Name Jerry

Student Number

STA 442/2101 F 2014 Quiz 9

1. U of T administration is very interested in whether the chances of success are different on the three campuses for undergraduate students with similar performance in High School. So, the Statistical Consulting Service carried out a logistic regression analysis in which

$$\log \frac{\pi}{1-\pi} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3,$$

where π is the probability of graduating within five years of admission, x_1 and x_2 are dummy variables for campus, and x_3 is High School Grade Point Average.

(a) (3 points) The table below shows how the dummy variables are defined. Write the odds of graduating within 5 years for each campus.

	x_1	x_2	Odds of Graduating
UTM	1	0	$e^{\beta_0+\beta_1+\beta_3 x_3}$
UTSC	0	1	$e^{\beta_0+\beta_2+\beta_3 x_3}$
St. George	0	0	$e^{\beta_0+\beta_3 x_3}$

(b) (1 point) Controlling for High School Grade Point Average, the odds of graduating within five years are ___ times as great for sudents on the UTM campus, compared to students on the UTSC campus. Write the answer in the space below in terms of β quantities.

$$\frac{\cancel{e^{\beta_0}}\, e^{\beta_1}\, \cancel{e^{\beta_3 x_3}}}{\cancel{e^{\beta_0}}\, e^{\beta_2}\, \cancel{e^{\beta_3 x_3}}} = e^{\beta_1 - \beta_2}$$

(c) (2 points) Suppose you concluded $\beta_2 < 0$. How would you express this in plain, non-statistical language? Use the word "chances" instead of "odds" or "probability," and begin with "Allowing for High School marks ..."

Allowing for High school marks, the chances of graduating within 5 years are less for students on the UTSC campus than for students on the St. George campus

2. In your analysis of the Bird-keeping data data, you fit a model in which the response variable was whether they got lung cancer (1=Yes, 0=No), and the explanatory variables were Gender (0=M, 1=F), Socioeconomic Status (0=Low, 1=High), Whether they are birdkeepers (1=Yes, 0=No) Age, How many years they have been smoking (including zero), and Cigarettes per day. Please base your answers on this full model.

   (a) (2 Points) Controlling for all the other variables in the model, being a bird-keeper multiplies the estimated odds of cancer by ...? Write the number in the space below.

$$e^{1.36259} = 3.9$$

   (b) (2 Points) Estimate the probability of lung cancer for a 30 year old male of low socioeconomic status who does not smoke and is not a bird-keeper. The answer is a number. Show a little work.

$$\hat{\pi} = \frac{e^{-1.93736 - (0.03976)30}}{1 + e^{-1.93736 - 1.1928}}$$

$$= \frac{0.0437}{1.0437} = 0.042$$

Please attach your R printout. You don't need to write anything on the printout this time eexcept your name and student number.

Quiz 9 Printout

```
> bird =
read.table("http://www.utstat.toronto.edu/~brunner/appliedf14/code_n_data/h
w/birdlung.data")
> colnames(bird) = c("cancer", "sex", "highses", "birdkeeper", "age",
"yrsmoke", "ncigs")
> head(bird)
  cancer sex highses birdkeeper age yrsmoke ncigs
1      1   0       0          1  37      19    12
2      1   0       0          1  41      22    15
3      1   0       1          0  43      19    15
4      1   0       0          1  46      24    15
5      1   0       0          1  49      31    20
6      1   0       1          0  51      24    15
> fullmod = glm(cancer ~ sex + highses + birdkeeper + age + yrsmoke +
ncigs, family=binomial,data=bird)
> summary(fullmod)

Call:
glm(formula = cancer ~ sex + highses + birdkeeper + age + yrsmoke +
    ncigs, family = binomial, data = bird)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5642  -0.8333  -0.4605   0.9808   2.2460

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.93736    1.80425  -1.074 0.282924
sex          0.56127    0.53116   1.057 0.290653
highses      0.10545    0.46885   0.225 0.822050
birdkeeper   1.36259    0.41128   3.313 0.000923 ***
age         -0.03976    0.03548  -1.120 0.262503
yrsmoke      0.07287    0.02649   2.751 0.005940 **
ncigs        0.02602    0.02552   1.019 0.308055
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 187.14  on 146  degrees of freedom
Residual deviance: 154.20  on 140  degrees of freedom
AIC: 168.2
```

```
Number of Fisher Scoring iterations: 5

> redmod = update(fullmod, . ~ . - birdkeeper)
> anova(redmod,fullmod,test="Chisq")
Analysis of Deviance Table

Model 1: cancer ~ sex + highses + age + yrsmoke + ncigs
Model 2: cancer ~ sex + highses + birdkeeper + age + yrsmoke + ncigs
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       141     165.87
2       140     154.20  1    11.67 0.0006352 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Name Jenny

Student Number

STA 442/2101 f2016 Quiz 5

1. (6 points) You have already proved that $(\mathbf{y}-X\boldsymbol{\beta})^\top(\mathbf{y}-X\boldsymbol{\beta}) = \mathbf{e}^\top\mathbf{e}+(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})^\top(X^\top X)(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$. Dividing both sides by σ^2, show that $\mathbf{e}^\top\mathbf{e}/\sigma^2 \sim \chi^2(n-p)$. Start with the distribution of the left side.

$$\frac{(y-X\beta)^T(y-X\beta)}{\sigma^2} = \frac{e^Te}{\sigma^2} + \frac{(\hat\beta-\beta)^T(X^TX)(\hat\beta-\beta)}{\sigma^2}$$

$$\Rightarrow \underbrace{(y-X\beta)^T(\sigma^2 I_n)^{-1}(y-X\beta)}_{W} = \underbrace{\frac{e^Te}{\sigma^2}}_{W_1} + \underbrace{(\hat\beta-\beta)^T(\sigma^2(X^TX)^{-1})^{-1}(\hat\beta-\beta)}_{W_2}$$

$$W = W_1 + W_2$$

Because $y \sim N(X\beta, \sigma^2 I_n)$, $W \sim \chi^2(n)$

Because $\hat\beta \sim N(\beta, \sigma^2(X^TX)^{-1})$, $W_2 \sim \chi^2(p)$

Because $\hat\beta$ & e are independent, W_1 & W_2 are independent

So by the formula sheet, $W_1 \sim \chi^2(n-p)$

2. (4 points) This question refers to your analysis of the SAT data. Controlling for Math score, is Verbal score related to first-year grade point average?

   (a) Give the null hypothesis in symbols. Assume your variables are in the same order as in the data file.

   $H_0: \beta_1 = 0$

   (b) In the space below, write the value of the test statistic. The answer is a number from your printout.

   $t = 4.178$

   (c) In the space below, write t the p-value. The answer is a number from your printout.

   $p = 4.41e{-05}$ or anything equivalent

   (d) Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

   Yes

   (e) Are the results statistically significant? Answer Yes or No.

   Yes

   (f) In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

   Allowing for math score, students who get higher Verbal scores tend to get higher 1st-year GPA

   Accept "controlling"

Guidelines

- One point for (a)
- 1/2 point each for (b) & (c)
- One point for (d & e) together. They must both be right, and consistent with (c).
- One point for f. Must be consistent with (c d e). If any one of (c d e) is wrong, zero for (f) as well.

You do *NOT* need to attach your printout this time.

Name Jerry

Student Number

STA 442/2101 f2016 Quiz Five

1. (6 points) Assuming the general linear model with normal errors (see formula sheet), let $\mathbf{a}$ be a $p \times 1$ vector of constants.

   (a) What is the distribution of $\mathbf{a}^\top \widehat{\boldsymbol{\beta}}$? Your answer includes both the expected value and the variance.

   $$a^T \hat{\beta} \sim N(a^T \beta, \sigma^2 a^T (X^T X)^{-1} a)$$

   (b) Now standardize the difference (subtract off the mean and divide by the standard deviation) to obtain a standard normal.

   $$Z = \frac{a^T \hat{\beta} - a^T \beta}{\sqrt{\sigma^2 a^T (X^T X)^{-1} a}}$$

   (c) Divide by the square root of a well-chosen chi-squared random variable, divided by its degrees of freedom, and simplify. Call the result T.

   $$T = \frac{(a^T \hat{\beta} - a^T \beta) / \sqrt{\sigma^2 a^T (X^T X)^{-1} a}}{\sqrt{\frac{e^T e}{\sigma^2} / (n-p)}}$$

   $$= \frac{a^T \hat{\beta} - a^T \beta}{\sqrt{MSE \; a^T (X^T X)^{-1} a}}$$

   (d) How do you know numerator and denominator are independent?

   Because $\hat{\beta}$ & e are independent

2. (4 points) This question refers to your analysis of the SAT data. Controlling for Verbal score, is Math score related to first-year grade point average?

   (a) Give the null hypothesis in symbols. Assume your variables are in the same order as in the data file.

   $H_0: \beta_2 = 0$

   (b) In the space below, write the value of the test statistic. The answer is a number from your printout.

   $t = 1.636$

   (c) In the space below, write t the p-value. The answer is a number from your printout.

   $p = 0.103$

   (d) Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

   No

   (e) Are the results statistically significant? Answer Yes or No.

   No

   (f) In plain, non-statistical language, what do you conclude? The answer is something about test scores and grade point average.

   Allowing for verbal score, there is not enough evidence to conclude that score on the math test is related to first-year GPA

   (If they say math score and GPA are unrelated no marks off, but write "Don't accept H_0")

Guidelines
- One point for (a)
- 1/2 point each for (b) & (c)
- One point for d & e together. They must both be right, and consistent with (c)
- One point for (f). Must be consistent with (c d e). If any one of (c d e) is wrong, zero for (f) as well

You do *NOT* need to attach your printout this time.

Name Jerry

Student Number

STA 442/2101 f2016 Quiz 7

1. Suppose you fit (estimate the parameters of) a regression model, obtaining $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{y}}$ and $\mathbf{e}$. Call this the *first model*. Then as an experiment, you fit a second regression model, using $\widehat{\mathbf{y}}$ from the first model as the response variable, and exactly the same $\mathbf{X}$ matrix as the first model. Call this the *second model*. **The following questions are pretty easy and you have more room than you need. Don't over-think this.**

   (a) (2 points) What is $\widehat{\boldsymbol{\beta}}$ for the second model? Denote it by $\widehat{\boldsymbol{\beta}}_2$.

   $$\hat{\beta}_2 = (X^T X)^{-1} X^T \hat{y} = (X^T X)^{-1} X^T X \hat{\beta} = \hat{\beta}$$

   (b) (2 points) What is $\widehat{\mathbf{y}}$ for the second model? Denote it by $\widehat{\mathbf{y}}_2$.

   $$\hat{y}_2 = X \hat{\beta}_2 = X \beta = \hat{y}$$

   (c) (2 points) What is $\mathbf{e}$ for the second model? Denote it by $\mathbf{e}_2$.

   $$e_2 = \hat{y} - \hat{y}_2 = \hat{y} - \hat{y} = 0$$

Name Jerry

Student Number

STA 442/2101 f2016 Quiz 7

1. Suppose you fit (estimate the parameters of) a regression model, obtaining $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{y}}$ and $\mathbf{e}$. Call this the *first model.* Then as an experiment, you fit a second regression model, using $\mathbf{e}$ from the first model as the response variable, and exactly the same $\mathbf{X}$ matrix as the first model. Call this the *second model.* **The following questions are pretty easy and you have more room than you need. Don't over-think this.**

   (a) (2 points) What is $\boldsymbol{\beta}$ for the second model? Denote it by $\widehat{\boldsymbol{\beta}}_2$. Show some work and simplify.

   $$\hat{\beta}_2 = (X^TX)^{-1}\underbrace{X^Te}_{0} = 0$$

   no marks off for not showing this

   (b) (2 points) What is $\widehat{\mathbf{y}}$ for the second model? Denote it by $\widehat{\mathbf{y}}_2$. Show some work and simplify.

   $$\hat{y}_2 = X\hat{\beta}_2 = X0 = 0$$

   (c) (2 points) What is $\mathbf{e}$ for the second model? Denote it by $\mathbf{e}_2$. Show some work and simplify.

   $$e_2 = e - \hat{y}_2 = e - 0 = e$$

2. (4 points) In your analysis of the chick weights data, you were asked to conduct a test for differences among mean weights for the five feed types *excluding* horsebean.

   (a) First, write the null hypothesis in terms of μ values. Assume the feeds are in alphabetical order: **`casein horsebean linseed meatmeal soybean sunflower`**

1 pt. $H_0: \mu_1 = \mu_3 = \mu_4 = \mu_5 = \mu_6$

   (b) Write the value of the test statistic in the space below. **On your printout, circle the number and write "Test statistic for Question 2" beside it.**

9.318

1 pt

   (c) Write the p-value in the space below. The answer is a number from your printout.

$p = 0.0000514$

   (d) Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No. Your answer must be consistent with 2e.

Yes

1 pt

   (e) Is there evidence of a difference in expected chick weight for the feeds other than Horsebean? Answer Yes or No. Your answer must be consistent with 2d.

Yes

   (f) What proportion of the remaining variation does the effect explain? The answer is a number between zero and one.

1 pt 0.364

Please **fold your R printout into the quiz**, with your name on the quiz paper showing. Make sure your name and student number also appear on your printout. *If you do not have your R printout, do not answer Question 2.*

Name Jerry

Student Number

STA 442/2101 f2016 Quiz 8

(10 points) One version of the delta method says that if $X_1, \ldots, X_n$ are a random sample from a distribution with mean μ and variance σ^2, and $g(x)$ is a function whose derivative is continuous in a neighbourhood of $x = \mu$, then $\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \stackrel{d}{\rightarrow} T \sim N(0, g'(\mu)^2\sigma^2)$. In many applications, both μ and σ^2 are functions of some parameter θ.

Let $X_1, \ldots, X_n$ be a random sample from an exponential distribution with parameter θ, so that $E(X_i) = \theta$ and $Var(X_i) = \theta^2$. Find a function $g(x)$ such that the limiting distribution of $Z_n = \sqrt{n}\left(g(\overline{X}_n) - g(\theta)\right)$ is *standard* normal — that is $Z_n \stackrel{d}{\rightarrow} Z \sim N(0, 1)$. Show your work. Finish your answer with the words "The function is ..." Write the function and **circle it**.

$$g'(\theta)^2 \cdot \theta^2 = 1 \iff g'(\theta)^2 = \frac{1}{\theta^2}$$

$$\Rightarrow g'(\theta) = \frac{1}{\theta}$$ (Just assume $g'(\theta) > 0$. No need to mention it.)

That is, $\frac{dg}{d\theta} = \frac{1}{\theta} \Rightarrow dg = \frac{1}{\theta}\, d\theta$

$$\Rightarrow \int dg = \int \frac{1}{\theta}\, d\theta = \ln(\theta) + c$$

let it $= 0$

So the function is $g(\theta) = \ln(\theta)$

They don't have to use separation of variables. They can even guess if they guess right.

Name Jerry

Student Number /

STA 442/2101 f2016 Quiz 8

(10 points) One version of the delta method says that if $X_1, \ldots, X_n$ are a random sample from a distribution with mean μ and variance σ^2, and $g(x)$ is a function whose derivative is continuous in a neighbourhood of $x = \mu$, then $\sqrt{n}\left(g(\overline{X}_n) - g(\mu)\right) \stackrel{d}{\rightarrow} T \sim N(0, g'(\mu)^2\sigma^2)$. In many applications, both μ and σ^2 are functions of some parameter θ.

Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with parameter λ, so that $E(X_i) = \lambda$ and $Var(X_i) = \lambda$. Find a function $g(x)$ such that the limiting distribution of $Z_n = \sqrt{n}\left(g(\overline{X}_n) - g(\lambda)\right)$ is *standard* normal — that is $Z_n \stackrel{d}{\rightarrow} Z \sim N(0,1)$. Show your work. Finish your answer with the words "The function is ..." Write the function and **circle it**.

$$g'(\lambda)^2 \cdot \lambda = 1 \Rightarrow g'(\lambda)^2 = \frac{1}{\lambda}$$

$$\Rightarrow g'(\lambda) = \lambda^{-\frac{1}{2}}$$ (Just assume $g'(\lambda) > 0$ No need to mention it.)

That is $\frac{dg}{d\lambda} = \lambda^{-\frac{1}{2}} \Rightarrow dg = \lambda^{-\frac{1}{2}} d\lambda$

$$\Rightarrow \int dg = \int \lambda^{-\frac{1}{2}} d\lambda = \frac{\lambda^{1/2}}{1/2} + c = 2\sqrt{\lambda}$$

Let it $= 0$

So The function is $g(\lambda) = 2\sqrt{\lambda}$

They don't have to use separation of variables. They can even guess if they guess right.

Name Jerry

Student Number /

STA 442/2101 f2016 Quiz 9

1. Consider a two-factor analysis of variance in which each factor has two levels. Use this regression model:

$$Y_i = \beta_0 + \beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 d_{i,1} d_{i,2} + \epsilon_i,$$

where $d_{i,1}$ and $d_{i,2}$ are zero-one indicator dummy variables for factors one and two respectively.

   (a) (2 points) In each cell of the table below, write the expected response in terms of β_j values. Naturally the *symbols* for the dummy variables should not be visible in your answer. They are either zero or one.

zero or 2

	$d_{i,2} = 0$	$d_{i,2} = 1$	
$d_{i,1} = 0$	β_0	$\beta_0 + \beta_2$	$\beta_0 + \beta_2/2$
$d_{i,1} = 1$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$	$\beta_0 + \beta_1 + \beta_2/2 + \beta_3/2$

no marks if (a) is wrong

   (b) (3 points) In terms of β_j values, what is the null hypothesis you would use to test for the main effect of Factor One?

$\cancel{\beta_0} + \cancel{\beta_2/2} = \cancel{\beta_0} + \beta_1 + \cancel{\beta_2/2} + \beta_3/2 \Rightarrow \beta_1 + \beta_3/2 = 0$

or $2\beta_1 + \beta_3 = 0$

2. (5 points) In your analysis of the sales data, you tested for whether the three software packages differ in their effectiveness for sales representatives with average (sample mean = 76.56) sales performance last quarter.

   (a) Write the value of the F statistic for the ordinary F-test (not the randomization test) in the space below. **On your printout, circle the number and write "Test statistic for Question 2" beside it.**

2 pts $F = 0.6122$

   (b) Write the p-value in the space below. The answer is a number from your printout.

1 pt $p = 0.5488$

2 pts

   (c) Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No. Your answer must be consistent with 2d.

No

   (d) Is there evidence of a difference? Answer Yes or No. Your answer must be consistent with 2c.

No

Name Jenny

Student Number

STA 442/2101 f2016 Quiz Nine

1. Consider a two-factor analysis of variance in which each factor has two levels. Use this regression model:

$$Y_i = \beta_0 + \beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 d_{i,1}d_{i,2} + \epsilon_i,$$

where $d_{i,1}$ and $d_{i,2}$ are zero-one indicator dummy variables for factors one and two respectively.

(a) (2 points) In each cell of the table below, write the expected response in terms of β_j values. Naturally the *symbols* for the dummy variables should not be visible in your answer. They are either zero or one.

zero or 2

	$d_{i,2} = 0$	$d_{i,2} = 1$
$d_{i,1} = 0$	β_0	$\beta_0 + \beta_2$
$d_{i,1} = 1$	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$
	$\beta_0 + \frac{1}{2}\beta_1$	$\beta_0 + \beta_2 + \frac{1}{2}(\beta_1 + \beta_3)$

No marks if (a) is wrong

(b) (3 points) In terms of β_j values, what is the null hypothesis you would use to test for the main effect of Factor Two?

$H_0: \beta_2 + \frac{1}{2}\beta_3 = 0$ or $2\beta_2 + \beta_3 = 0$

2. (5 points) In your analysis of the sales data, you tested for whether the three software packages differ in their effectiveness for sales representatives with average (sample mean = 76.56) sales performance last quarter.

2 pts

(a) Write the value of the F statistic for the ordinary F-test (not the randomization test) in the space below. **On your printout, circle the number and write "Test statistic for Question 2" beside it.**

$F = 0.6122$

1 pt

(b) Write the p-value in the space below. The answer is a number from your printout.

$p = 0.5488$

2 pts

(c) Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No. Your answer must be consistent with 2d.

No

(d) Is there evidence of a difference? Answer Yes or No. Your answer must be consistent with 2c.

No

	Low	Med	High
Vit C	β_1	β_2	β_3
OJ	β_4	β_5	β_6

not required

Name Jerry

Student Number

STA 442/2101 F 2017 Quiz Ten

In the Tooth Growth data, guinea pigs were randomly assigned to a combination of Supplement Type (Vitamin C or orange juice), and Dosage Level (Low, Medium, High). The response variable was tooth length after a certain length of time.

1. (1 point) Write the regression equation ~~with an E~~ for a regression model with *cell means coding*. That's the model with no intercept, and one indicator dummy variable for each treatment combination. You don't have to say how the dummy variables are defined. That will become clear in the next part. Just give the regression equation.

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \epsilon$$

Writing just $E(y)$ is OKAY.

2. (3 points) Make a table with six rows, one for each treatment combination. Make columns indicating how your dummy variables are defined. Make one more column containing $E(y)$ in terms of the β values from your regression model.

Supplement	Dose	x_1	x_2	x_3	x_4	x_5	x_6	$E(y)$
Vit C	Low	1	0	0	0	0	0	β_1
Vit C	Med	0	1	0	0	0	0	β_2
Vit C	High	0	0	1	0	0	0	β_3
OJ	Low	0	0	0	1	0	0	β_4
OJ	Med	0	0	0	0	1	0	β_5
OJ	high	0	0	0	0	0	1	β_6

3. (2 points) Suppose you want to test whether, averaging over Dosage Level, Supplement Type has an effect on average tooth growth. Give the null hypothesis in terms of β values from your regression equation.

$$H_0: \beta_1 + \beta_2 + \beta_3 = \beta_4 + \beta_5 + \beta_6$$

4. (2 points) Suppose you want to test whether the effect of Dosage Level depends on Supplement Type. Give the null hypothesis in terms of β values from your regression equation.

$$H_0 \; \beta_1 - \beta_4 = \beta_2 - \beta_5 = \beta_3 - \beta_6$$

5. (2 points) Suppose you want to test whether, averaging over Supplement Type, Dosage Level has an effect on average tooth growth. Give the null hypothesis in terms of β values from your regression equation.

$$H_0: \beta_1 + \beta_4 = \beta_2 + \beta_5 = \beta_3 + \beta_6$$

Name ____________________

Student Number ____________________

STA 442/2101 f 2017 Quiz Ten

In the Tooth Growth data, guinea pigs were randomly assigned to a combination of Supplement Type (Vitamin C or orange juice), and Dosage Level (Low, Medium, High). The response variable was tooth length after a certain length of time.

1. (1 point) Write the regression equation for a regression model with *effect coding*. That's the model with an intercept, and zeros, ones and minus ones. You don't have to say how the dummy variables are defined. That will become clear in the next part. Just give the regression equation.

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon$$

Writing just $E(y)$ is okay

2. (3 points) Make a table with six rows, one for each treatment combination. Make columns showing how your dummy variables are defined. You are *not* being asked for the expected values. They are too messy.

Supplement	Dose	x_1	x_2	x_3	
Vit C	Low	1	1	0	
Vit C	Med	1	0	1	
Vit C	High	1	-1	-1	
OJ	Low	-1	1	0	
OJ	Med	-1	0	1	
OJ	High	-1	-1	-1	

3. (2 points) Suppose you want to test whether, averaging over Dosage Level, Supplement Type has an effect on average tooth growth. Give the null hypothesis in terms of β values from your regression equation.

$H_0: \beta_1 = 0$

4. (2 points) Suppose you want to test whether the effect of Dosage Level depends on Supplement Type. Give the null hypothesis in terms of β values from your regression equation.

$H_0: \beta_4 = \beta_5 = 0$

5. (2 points) Suppose you want to test whether, averaging over Supplement Type, Dosage Level has an effect on average tooth growth. Give the null hypothesis in terms of β values from your regression equation.

$H_0: \beta_2 = \beta_3 = 0$

Name Jerry

Student Number

STA 442/2101 F 2017 Quiz Nine

1. (4 points) A Poisson regression model has an intercept and two explanatory variables. To obtain a confidence interval for the expected response at particular values x_1 and x_2, you plan to use the multivariate delta method.

   (a) What is the function $g(\beta_0, \beta_1, \beta_2)$ you seek to estimate?

$$g(\beta_0, \beta_1, \beta_2) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}$$

   (b) What is $\dot{g}(\beta_0, \beta_1, \beta_2)$? ~~Show a little work and~~ *circle your ~~final~~ answer.*

$$\dot{g}(\beta_0, \beta_1, \beta_2) = \left(\frac{\partial g}{\partial \beta_0}, \frac{\partial g}{\partial \beta_1}, \frac{\partial g}{\partial \beta_2} \right)$$

$$= \left(e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}, x_1 e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2}, x_2 e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} \right)$$

$$= e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2} (1, x_1, x_2)$$

2. (6 points) In your analysis of the Heart data (homework Question 3), you fit a model with just age, cholesterol level, and family history of heart disease as the explanatory variables. You were asked to test each variable controlling for the other two. Write the G^2 statistics and p-values in the spaces below. **Circle the numbers on your printout.**

(a) Test of age controlling for cholesterol level and family history of heart disease.

G^2	p-value
5.87	0.053

(b) Test of cholesterol level controlling for age and family history of heart disease.

G^2	p-value
4.67	0.097

(c) Test of family history of heart disease controlling for age and cholesterol level.

G^2	p-value
3.58	0.167

Please fold the printout for the Heart data into your quiz paper.

```
> # Heart attack
>
> rm(list=ls()); options(scipen=999) # To avoid scientific notation
> # install.packages("mlogit", dependencies=TRUE) # Only need to do this once
> library(mlogit) # Load the package every time
Loading required package: Formula
Loading required package: maxLik
Loading required package: miscTools

Please cite the 'maxLik' package as:
Henningsen, Arne and Toomet, Ott (2011). maxLik: A package for maximum likelihood estimation
in R. Computational Statistics 26(3), 443-458. DOI 10.1007/s00180-010-0217-1.

If you have questions, suggestions, or comments regarding the 'maxLik' package, please use a
forum or 'tracker' at maxLik's R-Forge site:
https://r-forge.r-project.org/projects/maxlik/
> heart = read.table("http://www.utstat.toronto.edu/~brunner/data/illegal/attack.data.txt")
> # heart = read.table("attack.data.txt") # Local copy
> head(heart)
  age diastol cholest ncigs height weight famhist   school   outcome
1  40      70     321     0   68.8    190     Yes  PostSec   Alive10
2  49      87     246    60   72.2    204      No       HS   Alive10
3  43      89     262     0   69.0    162      No       HS DiedFirst
4  50     105     275    15   62.5    152     Yes GradeSch   Alive10
5  50      88     261    30   68.0    142      No GradeSch    Dead10
6  47      79     372    30   67.0    193      No       HS   Alive10
> # Compute BMI
> height = heart$height; weight = heart$weight
> bmi = 703 * weight/height^2
> # Make a data frame with just the variables in the model
> datta = data.frame(heart[,1:4],bmi,heart[7:9])
> head(datta)
  age diastol cholest ncigs      bmi famhist   school   outcome
1  40      70     321     0 28.21838     Yes  PostSec   Alive10
2  49      87     246    60 27.51130      No       HS   Alive10
3  43      89     262     0 23.92060      No       HS DiedFirst
4  50     105     275    15 27.35514     Yes GradeSch   Alive10
5  50      88     261    30 21.58867      No GradeSch    Dead10
6  47      79     372    30 30.22477      No       HS   Alive10
> # Make an mlogit data frame in long format
> long1 = mlogit.data(datta,shape="wide",choice="outcome")
> # Fit the full model
> fullmod1 = mlogit(outcome ~ 0 | age+diastol+cholest+ncigs+bmi+famhist+school, data=long1)
> summary(fullmod1)

Call:
mlogit(formula = outcome ~ 0 | age + diastol + cholest + ncigs +
    bmi + famhist + school, data = long1, method = "nr", print.level = 0)

Frequencies of alternatives:
  Alive10    Dead10 DiedFirst
  0.53846   0.12500   0.33654

nr method
6 iterations, 0h:0m:0s
g'(-H)^-1g = 5.45E-06
successive function values within tolerance limits

Coefficients :
                           Estimate  Std. Error t-value Pr(>|t|)
Dead10:(intercept)     -19.09507993  6.86814066 -2.7802 0.005432 **
DiedFirst:(intercept)   -4.83994919  4.26471178 -1.1349 0.256424
Dead10:age               0.21938523  0.10023356  2.1887 0.028616 *
DiedFirst:age            0.12209406  0.06436732  1.8968 0.057850 .
Dead10:diastol           0.04015460  0.02786461  1.4411 0.149568
DiedFirst:diastol        0.02576623  0.02057755  1.2522 0.210514
Dead10:cholest          -0.00085502  0.00569787 -0.1501 0.880718
DiedFirst:cholest       -0.00884484  0.00434494 -2.0357 0.041784 *
```

```
Dead10:ncigs                0.01971071   0.02674279  0.7370 0.461093
DiedFirst:ncigs             0.00573135   0.02001375  0.2864 0.774594
Dead10:bmi                  0.07439465   0.11514935  0.6461 0.518233
DiedFirst:bmi              -0.05419410   0.08485427 -0.6387 0.523036
Dead10:famhistYes          -0.40723225   0.69927500 -0.5824 0.560322
DiedFirst:famhistYes       -1.04692473   0.52841978 -1.9812 0.047565 *
Dead10:schoolHS             1.40738650   1.17731847  1.1954 0.231924
DiedFirst:schoolHS          0.12318010   0.58897076  0.2091 0.834335
Dead10:schoolPostSec        2.14631341   1.22972923  1.7454 0.080923 .
DiedFirst:schoolPostSec     0.34817071   0.69039974  0.5043 0.614048
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -88.589
McFadden R^2:  0.11247
Likelihood ratio test : chisq = 22.452 (p.value = 0.12917)
>
> # Smaller model with just age, cholesterol level and family history
> # Make data frame
> long2 = mlogit.data(datta[,c(1,3,6,8)],shape="wide",choice="outcome")
> head(long2)
            age cholest famhist outcome chid       alt
1.Alive10    40     321     Yes    TRUE    1   Alive10
1.Dead10     40     321     Yes   FALSE    1    Dead10
1.DiedFirst  40     321     Yes   FALSE    1 DiedFirst
2.Alive10    49     246      No    TRUE    2   Alive10
2.Dead10     49     246      No   FALSE    2    Dead10
2.DiedFirst  49     246      No   FALSE    2 DiedFirst
> fullmod2 = mlogit(outcome ~ 0 | age+cholest+famhist, data=long2)
> summary(fullmod2)

Call:
mlogit(formula = outcome ~ 0 | age + cholest + famhist, data = long2,
    method = "nr", print.level = 0)

Frequencies of alternatives:
  Alive10    Dead10 DiedFirst
  0.53846   0.12500   0.33654

nr method
5 iterations, 0h:0m:0s
g'(-H)^-1g = 5.32E-05
successive function values within tolerance limits

Coefficients :
                        Estimate Std. Error t-value Pr(>|t|)
Dead10:(intercept)    -9.2358517  4.5775569 -2.0176  0.04363 *
DiedFirst:(intercept) -3.1529671  3.0567050 -1.0315  0.30231
Dead10:age             0.1634238  0.0895093  1.8258  0.06788 .
DiedFirst:age          0.1077304  0.0595888  1.8079  0.07062 .
Dead10:cholest        -0.0004808  0.0051048 -0.0942  0.92496
DiedFirst:cholest     -0.0085263  0.0042825 -1.9910  0.04649 *
Dead10:famhistYes     -0.2413608  0.6552984 -0.3683  0.71263
DiedFirst:famhistYes  -0.9337608  0.5108499 -1.8279  0.06757 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Log-Likelihood: -93.041
McFadden R^2:  0.067872
Likelihood ratio test : chisq = 13.549 (p.value = 0.035095)
>
> # Test age controlling for cholest and famhist
> no_age = mlogit(outcome ~ 0 | cholest+famhist, data=long2) # It can ignore vars, good
> G2_age =  -2 * as.numeric(no_age$logLik - fullmod2$logLik)
> pval_age = 1-pchisq(G2_age, df=2) # 2 betas
> c(G2_age,pval_age)
[1] 5.865214 0.053258
>
```

```
> # Test cholest controlling for age and famhist
> no_cholest = mlogit(outcome ~ 0 | age+famhist, data=long2)
> G2_cholest =  -2 * as.numeric(no_cholest$logLik - fullmod2$logLik)
> pval_cholest = 1-pchisq(G2_cholest, df=2) # 2 betas
> c(G2_cholest,pval_cholest)
[1] 4.67362108 0.09663536
>
> # Test famhist controlling for age and cholest
> no_famhist = mlogit(outcome ~ 0 | age+cholest, data=long2)
> G2_famhist =  -2 * as.numeric(no_famhist$logLik - fullmod2$logLik)
> pval_famhist = 1-pchisq(G2_famhist, df=2) # 2 betas
> c(G2_famhist,pval_famhist)
[1] 3.5831202 0.1666999
>
```

Name Jerry

Student Number ______________________

STA 442/2101 F 2017 Quiz Eight

Please do *not* turn any printouts in this time.

1. For your analysis of the bird keeping data (Question 4),

   (a) (2 points) ~~All else being equal~~ controlling for other variables, the estimated odds of cancer are __ as great for women. Answer the question without considering statistical significance. The answer is a number. Show a little work (but no proof) and answer the question in the space below. **Circle your answer.**

   $$e^{0.56127} = 1.75$$

   (b) (2 points) The main question in this study is whether, all else being equal, keeping birds is associated with an increased risk of cancer. In plain, non-statistical language including *no numbers*, answer the question in the space below. You have more room than you need.

   Keeping birds is associated with an increased risk of cancer.

2. (6 points) Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta X_i + \epsilon_i,$$

where $X_i \sim N(0, \sigma_x^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. Because of omitted variables that influence both X_i and Y_i, we have $Cov(X_i, \epsilon_i) = c \neq 0$.

Let $\widehat{\beta}_n = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$. Is $\widehat{\beta}_n$ a consistent estimator of β? Prove your answer. After the calculations, write "Yes, consistent" or "No, not consistent.

$$\begin{aligned} E(X_i Y_i) &= E\left(X_i(\beta X_i + \varepsilon_i)\right) \\ &= E(\beta X_i^2 + X_i \varepsilon_i) \\ &= \beta E(X_i^2) + E(X_i \varepsilon_i) \\ &= \beta \sigma_x^2 + c \end{aligned}$$

$$E(X_i^2) = \sigma_x^2$$

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n X_i Y_i}{\frac{1}{n}\sum_{i=1}^n X_i^2} \xrightarrow{as} \frac{E(X_i Y_i)}{E(X_i^2)}$$

By LLN and continuous mapping

$$= \frac{\beta \sigma_x^2 + c}{\sigma_x^2} = \beta + \frac{c}{\sigma_x^2} \neq \beta$$

No, not consistent

Name Jerry

Student Number ______________

STA 442/2101 F 2017 Quiz eight

Please do *not* turn any printouts in this time.

1. For your analysis of the bird keeping data (Question 4),

   (a) (2 points) ~~All else being equal~~ Controlling for other variables, the estimated odds of cancer are __ as great for participants of high socioeconomic status. Answer the question without considering statistical significance. The answer is a number. Show a little work (but no proof) and answer the question in the space below. **Circle your answer.**

   $$e^{0.10545} = 1.11$$

   (b) (2 points) The main question in this study is whether, all else being equal, keeping birds is associated with an increased risk of cancer. In plain, non-statistical language including *no numbers*, answer the question in the space below. You have more room than you need.

   Keeping birds is associated with an increased risk of cancer.

2. (6 points) Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta X_i + \epsilon_i \\ W_i &= X_i + \delta_i \end{aligned}$$

where all random variables are normal with expected value zero, $Var(X_i) = \cancel{\phi}\ \sigma_x^2 > 0$, $Var(\epsilon_i) = \sigma_\epsilon^2 > 0$, $Var(\delta_i) = \sigma_\delta^2 > 0$, and ϵ_i, δ_i and X_i are all independent. The variables W_i and Y_i are observable, while X_i is latent (unobservable).

Let $\widehat{\beta}_n = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2}$. Is $\widehat{\beta}_n$ a consistent estimator of β? Prove your answer. After the calculations, write "Yes, consistent" or "No, not consistent.

$$\begin{aligned} E(W_i Y_i) &= E(X_i + \delta_i)(\beta X_i + \epsilon_i) \\ &= \beta E(X_i^2) + 0 + 0 + 0 \\ &= \beta \sigma_x^2 \end{aligned}$$

$$E(W_i^2) = Var(W_i) = \sigma_x^2 + \sigma_\delta^2$$

$$\hat{\beta}_n = \frac{\frac{1}{n}\sum_{i=1}^n W_i Y_i}{\frac{1}{n}\sum_{i=1}^n W_i^2} \xrightarrow{as} \frac{E(W_i Y_i)}{E(W_i^2)}$$

By SLLN and continuous mapping

$$= \frac{\beta \sigma_x^2}{\sigma_x^2 + \sigma_\delta^2} \neq \beta$$

Unless $\beta = 0$ but they don't have to say it.

No, not consistent