   2.354  0.01856 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

    Null deviance: 234.67  on 188  degrees of freedom
Residual deviance: 215.01  on 184  degrees of freedom
AIC: 225.01

Number of Fisher Scoring iterations: 4

> mod1 = glm(low ~ lwt+smoke, family=binomial)
> anova(mod1,baby,test=’Chisq’)
Analysis of Deviance Table

Model 1: low ~ lwt + smoke
Model 2: low ~ lwt + smoke + race
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       186     224.34
2       184     215.01  2    9.326 0.009438 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
> mod2 = glm(low ~ 1 , family=binomial)
> anova(mod2,baby,test=’Chisq’)
Analysis of Deviance Table

Model 1: low ~ 1
Model 2: low ~ lwt + smoke + race
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       188     234.67
2       184     215.01  4   19.657 0.0005835 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

*continued on page 11*

(a) Controlling for race and weight, the estimated odds of a low birth weight baby are _________ times as great for a mother who smokes. The answer is a number. Write your answer in the space below. **Circle your answer**.

(b) Controlling for smoking and weight, the estimated odds of a low birth weight baby are _________ times as great for a Black mother as for a White mother. The answer is a number. Write your answer in the space below. **Circle your answer**.

(c) We want to know whether, controlling for smoking and weight, race is related to the chances of having a low birth weight baby.

   i. Give the value of test statistic (not the $p$-value). The answer is a number.

   ii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes of No.

   iii. Controlling for smoking and weight, is there evidence that race is related to the chances of having a low birth weight baby? Answer Yes or No.

(d) We want to know whether, controlling for smoking and weight, White and Other mothers differ in their chances of having a low birth weight baby.

   i. Give the value of test statistic (not the $p$-value). The answer is a number.

   ii. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes of No.

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

(e) Estimate the probability of a low birth weight bably for a 130-pound, non-smoking White mother. The answer is a number. **Circle your answer**.

Total Marks = 100 points