

UNIVERSITY OF TORONTO
Faculty of Arts and Science

December 2014 Final Examination

STA442H1F/2101H1F

Methods of Applied Statistics

Jerry Brunner

Duration - 3 hours

Aids: Calculator Model(s): Any calculator without wireless capability is okay.
Formula sheet supplied.

Last/Family Name (Print): _____

First/Given Name (Print): _____

Student Number: _____

Signature: _____

Qn. #	Value	Score
1	10	
2	12	
3	12	
4	14	
5	10	
6	18	
7	12	
8	12	
Total = 100 Points		

1. (10 points) Mantids are insects, like crickets or grasshoppers. When frightened, they emit loud noises that function as alarm calls. I believe they make the sounds by rubbing their hind legs together. The frequency (number of calls per minute) may indicate how alarmed the mantids are. In one study, caged mantids (either Female or Male) were randomly assigned to be exposed to one of four predators (birds), and the number of alarm calls per minute was recorded.
 - (a) Write a regression equation ($Y_i = \dots$ and so on) in which expected number of alarm calls is a function of Sex and Predator. You will use *cell means coding*, with an indicator for each treatment combination. Just give the regression equation in the space below.
 - (b) Make a table with one row for each treatment combination, showing how the dummy variables are defined. Make one more column with $E(Y|\mathbf{x})$.
 - (c) In terms of the β values from your regression model, give the null hypothesis you would test to answer each of the following questions.
 - i. Are there *any* differences among the expected values for the eight treatments?
 - ii. Does the effect of Predator depend upon Sex of Mantid?
 - iii. Averaging over predators, is there a Sex difference in the average frequency of alarm calls the predators elicit?
 - iv. Averaging over males and females, are there any differences among predators in the expected frequency of alarm calls they elicit?

2. (12 points) Independently for $i = 1, \dots, n$, let

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

where $E(X_i) = E(\epsilon_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i . Let W_1, \dots, W_n be independent random variables with $E(W_i) = \mu_w \neq 0$ and $Var(W_i) = \sigma_w^2$. The random variables W_1, \dots, W_n are independent of X_1, \dots, X_n and $\epsilon_1, \dots, \epsilon_n$. Is

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

a consistent estimator of β_1 ? Answer Yes or No and prove your answer.

3. (12 points) This question is based on the following R output for the Birth Weight Study.

```
> # 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

```
> # For Wald Tests: Wtest = function(L,Tn,Vn,h=0)
> source("http://www.utstat.utoronto.ca/~brunner/Rfunctions/Wtest.txt")
> betahat = baby$coefficients; Vhat = vcov(baby)
> L1 = rbind(c(0,1,0,0,0),
+           c(0,0,1,0,0))
> Wtest(L1,betahat,Vhat)
           W           df      p-value
12.960120803 2.000000000 0.001533718
> L2 = rbind(c(0,0,0,1,0),
+           c(0,0,0,0,1))
> Wtest(L2,betahat,Vhat)
           W           df      p-value
8.73050440 2.000000000 0.01271145
> L3 = rbind(c(0,0,0,1,-1))
> Wtest(L3,betahat,Vhat)
           W           df      p-value
0.3696971 1.0000000 0.5431694
```

- (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 compared to a White mother. The answer is a number. Write your answer in the space below. **Circle your answer.**
- (c) We want to know whether *any* of the explanatory variables are related to the chances of having a low birth weight baby.
- Give the value of test statistic. The answer is a number.
 - What is the critical value at $\alpha = 0.05$? The answer is a number.
 - Do you reject the null hypothesis? Answer Yes or No.
 - Are any of the explanatory variables 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, race is related to the chances of having a low birth weight baby.
- Give the value of test statistic. The answer is a number.
 - What is the critical value at $\alpha = 0.05$? The answer is a number.
 - Do you reject the null hypothesis? Answer Yes or No.
- (e) Carry out Bonferroni-corrected pairwise comparisons of the three racial groups.
- Give the three Bonferroni-corrected p -values.
 - Guided *strictly* by the joint $\alpha = 0.05$ significance level, what do you conclude? Use plain, non-statistical language.

4. (*14 points*) For the linear model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$,
- (a) Calculate $\mathbf{X}^\top \mathbf{e}$ 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 add up to zero?

- (c) When a regression model has no intercept, residuals do not always add up to zero, but sometimes they do. Now we will see that if a linear combination of the explanatory variables equals a column of ones, the the sum of residuals is equal to zero.

Let $\mathbf{1}$ denote an $n \times 1$ vector of ones. Show that if there is a $p \times 1$ vector \mathbf{v} with $\mathbf{X}\mathbf{v} = \mathbf{1}$, then $\sum_{i=1}^n e_i = 0$.

- (d) Suppose that a regression model has no intercept, but it does include a categorical explanatory variable with cell means dummy variable coding. How do you know that the sum of the residuals is zero?

5. (10 points) This question is based on the following R output for the Bunnies data.

```
> # Bunnies: Response variable will be Force required to pull out the implant.
> bunnies = read.table("http://www.utstat.toronto.edu/~brunner/appliedf14/code_n_data/hw/bunnies2.data",header=T)
> attach(bunnies)
> aggregate(Force,list(Time,Drug),FUN=mean)
  Group.1 Group.2      x
1       3       0 56.22
2       6       0 70.68
3       9       0 191.18
4      12       0 275.86
5       3       1 36.62
6       6       1 52.62
7       9       1 71.26
8      12       1 89.76
> table(Drug,Time)
      Time
Drug 3 6 9 12
     0 5 5 5 5
     1 5 5 5 5
> Time = factor(Time)
> rabbit = lm(Force ~ Drug + Time + Drug:Time); anova(rabbit)
Analysis of Variance Table

Response: Force
      Df Sum Sq Mean Sq F value    Pr(>F)
Drug    1  73822   73822 11.0417 0.002239 **
Time    3 120516   40172  6.0086 0.002287 **
Drug:Time 3  50488   16829  2.5172 0.075737 .
Residuals 32 213946    6686
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The main question in this study is whether the HEBP drug helps the dental implants become more firmly attached to the bone. What do you conclude about this question? Use plain, non-statistical language. Your answer is a statement about the attachment of dental implants to bone. Do not use the word “force” in your answer. If you do, you get zero marks for this question.

6. (18 points) A forestry company has developed a regression equation for predicting the amount of useable wood that they will get from a tree, based on a set of measurements that can be taken without cutting the tree down. They are convinced that a model with normal error terms is right. They have $\hat{\beta}$ and MSE based on a set of n trees they measured first and then cut down, and they know how to calculate a predicted Y and a prediction interval for the amount of wood they will get from a single tree.

But that's not what they want. They have a set of k more trees they are planning to cut down, and they have measured several explanatory variables for each tree, yielding $\mathbf{x}_{n+1} \dots, \mathbf{x}_{n+k}$. What they want is a prediction of the *total* amount of wood they will get from these trees, along with a 95% prediction interval for the total.

- (a) The quantity they want to predict is $W = \sum_{j=n+1}^{n+k} Y_j$, where $Y_j = \mathbf{x}_j^\top \beta + \epsilon_j$. What is the distribution of W ? You can just write down the answer without showing any work.
- (b) Let \widehat{W} denote the prediction of W . It is calculated using the company's regression data along with $\mathbf{x}_{n+1} \dots, \mathbf{x}_{n+k}$. Give a formula for \widehat{W} .
- (c) What is the distribution of $W - \widehat{W}$? Show your work, but don't use moment-generating functions. Just write down expected value and calculate the variance.

- (d) Now standardize $W - \widehat{W}$ to obtain a standard normal. Call it Z .
- (e) Divide Z by the square root of a chi-squared random variable, divided by its degrees of freedom, and simplify. Call it T . What are the degrees of freedom?
- (f) How do you know that numerator and denominator are independent?

- (g) Using your formula for T , derive the $(1 - \alpha) \times 100\%$ prediction interval for W . Please use the symbol $t_{\alpha/2}$ for the critical value.

7. (12 points) This question is based on the following R output for the Salmon data.

```
> fish = read.table("http://www.utstat.toronto.edu/~brunner/appliedf14/code_n_data/hw/salmon2.data", header=T)
> head(fish); attach(fish)
  Country Gender Fresh Marine
1 Alaskan  Male   108   368
2 Alaskan Female   131   355
3 Alaskan Female   105   469
4 Alaskan  Male    86   506
5 Alaskan Female    99   402
6 Alaskan  Male    87   423
> table(Country,Gender)
      Gender
Country  Female Male
Alaskan     26   24
Canadian     26   24
> AveGrowth = (Fresh+Marine)/2; Diff = Fresh-Marine
> # Effect coding (with the minus ones). R uses ALPHABETICAL ORDER
> contrasts(Country) = contr.sum(2); contrasts(Gender) = contr.sum(2)
> Amodel = lm(AveGrowth ~ Country + Gender + Country:Gender); summary(Amodel)
```

Call:

```
lm(formula = AveGrowth ~ Country + Gender + Country:Gender)
```

Residuals:

Min	1Q	Median	3Q	Max
-35.62	-13.36	-1.76	12.41	38.73

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	258.0649	1.8509	139.426	< 2e-16 ***
Country1	6.1178	1.8509	3.305	0.00134 **
Gender1	-0.8726	1.8509	-0.471	0.63840
Country1:Gender1	-3.1947	1.8509	-1.726	0.08756 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.49 on 96 degrees of freedom

Multiple R-squared: 0.1248, Adjusted R-squared: 0.09747

F-statistic: 4.564 on 3 and 96 DF, p-value: 0.004939

```
> Dmodel = lm(Diff ~ Country + Gender + Country:Gender); summary(Dmodel)
```

Call:

```
lm(formula = Diff ~ Country + Gender + Country:Gender)
```

Residuals:

Min	1Q	Median	3Q	Max
-91.167	-26.135	0.269	23.266	103.077

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-280.301	3.896	-71.951	<2e-16 ***
Country1	-51.154	3.896	-13.131	<2e-16 ***
Gender1	2.032	3.896	0.522	0.603
Country1:Gender1	2.346	3.896	0.602	0.548

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38.93 on 96 degrees of freedom
 Multiple R-squared: 0.6427, Adjusted R-squared: 0.6316
 F-statistic: 57.57 on 3 and 96 DF, p-value: < 2.2e-16

```
>
> # Want to look at Country by Environment
> aggregate(cbind(Fresh,Marine),list(Country),FUN=mean)
  Group.1  Fresh Marine
1  Alaskan  98.38 429.66
2 Canadian 137.46 366.62
>
> t.test(Diff[Country=="Alaskan"]) # Matched t-test for Alaskan salmon
```

One Sample t-test

```
data: Diff[Country == "Alaskan"]
t = -51.9164, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -344.1032 -318.4568
sample estimates:
mean of x
 -331.28
```

```
> t.test(Diff[Country=="Canadian"]) # Matched t-test for Canadian salmon
```

One Sample t-test

```
data: Diff[Country == "Canadian"]
t = -52.5082, df = 49, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -237.9303 -220.3897
sample estimates:
mean of x
 -229.16
```

- (a) Think of this as a three-factor design: Country by Gender by Environment (Fresh versus Marine). Identify each factor as either within-cases or between-cases.

Country

Gender

Environment

- (b) Write numbers in the table below. The test statistics may be either t or F . Scientific notation is okay for the p -values if you wish. .

Effect	Test Statistic	p-value
Country		
Gender		
Environment		
Country \times Gender		
Country \times Environment		
Gender \times Environment		
Country \times Gender \times Environment		

- (c) One of the main effects is statistically significant at $\alpha = 0.05$, but you would *not* want to interpret it. What's the effect and why should you not discuss it?

- (d) Express the conclusions of this analysis in plain, non-statistical language. Be guided by the $\alpha = 0.05$ significance level, but do not mention any statistical terms, and do not give any numbers. You are being asked for a few sentences about the growth of fish.

8. (12 points) Let Y_1, \dots, Y_n be independent and identically distributed $N(\mu, \sigma^2)$ random variables. We are interested in the power of a test of $H_0 : \mu = \mu_0$ against the alternative that $\mu \neq 0$. The easy way out is to use a regression model with an intercept but no explanatory variables, so that the null hypothesis is a statement about the intercept.

(a) Calculate the non-centrality parameter λ and simplify.

(b) What is “effect size” for this problem?