UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2017 Examinations
STA442H1F/2101H1F
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours

Aids Allowed: Any calculator without wireless capability.
Formula sheet supplied.

Last/Family Name (Print): ______________________

First/Given Name (Print): ______________________

Student Number: ______________________

Signature: ______________________

Qn. #	Value	Score
1	12	
2	10	
3	15	
4	15	
5	8	
6	8	
7	15	
8	17	
Total = 100 Points		

1. (12 *points*) In an experiment with three experimental treatments, n_1 sampling units receive treatment one, n_2 sampling units receive treatment two, and n_3 sampling units receive treatment three. The total sample size is $n = n_1 + n_2 + n_3$. Expected responses to the treatments are μ_1, μ_2 and μ_3, and the corresponding sample means are $\overline{y}_1$, $\overline{y}_2$ and $\overline{y}_3$. To set this up as a regression model, let $x_{i,1}$ be an indicator dummy variable for treatment one, let $x_{i,2}$ be an indicator dummy variable for treatment two, and let $x_{i,3}$ be an indicator dummy variable for treatment three.

   (a) First consider a model with an intercept and two dummy variables. In scalar form, the model equation is $y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \epsilon_i$. For the matrix version $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, what is the $\mathbf{X}^\top\mathbf{X}$ matrix? It is a specific 3×3 matrix, and each element of the matrix is either a zero or a function of n_1, n_2 and n_3.

   (b) Now consider a cell means model with all three dummy variables and no intercept. In scalar form, the model equation is $y_i = \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,3} + \epsilon_i$.

      i. What is the $\mathbf{X}^\top\mathbf{X}$ matrix? Again, each element of the matrix is either a zero or a function of n_1, n_2 and n_3.

      ii. What is $(\mathbf{X}^\top\mathbf{X})^{-1}$?

      iii. (What is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}$? It is a specific matrix containing symbols that have been previously mentioned in this question.

continued on page 3

2. (10 *points*) The formula sheet has a formula for the Wald test statistic W_n, which may be used to test the null hypothesis $H_0 : \mathbf{L}\boldsymbol{\theta} = \mathbf{h}$. Let $\mathbf{A}$ be an $r \times r$ matrix with an inverse. Clearly, $H_0 : \mathbf{A}\mathbf{L}\boldsymbol{\theta} = \mathbf{A}\mathbf{h}$ is true if and only if $\mathbf{L}\boldsymbol{\theta} = \mathbf{h}$. How does this way of re-expressing H_0 affect W_n? Show your work. *Make sure you answer the question.*

continued on page 4

3. (*15 points*) Most standard tests are for linear hypotheses, but non-linear hypotheses can be very useful too. For example in a regression, if $H_0 : \beta_1\beta_2 = 0$ is rejected, it means that both regression coefficients are non-zero. Here is a simple example from the `Math` data. The questions come after the printout.

```
> # Testing a non-linear hypothesis suggested by Vishak Patel
>
> mathengl = math[,5:7]; mathengl = na.omit(mathengl)
> head(mathengl); attach(mathengl)
  hscalc hsengl ucalc
1     65     80    39
2     54     75    57
3     77     70    62
4     80     67    76
5     87     80    86
6     53     75    60
>
> mod = lm(ucalc ~ hscalc + hsengl); summary(mod)

Call:
lm(formula = ucalc ~ hscalc + hsengl)

Residuals:
    Min      1Q  Median      3Q     Max
-53.721  -8.016   1.745  10.337  35.005

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) -22.19111    9.24571  -2.400              0.0170 *
hscalc        0.85098    0.07698  11.055 <0.0000000000000002 ***
hsengl        0.20396    0.10053   2.029              0.0433 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 15.75 on 324 degrees of freedom
Multiple R-squared:  0.2915,Adjusted R-squared:  0.2872
F-statistic: 66.67 on 2 and 324 DF,  p-value: < 0.00000000000000022

> betahat = coefficients(mod); betahat
(Intercept)      hscalc      hsengl
-22.1911146   0.8509789   0.2039602
> V = vcov(mod); V # Good approximate asymptotic covariance matrix
            (Intercept)        hscalc        hsengl
(Intercept)  85.4832405 -0.3912663582 -0.7069063563
hscalc       -0.3912664  0.0059253409 -0.0008981144
hsengl       -0.7069064 -0.0008981144  0.0101055340
> gdot = cbind(0,betahat[3],betahat[2]) # A row vector
> hummm = gdot %*% V %*% t(gdot); hummm   # hummm, I wonder what this is.
            [,1]
[1,] 0.004399549
```

continued on page 5

Base your answers on the printout just given.

(a) With a Bonferroni correction for two separate tests, are you able to conclude that both β_1 and β_2 are non-zero at $\alpha = 0.05$? Just answer Yes or No.

(b) Using a calculator and numbers from the printout, calculate a single test statistic for testing $H_0 : \beta_1\beta_2 = 0$. Your answer is a number. Show a little work (there's not much) and **circle your answer**.

(c) With this test, are you able to conclude that both β_1 and β_2 are non-zero at $\alpha = 0.05$? Just answer Yes or No.

continued on page 6

4. (*15 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 response variable. Explanatory variables are Drug type, Mother's live adult weight and Father's live adult weight.

   (a) Write the regression equation for the full model, including ϵ_i. Let x_1 = mother's weight and x_2 = father's weight. There are no product terms yet.

   (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) What is the expected dressed weight of a pig getting Drug 2, whose mother weighed 140 pounds, and whose father weighed 185 pounds? Your answer is a formula involving some β values.

   (d) 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. Allowing 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?

continued on page 7

(e) The table in Part 4b gives the equations for three parallel planes, one for each drug. Write the regression equation for a model in which the planes might not be parallel. You do *not* have to make a table this time.

(f) Suppose you wanted to test whether the three planes are parallel. Give the null hypothesis in symbols. Your answer is a statement or set of statements involving the β values from your regression equation. You don't have to show any work. The most natural guess is correct.

5. (*8 points*) The Titanic was a passenger ship that hit an iceberg and sank on its very first voyage in 1912. It was the largest passenger ship in the world at the time, and supposedly unsinkable. More than 1,500 of the roughly 2,200 passengers and crew died. Data are available, actually there's a built-in R data set in case you want to play with it over the holiday.

Passengers were either in 1st class (where there were some lifeboats), 2nd class or 3d class, and they either lived or died. Let c_2 be an indicator dummy variable for 2nd class and c_3 be an indicator dummy variable for 3d class; $y = 1$ means the passenger survived.

(a) Write a regression equation for the log odds of survival. There is an intercept.

(b) Compared to the odds of survival for a passenger in 1st class, the odds of survival for a passenger in 3d class are _____ times as great. Give the answer in terms of the β values from your regression model. Write your answer in the space below. You don't have to prove it or show any work.

(c) What null hypothesis would you test to determine whether Class (1st versus 2nd versus 3d) was related to survival? Give the answer in terms of the β values from your regression model. Write your answer in the space below.

(d) What null hypothesis would you test to determine whether passengers in 2nd class had a better chance of survival than passengers in 3d class? Give the answer in terms of the β values from your regression model. Write your answer in the space below.

continued on page 8

6. (*8 points*) In the Noise data, subjects of different ages listened to brief political discussions under 5 levels of background noise. The response variable is discrimination score, a measure of how well they could tell what was being said. Interest in the topic being discussed was a covariate. The questions come after the printout, which shows just a small part of the analysis.

```
> rm(list=ls()); options(scipen=999) # To avoid scientific notation
> library(lme4); library(car)
Loading required package: Matrix
> loud = read.table("http://www.utstat.toronto.edu/~brunner/data/legal/noise.data.txt")
> colnames(loud) = c("ident", "interest", "sex", "age", "noise", "time", "discrim")
> head(loud)
  ident interest sex age noise time discrim
1     1      2.5   1   2     1    4    50.7
2     1      2.5   1   2     2    1    27.4
3     1      2.5   1   2     3    3    39.1
4     1      2.5   1   2     4    2    37.5
5     1      2.5   1   2     5    5    35.4
6     2      1.9   1   2     1    3    40.3
> attach(loud); agefactor = factor(age); noisefactor=factor(noise)
> fullmodel = lmer(discrim ~ interest + agefactor*noisefactor + (1 | ident)); Anova(fullmodel, test=’F’)
Analysis of Deviance Table (Type II Wald F tests with Kenward-Roger df)

Response: discrim
                            F Df Df.res          Pr(>F)
interest               9.1415  1     56        0.003766 **
agefactor              7.5322  2     56        0.001268 **
noisefactor           14.1164  4    228 0.0000000002622 ***
agefactor:noisefactor  1.0311  8    228        0.413338
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

(a) Is `interest` a between-cases variable or a within-cases variable?

(b) Is `agefactor` a between-cases factor or a within-cases factor?

(c) Is `noisefactor` a between-cases factor or a within-cases factor?

(d) For each question below, give the F statistic that would be used to help provide an answer. The answers are numbers from the printout.

i. Controlling for interest in the topic and averaging across noise levels, is age related to discrimination score?

ii. Controlling for interest in the topic, does the effect of noise level depend on age?

iii. Controlling for interest in the topic, does the effect of age depend on noise level?

iv. Controlling for interest in the topic, is there an interaction between age and noise level?

v. Controlling for noise level and age, is interest in the topic related to discrimination score?

continued on page 9

7. (*15 points*) When explanatory variables in a regression are random and measured with error (very common), the usual least-squares estimates ignoring measurement error are generally inconsistent. However, with two imperfect but independent measurements of each explanatory variable, things can be better. Here is perhaps the simplest example. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} W_{i,1} &= X_i + \delta_{i,1} \\ W_{i,2} &= X_i + \delta_{i,2} \\ Y_i &= \beta X_i + \epsilon_i, \end{aligned}$$

where

- X_i is a latent, unobservable variable. We can observe only the values of $W_{i,1}$, $W_{i,2}$ and Y_i.
- X_i, $\delta_{i,1}$, $\delta_{i,2}$ and ϵ_i are all independent with expected value zero.
- $Var(X_i) = \sigma^2_x > 0$, $Var(\delta_{i,1}) = \sigma^2_1$, $Var(\delta_{i,2}) = \sigma^2_2$, and $Var(\epsilon_i) = \sigma^2_\epsilon$.

(a) Calculate the variance-covariance matrix of the vector of observable variables $(W_{i,1}, W_{i,2}, Y_i)^\top$. This is a set of scalar calculations that you put in a matrix when you are done. You may leave the lower triangle blank if you wish. Don't show more work than you need to.

continued on page 10

(b) Suggest an estimator of β. Call it $\widehat{\beta}_n$. Remember, $\widehat{\beta}_n$ must be a function of the observable data.

(c) Prove that $\widehat{\beta}_n$ is a consistent estimator of β. You have more room than you need.

continued on page 11

8. (*17 points*) Using the `Math` data, we investigate choice of university Calculus course as a function of High School Calculus mark and sex. The questions come after the printout.

```
> # Choice of university course based on High School data, sex and first language
> # install.packages("mlogit", dependencies=TRUE) # Only need to do this once
> library(mlogit) # Load the package every time
>
> datta = math[,c(1,5,9)] # Just course, hscalc and sex
> datta = na.omit(datta)
> summary(datta); attach(datta)
      course        hscalc       sex
 Catch-up: 20   Min.   : 50.00   F:193
 Elite   : 28   1st Qu.: 67.00   M:186
 Mainstrm:331   Median : 77.00
                Mean   : 76.09
                3rd Qu.: 86.00
                Max.   :100.00
>
> # Make Mainstream the reference category for course by changing alphabetical order.
> n = length(course); Course = character(n)
> Course[course==’Mainstrm’] = ’1_Mainstrm’
> Course[course==’Elite’] = ’2_Elite’
> Course[course==’Catch-up’] = ’3_Catch-up’
> Course = factor(Course); table(Course)
Course
1_Mainstrm    2_Elite 3_Catch-up
       331         28         20
> datta$course = Course # Put the fixed-up version back in the data frame
>
> # Make an mlogit data frame in long format
> long = mlogit.data(datta,shape="wide",choice="course")
>
> # Fit full model
> full = mlogit(course ~ 0 | hscalc + sex, data=long)
> summary(full)

Call:
mlogit(formula = course ~ 0 | hscalc + sex, data = long, method = "nr",
    print.level = 0)

Frequencies of alternatives:
1_Mainstrm    2_Elite 3_Catch-up
  0.873351   0.073879   0.052770

nr method
6 iterations, 0h:0m:0s
g’(-H)^-1g = 1.94E-07
gradient close to zero

Coefficients :
                        Estimate Std. Error t-value  Pr(>|t|)
2_Elite:(intercept)    -6.277453   1.573941 -3.9884 6.653e-05 ***
3_Catch-up:(intercept)  4.213750   1.472793  2.8611  0.004222 **
2_Elite:hscalc          0.036789   0.018833  1.9535  0.050762 .
3_Catch-up:hscalc      -0.107888   0.023086 -4.6732 2.965e-06 ***
2_Elite:sexM            1.411205   0.475800  2.9660  0.003017 **
3_Catch-up:sexM         0.733976   0.497212  1.4762  0.139895
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

continued on page 12

```
Log-Likelihood: -153.11
McFadden R^2:  0.13307
Likelihood ratio test : chisq = 47.003 (p.value = 1.5226e-09)
>
> # Restricted models
> NoCalculus = mlogit(course ~ 0 | sex, data=long)
> summary(NoCalculus)

Call:
mlogit(formula = course ~ 0 | sex, data = long, method = "nr",
    print.level = 0)

Frequencies of alternatives:
1_Mainstrm    2_Elite 3_Catch-up
  0.873351   0.073879   0.052770

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 3.08E-08
gradient close to zero

Coefficients :
                       Estimate Std. Error t-value  Pr(>|t|)
2_Elite:(intercept)    -3.39563    0.41503 -8.1816  2.22e-16 ***
3_Catch-up:(intercept) -3.10794    0.36137 -8.6005 < 2.2e-16 ***
2_Elite:sexM            1.46279    0.47359  3.0887   0.00201 **
3_Catch-up:sexM         0.56897    0.46957  1.2117   0.22564
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -170.31
McFadden R^2:  0.035674
Likelihood ratio test : chisq = 12.601 (p.value = 0.0018356)

> anova(NoCalculus, full)
Error in UseMethod("anova") :
  no applicable method for 'anova' applied to an object of class "mlogit"

> NoSex      = mlogit(course ~ 0 | hscalc, data=long)
> summary(NoSex)

Call:
mlogit(formula = course ~ 0 | hscalc, data = long, method = "nr",
    print.level = 0)

Frequencies of alternatives:
1_Mainstrm    2_Elite 3_Catch-up
  0.873351   0.073879   0.052770

nr method
6 iterations, 0h:0m:0s
g'(-H)^-1g = 1.56E-07
gradient close to zero
```

continued on page 13

```
Coefficients :
                        Estimate Std. Error t-value  Pr(>|t|)
2_Elite:(intercept)    -5.641500   1.514151 -3.7258 0.0001947 ***
3_Catch-up:(intercept)  4.472142   1.453862  3.0760 0.0020977 **
2_Elite:hscalc          0.040052   0.018460  2.1696 0.0300373 *
3_Catch-up:hscalc      -0.106000   0.022873 -4.6343 3.581e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Log-Likelihood: -159.34
McFadden R^2:  0.097752
Likelihood ratio test : chisq = 34.528 (p.value = 3.1797e-08)
```

(a) The tests `summary` output for the full model include two tests (excluding tests for the intercepts) that are statistically significant. In plain, non-statistical language and mentioning *no numbers*, give the conclusions from these two tests. You have more room than you need.

continued on page 14

(b) We seek a *single* test of the relationship between sex and choice of university Calculus course, controlling for mark in High School Calculus.

   i. Write the numerical value of the test statistic in the space below. The answer is a number. If you need to calculate this number from material on the printout, show a little work. **Circle the number**.

   ii. What is the critical value? The answer is a number from the formula sheet.

   iii. Do you reject the null hypothesis? Answer Yes or No.

   iv. Is there evidence that sex is related to choice of Calculus course, controlling for High School performance? Just answer Yes or No.

Total Marks = 100 points