UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2011 Final Examination
STA442H1F/2101HF
Methods of Applied Statistics
Jerry Brunner
Duration - 3 hours

Aids: Calculator Model(s): Any calculator is okay

Last/Surname (Print): ______________________

First/Given Name (Print): ______________________

Student Number: ______________________

Signature: ______________________

Qn. #	Value	Score
1	20	
2	20	
3	10	
4	15	
5	5	
6	12	
7	8	
8	10	
Total = 100 Points		

20 points

1. Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where $E(X_i) = \mu$, $Var(X_i) = \sigma^2_x$, $E(\epsilon_i) = 0$, $Var(\epsilon_i) = \sigma^2_\epsilon$, and ϵ_i is independent of X_i. Let

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

(a) Is $\widehat{\beta}_1$ a consistent estimator of β_1? Answer Yes or No and **Circle Yes or No**. Prove your answer.

(b) For some special cases we have $\widehat{\beta}_1 \stackrel{a.s.}{\rightarrow} \beta_1$. When does this happen?

continued on page 3

20 points

2. Let $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2\mathbf{I}_n$, where $\sigma^2 > 0$ is an unknown constant. You may use without proof the fact that if $\mathbf{T} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{AT} \sim N(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}')$.

   (a) In the regression model, what is the distribution of $\mathbf{Y}$? No proof is needed.

   (b) The maximum likelihood estimate of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ (no proof needed). What is the distribution of $\widehat{\boldsymbol{\beta}}$? Show the calculations.

continued on page 4

10 points

3. In a study of math education in elementary school, equal numbers of boys and girls were randomly assigned to one of three training programmes designed to improve spatial reasoning. After five school days of training, the students were given a standardized test of spatial reasoning. Score on the spatial reasoning test is the response variable. You will define a regression model for this factorial analysis of variance. Don't write the model yet.

   (a) In the table below, show how your dummy variables are defined. *Use effect coding.* Write the name of each dummy variable at the head of its column.

Girls, Programme 1	
Girls, Programme 2	
Girls, Programme 3	
Boys, Programme 1	
Boys, Programme 2	
Boys, Programme 3	

   (b) Give $E[Y_i|\mathbf{X}_i = \mathbf{x}_i]$ for the full model. Include the interaction terms.

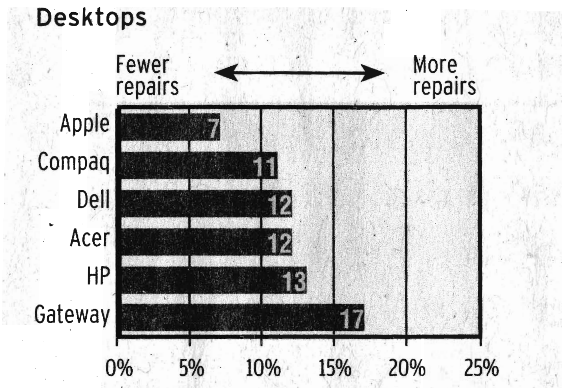
   (c) Suppose you want to test whether, averaging across training programmes, there is a difference between girls and boys in their average performance on the spatial reasoning test. State the null hypothesis in terms of the β values in your model.

   (d) Suppose you want to test whether the average sex difference in performance depends on which training programme the children are in. State the null hypothesis in terms of the β values in your model.

continued on page 5

15 points

4. The following was scanned from the 2012 *Consumer Reports Buying Guide*. It shows repair history for several desktop computer brands, and seems to be based on a fairly sophisticated statistical analysis.



Gateway has been among the more repair-prone brands of desktop computers and Apple has been among the least. That's what we found when we asked almost 20,000 readers who bought a desktop computer between 2008 and 2011 about their experiences. The graph shows the percentage of models for each brand that were repaired or had a serious problem. Differences of fewer than 5 points aren't meaningful, and we've adjusted the data to eliminate differences linked solely to the age of the desktop computer. Models within a brand can vary, and design or manufacture changes might affect future reliability.

(a) What statistical method do you guess they employed?

(b) What words above suggest that they did some kind of hypothesis tests?

(c) What was the covariate?

(d) How would you set up dummy variables for the Brand of Computer? *Make a table in the space beside the scanned material.*

(e) Assuming there is no interaction, write an expression (a function of the $\widehat{\beta}$ values) that would give you the number 17 for Gateway computers. Denote the covariate by x.

continued on page 6

5 points

5. On the `Computer Printout`, the output for the `Beta data` is based on a random sample of size $n = 50$ from a beta distribution.

   (a) What is the maximum likelihood estimate $\widehat{\alpha}$? The answer is a number from the printout.

   (b) Carry out a test of $H_0 : \alpha = 5$ versus $H_1 : \alpha \neq 5$. Calculate the test statistic. The answer is a number. Show a little work. **Circle the number.** Do you reject H_0 at $\alpha = 0.05$? **Answer Yes or No.** You have more room than you need.

continued on page 7

12 points

6. On the `Computer Printout`, the output for the `Cars data` is based on the same metric cars data you analyzed for homework. Potentially there are three regression lines relating weight of car to fuel efficiency.

   (a) We wish to know whether country differences in fuel efficiency depend on the weight of the car. Fill in the table below.

F Statistic	Degrees of Freedom (2 numbers)	p-value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

   (b) Do country differences in fuel efficiency depend on the weight of the car? Answer Yes or No.

   (c) Are the three regression lines parallel *in the population*? Answer Yes or No.

   (d) What is the estimated expected fuel efficiency for a U.S. car of average weight (meaning average for the entire sample)? The answer is a single number.

   (e) What is the estimated slope of the regression line for U.S. cars? The answer is a single number.

   (f) To show which slopes are different from one another, make a table whose i, j element is the Bonferroni-adjusted p-value for the tests of difference between the slope for country i and country j. Just fill in the upper triangular part of the table. Use your calculator to convert p-values on the printout to Bonferroni-adjusted p-values.

   (g) Based on the multiple comparisons, which slopes are really different? Don't just say they're different; say which one is steeper (going down faster).

continued on page 8

8 points

7. On the `Computer Printout`, the output for the `Birth Weight Data` is based on the same data discussed in lecture.

   (a) The estimated odds of a low birth weight baby are _____ times as great for a mother with a history of premature labour. The answer is a single number; write it on the line.

   (b) You want to know whether *any* of the variables in the model are related to the chances that a baby will have low birth weight. Fill in the table below.

Wald χ^2 Statistic	Degrees of Freedom	p-value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

   (c) Does the preceding test indicate that at least one of the explanatory variables is related to the response variable? Answer Yes or No.

   (d) Give an approximate 95% confidence interval for the regression coefficient corresponding to mother's weight. Your answer is a pair of numbers, and there is more than one way to calculate them from the numbers on the printout. You have more room than you need.

continued on page 9

10 points

8. On the `Computer Printout`, the output for the `Dichotic Listening Study` comes from data given in the Final Assignment. Focus on this question. Does mode of presentation (Left *versus* Right *versus* Both) influence performance?

   (a) There is a single test for the question of interest. Record the numbers below.

F Statistic	Degrees of Freedom (2 numbers)	p-value	Reject H_0 at $\alpha = 0.05$? (Yes or No)

   (b) Only one of the pairwise comparisons of marginal means is statistically significant using a multiple comparison method. Give the Bonferroni-adjusted p-value. The answer is a single number.

   (c) Describe the difference in simple, non-statistical language. Be sure you say which mean is bigger.

STA442/2101 Final Exam Printout

Table of Contents Page

Critical Values

```
> qnorm(0.975)
[1] 1.959964
> DF = 1:5
> CritVal = qchisq(0.95,DF); cbind(DF,CritVal)
     DF   CritVal
[1,]  1  3.841459
[2,]  2  5.991465
[3,]  3  7.814728
[4,]  4  9.487729
[5,]  5 11.070498
```

Beta Data

```
> x <- scan("http://www.utstat.toronto.edu/~brunner/appliedf11/data/beta.data")
Read 50 items
> BLL <- function(ab,datta) # - Loglike of beta
+     {
+     n <- length(datta)
+     BLL <- n*lgamma(ab[1]) + n*lgamma(ab[2]) - n*lgamma(sum(ab)) -
+     (ab[1]-1)*sum(log(datta)) - (ab[2]-1)*sum(log(1-datta))
+     if(ab[1] <= 0) BLL <- Inf ;  if(ab[2] <= 0) BLL <- Inf
+     BLL
+     }
> fit1 <- nlminb(c(1,1),objective=BLL,datta=x); fit1
$par
[1] 13.96757 27.27781

$objective
[1] -60.26451

$convergence
[1] 0

$message
[1] "relative convergence (4)"

$iterations
[1] 20

$evaluations
function gradient
      21       50
```

```
> fit2 <- nlm(BLL,fit1$par,hessian=T,datta=x); fit2
$minimum
[1] -60.26451

$estimate
[1] 13.96757 27.27781

$gradient
[1] -1.008464e-07 -4.071882e-08

$hessian
          [,1]       [,2]
[1,]  2.483506 -1.2270086
[2,] -1.227009  0.6398216

$code
[1] 3

$iterations
[1] 1

> huh = solve(fit2$hessian); huh
          [,1]     [,2]
[1,]  7.667046 14.70337
[2,] 14.703367 29.76010
>
```

Cars Data

```
/********************** FinalCars.sas *************************/
options linesize=79 pagesize=100 noovp formdlim='-' nodate;
title 'Metric Cars Data: STA442/2101 Fall 2011 Final Exam';

data auto;
     infile 'mcars2.data' firstobs=2 ;     /* Skipping the header on line 1 */
     input id country $ kpl weight length;
     if country = 'US' then c1=1;
        else if country = 'Japan' then c1=0;
        else if country = 'Europ' then c1=0;
     if country = 'Europ' then c2=1;
        else if country = 'US' then c2=0;
        else if country = 'Japan' then c2=0;
     cweight = weight - 1413.45; /* Subtract off mean weight */
     cwc1 = cweight*c1;
     cwc2 = cweight*c2;
     label country = 'Country of Origin'
           kpl = 'Kilometers per Litre'
           weight = 'Weight in kg'
           cweight = 'Centered Weight'
           length = 'Length in cm';
```

```
proc reg;
     model kpl = cweight c1 c2 cwc1 cwc2;
     C1_eq_C2:          test c1=c2;
     C1_eq_C2_eq_0:     test c1=c2=0;
     CWC1_eq_CWC2:      test cwc1=cwc2;
     CWC1_eq_CWC2_eq_0: test cwc1=cwc2=0;
```

The REG Procedure
Model: MODEL1
Dependent Variable: kpl Kilometers per Litre

Number of Observations Read	100
Number of Observations Used	100

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	489.27223	97.85445	42.46	<.0001
Error	94	216.61706	2.30444		
Corrected Total	99	705.88930			

Root MSE	1.51804	R-Square	0.6931
Dependent Mean	8.79480	Adj R-Sq	0.6768
Coeff Var	17.26062		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value
Intercept	Intercept	1	3.36821	1.53516	2.19
cweight	Centered Weight	1	-0.01827	0.00418	-4.37
c1		1	5.45383	1.54696	3.53
c2		1	3.73906	1.69123	2.21
cwc1		1	0.01304	0.00422	3.09
cwc2		1	0.00611	0.00453	1.35

Parameter Estimates

Variable	Label	DF	Pr > |t|
Intercept	Intercept	1	0.0307
cweight	Centered Weight	1	<.0001
c1		1	0.0007
c2		1	0.0295
cwc1		1	0.0026
cwc2		1	0.1810

Metric Cars Data: STA442/2101 f2011 Final Exam 2

The REG Procedure
Model: MODEL1

Test C1_eq_C2 Results for Dependent Variable kpl

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	12.55043	5.45	0.0217
Denominator	94	2.30444		

Metric Cars Data: STA442/2101 f2011 Final Exam 3

The REG Procedure
Model: MODEL1

Test C1_eq_C2_eq_0 Results for Dependent Variable kpl

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	20.01055	8.68	0.0003
Denominator	94	2.30444		

Metric Cars Data: STA442/2101 f2011 Final Exam 4

The REG Procedure
Model: MODEL1

Test CWC1_eq_CWC2 Results for Dependent Variable kpl

Source	DF	Mean Square	F Value	Pr > F
Numerator	1	33.02284	14.33	0.0003
Denominator	94	2.30444		

Metric Cars Data: STA442/2101 f2011 Final Exam 5

The REG Procedure
Model: MODEL1

Test CWC1_eq_CWC2_eq_0 Results for Dependent Variable kpl

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	26.53036	11.51	<.0001
Denominator	94	2.30444		

Birth Weight Data

```
title2 'STA442/2101f11 Final Exam';
%include 'bweightread.sas';
     label lwt   =  'Weight at Last Period'
           ptl   =  'History of Premature Labour (1=Yes, 0=No)'
           ht    =  'History of Hypertension (1=Yes, 0=No)';

proc logistic;
     model low (event='Under 2500 g') = lwt ptl ht / covb;
```

```
                            Low Birth Weight Data                           1

                           The LOGISTIC Procedure

                              Model Information

      Data Set                      WORK.BIGBABY
      Response Variable             low                    Low Birth Weight
      Number of Response Levels     2
      Model                         binary logit
      Optimization Technique        Fisher's scoring


                  Number of Observations Read         189
                  Number of Observations Used         189


                               Response Profile

                   Ordered                          Total
                     Value     low              Frequency

                         1     2500 g +               130
                         2     Under 2500 g            59

                Probability modeled is low='Under 2500 g'.


                         Model Convergence Status

              Convergence criterion (GCONV=1E-8) satisfied.


                            Model Fit Statistics

                                                Intercept
                                 Intercept            and
                  Criterion           Only     Covariates

                  AIC              236.672        218.123
                  SC               239.914        231.090
                  -2 Log L         234.672        210.123
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	24.5486	3	<.0001
Score	24.2151	3	<.0001
Wald	20.1449	3	0.0002

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.0171	0.8533	1.4209	0.2333
lwt	1	-0.0173	0.00679	6.4812	0.0109
ptl	1	1.4067	0.4285	10.7778	0.0010
ht	1	1.8939	0.7211	6.8984	0.0086

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
lwt	0.983	0.970	0.996
ptl	4.083	1.763	9.455
ht	6.645	1.617	27.308

Association of Predicted Probabilities and Observed Responses

Percent Concordant	71.1	Somers' D	0.438
Percent Discordant	27.3	Gamma	0.445
Percent Tied	1.6	Tau-a	0.189
Pairs	7670	c	0.719

Estimated Covariance Matrix

Parameter	Intercept	lwt	ptl	ht
Intercept	0.728149	-0.00564	-0.04289	0.17753
lwt	-0.00564	0.000046	0.000051	-0.00173
ptl	-0.04289	0.000051	0.183611	0.018931
ht	0.17753	-0.00173	0.018931	0.519955

Dichotic Listening Data

```
> ear =
read.table("http://www.utstat.toronto.edu/~brunner/appliedf11/data/Dichotic.data")
> attach(ear)
> # Hotelling's T-squared for H0: L mu = h
> HTest = function(datta,L,h=0)
+     {
+     HTest = numeric(5)
+     names(HTest) = c("T-squared","F","df1","df2","p-value")
+     xbar = apply(datta,2,mean)
+     n = dim(datta)[1]; k = dim(datta)[2]; r = dim(L)[1]
+     if(dim(L)[2] != k) stop("L and data matrix incompatible sizes")
+     T2 = n * t(L%*%xbar-h) %*% solve(L%*%var(datta)%*%t(L)) %*% (L%*%xbar-h)
+     T2 = as.numeric(T2); F = (n-r)/(r*(n-1)) * T2
+     pval = 1-pf(F,r,n-r)
+     HTest = c(T2,F,r,n-r,pval)
+     names(HTest) = c("T-squared","F","df1","df2","p-value")
+     round(HTest,5)
+     } # End function HTest
>
> ear[1:5,]
  test11 test12 test13 test21 test22 test23 test31 test32 test33
1     13     12     10     15     14     14     14     13     14
2      4      8      8      6      5      8      6      3      4
3     13     15     11     11     13     15     11     13     12
4      7      5      4      6      7      3      6      7      6
5     11     12     14      9     11      8     12     10     11
>
> # First some descriptive statistics
> Xbar = apply(ear,2,mean); Xbar
  test11   test12   test13   test21   test22   test23   test31   test32   test33
9.444444 9.592593 9.197531 9.111111 9.654321 8.950617 8.851852 9.308642 8.567901
> mean(test12)
[1] 9.592593
> cellmeans = Xbar; dim(cellmeans) <- c(3,3); cellmeans
         [,1]     [,2]     [,3]
[1,] 9.444444 9.111111 8.851852
[2,] 9.592593 9.654321 9.308642
[3,] 9.197531 8.950617 8.567901
> cellmeans = t(cellmeans)
> rownames(cellmeans) <- c("Left","Right","Both")
> colnames(cellmeans) <- c("HipHop","Classc","Radio")
>
```

```
> Xbar
  test11   test12   test13   test21   test22   test23   test31   test32   test33
9.444444 9.592593 9.197531 9.111111 9.654321 8.950617 8.851852 9.308642 8.567901
> cellmeans
        HipHop   Classc    Radio
Left  9.444444 9.592593 9.197531
Right 9.111111 9.654321 8.950617
Both  8.851852 9.308642 8.567901
> # Marginal Means
> apply(cellmeans,1,mean)
    Left    Right     Both
9.411523 9.238683 8.909465
> apply(cellmeans,2,mean)
  HipHop   Classc    Radio
9.135802 9.518519 8.905350
>
> # Tests
>
> C1 = rbind(c(1,1,1,-1,-1,-1,0,0,0),
+            c(0,0,0,1,1,1,-1,-1,-1)  )
> HTest(ear,C1)
T-squared         F       df1       df2   p-value
  8.57581   4.23430   2.00000  79.00000   0.01791
>
> C2 = rbind(c(1,-1,0,1,-1,0,1,-1,0),
+            c(0,1,-1,0,1,-1,0,1,-1)  )
> HTest(ear,C2)
T-squared         F       df1       df2   p-value
 18.03113   8.90287   2.00000  79.00000   0.00033
>
> C3 = rbind(c(1,-1,0,-1,1,0,0,0,0),
+            c(0,1,-1,0,-1,1,0,0,0),
+            c(0,0,0,1,-1,0,-1,1,0),
+            c(0,0,0,0,1,-1,0,-1,1))
> HTest(ear,C3)
T-squared         F       df1       df2   p-value
  1.59136   0.38292   4.00000  77.00000   0.82021
>
> C4 = rbind(c(1,1,1,-1,-1,-1,0,0,0))
> HTest(ear,C4)
T-squared         F       df1       df2   p-value
  1.19640   1.19640   1.00000  80.00000   0.27733
> C5 = rbind(c(1,1,1,0,0,0,-1,-1,-1))
> HTest(ear,C5)
T-squared         F       df1       df2   p-value
  8.55538   8.55538   1.00000  80.00000   0.00448
> C6 = rbind(c(0,0,0,1,1,1,-1,-1,-1))
> HTest(ear,C6)
T-squared         F       df1       df2   p-value
  3.41493   3.41493   1.00000  80.00000   0.06831
>
> C7 = rbind(c(1,-1,0,1,-1,0,1,-1,0))
> HTest(ear,C7)
T-squared         F       df1       df2   p-value
  6.73559   6.73559   1.00000  80.00000   0.01124
> C8 = rbind(c(1,0,-1,1,0,-1,1,0,-1))
> HTest(ear,C8)
T-squared         F       df1       df2   p-value
  1.66406   1.66406   1.00000  80.00000   0.20077
> C9 = rbind(c(0,1,-1,0,1,-1,0,1,-1))
> HTest(ear,C9)
T-squared         F       df1       df2   p-value
 15.85559  15.85559   1.00000  80.00000   0.00015
```