First Year Qualifying Exam

Methods 210, 210B, 210C

Monday, June 21, 2021
9:00 am-12:00 pm

- There are 4 questions on the examination. You are to do 3 of 4 questions.
- Your solutions to each problem should be written on separate sheets of paper. Label each sheet with your student identification number, the problem number, and the page number of that solution written in the upper right hand corner. For example, the labeling on a page may be:

ID# 912346378
Problem 2, page 3

- You have 3 hours to complete your solution. Please be prepared to turn in your exam at 12:00pm.

1. Evolutionary biologists are often interested in the key characteristics that allow a species to survive against the pressure of evolution, and one key characteristic is brain size. While large brain size may be an advantage for a species, it also introduces certain penalties, such as the need for longer pregnancies and fewer offspring. In this example, we will examine how the size of brain is associated with different species characteristics.

   The dataset we will use here includes the average values of **brain weight (response variable), body weight, gestation length** (length of pregnancy), and **litter size** (the number of offspring produced at one birth by an animal) for **96 species of mammals**. Check the following table for the units of each variables and a small portion of the full data set:

Species	**Brain Weight**	Body Weight	Gestation Period	Litter Size
	(grams)	(kilograms)	(days)	
Dog	70.2	8.5	63	4.0
Domestic Cat	28.4	2.5	63	4.0
Human	1300	65	270	1.0
...	...	...	...	...

   The following linear regression model is used to study this data set:

$$\text{BRAIN} = \beta_0 + \beta_1 \text{ BODY} + \beta_2 \text{ GESTATION} + \beta_3 \text{ LITTER} + \varepsilon.$$

   Here is the output from R for fitting the above linear regression model (for simplicity, some numbers in the output were round off to two decimal digits)

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   -225.29     83.06    -2.71  0.00797 **
BODY             0.96      0.09    10.46  < 2e-16 ***
GESTATION        1.81      0.35     5.10 1.79e-06 ***
LITTER          27.65     17.41     1.59  0.11579
---
Residual standard error: 224.6 on 92 degrees of freedom
Multiple R-squared:  0.81,Adjusted R-squared:  0.8038
F-statistic:  --- on - and -- DF,  p-value: < 2.2e-16
```

   Answer the following questions. (*In order to receive full credit, include the formula/reasoning you use for obtaining the results*):

   (a) What are the fitted values and residuals of the brain weight for humans?

   (b) Interpret the meaning of $\hat{\beta}_2 = 1.81$ **in the context of the problem being addressed**.

   (c) Construct a 95% confidence interval for the regression coefficient associated with GESTATION.

   (d) Some output pertaining to the F-statistic is missing in the R output. Calculate the F-statistic value as well as the corresponding degrees of freedom. Also write down the null hypothesis of this F-test, and interpret the conclusion of this test in the context of the problem being addressed.

   (e) The ANOVA table of the regression performed above is show below:

```
Analysis of Variance Table
Response: BRAIN
          Df   Sum Sq  Mean Sq  F value    Pr(>F)
BODY       1 18228007 18228007 361.4682 < 2.2e-16 ***
GESTATION  1  1422101  1422101  28.2008 7.555e-07 ***
LITTER     1   127117   127117   2.5208    0.1158
Residuals 92  4639348    50428
```

Suppose we wish to conduct a general F-test to see if mean brain size is related to **either the period of gestation or the size of litter**, after **accounting for the effect of body weight**. Answer the following problems:

i. Specify the null hypothesis of this F-test.

ii. What is the extra amount of the variation of the response variable explained by adding variables GESTATION and LITTER to the model with only variable BODY?

iii. Calculate the F-test statistics.

(f) The variance inflation factor of the new model is listed as follows:

```
BODY     GESTATION  LITTER
2.4927    2.8874    1.5878
```

Based on this information, answer the following questions:

i. What can we say about the potential multicollinearity in the predictors?

ii. Suppose we perform a multiple linear regression using BODY as the response variable, GESTATION and LITTER as the predictor, what will be the value of R^2?

END OF QUESTION (1)

2. Cardiothoracic surgery is a major medical procedure that carries the possibility of increased risk of adverse events in patients. A major risk factor associated with poor outcomes in surgery patients is the total time of surgery until skin closure. As such, it is of interest to investigate factors associated with increased surgical time. In this question, we will consider data from $N = 156$ cardiothoracic surgery patients seen at a Texas hospital. We will focus on estimating the association between age and total surgical time.

   (a) We will start by considering a regression model of the following form:

$$Y_i = \beta_0 + \beta_1 \texttt{age.c}_i + \epsilon_i, \quad i = 1, \dots, 156. \tag{1}$$

   with Y_i denotes the total surgical time for patient i and $\texttt{age.c}_i$ denotes the age of the patient in years **and has been centered at age 65 (i.e. 65 has been subtracted from each patient's age)**. Provide precise interpretations of β_0 and β_1.

   (b) Let $\widehat{\beta}_1$ denote the ordinary least squares estimator (OLS) of β_1. What assumptions regarding ϵ_i, $i = 1, \dots, 156$, are required for $\widehat{\beta}_1$ to be unbiased of β_1? Justify your answer.

   (c) Output of the OLS estimates after fitting model (1) to the data are given below. Based upon this model, provide a 95% CI for the expected difference in mean surgical times comparing two subpopulations of participants differing in age by 10 years. Note that elements of the output have been purposefully omitted, but sufficient information is provided to answer the problem. To save time, in all parts using numeric output you may leave your answer unevaluated.

```
> fit1 <- lm( surgtime ~ age.c, data=surgery )
> summary(fit1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   63.642     13.535     4.7  5.7e-06 ***
age.c          0.946     ------     4.1  6.7e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 35.6 on 154 degrees of freedom
Multiple R-squared:  0.0984,Adjusted R-squared:  0.0926
F-statistic: 16.8 on 1 and 154 DF,  p-value: 6.65e-05
```

   (d) It is thought that surgical procedure type (in this case an aortic procedure vs. a non-aortic procedure) may confound the relationship between age and surgical time. Based upon the summary fits below, is there empirical evidence that this may be true? Explain.

```
> fit2a <- lm( surgtime ~ I(procedure=="Aortic"), data=surgery )
> summary(fit2a)

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    116.38       2.99    38.9   <2e-16 ***
I(procedure == "Aortic")TRUE    39.62      15.27     2.6     0.01 *


> fit2b <- lm( age.c ~ I(procedure=="Aortic"), data=surgery )
> summary(fit2b)

Coefficients:
```

```
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    56.994      0.963    59.2   <2e-16 ***
I(procedure == "Aortic")TRUE   12.006      5.003     2.4    0.018 *
```

(e) Consider the following model that adjusts for an indicator of aortic procedure:

$$Y_i = \gamma_0 + \gamma_1 \texttt{age.c}_i + \gamma_2 I_{\texttt{Aortic},i} + \epsilon_i, \quad i = 1, \ldots, 156. \tag{2}$$

Let $\widehat{\gamma}_1$ denote the OLS estimate of γ_1 and let β_1 denote the coefficient associated with age in model (1). From the regression output given in Part (d), which of the following will be true:

i. $E[\widehat{\gamma}_1] = \beta_1$
ii. $E[\widehat{\gamma}_1] < \beta_1$
iii. $E[\widehat{\gamma}_1] > \beta_1$

Completely justify your answer. Given your answer, discuss the need for considering adjustment for confounding factors when estimating the associating between surgical time and age.

(f) Sex has previously been shown to be related to surgical time. It has further been hypothesized that the association between mean surgical time and age may differ by sex. As such, an interaction is included in the model as follows:

$$Y_i = \delta_0 + \delta_1 \texttt{age.c}_i + \delta_2 I_{\texttt{Aortic},i} + \delta_3 \texttt{age.c}_i \times I_{\texttt{female},i} + \epsilon_i, \quad i = 1, \ldots, 156, \tag{3}$$

where $I_{\texttt{female},i}$ is an indicator of female sex for patient i. Note that no main effect for sex is included in the model. Based upon this model, what is the estimated association between sex and mean surgical time among individuals aged 65 years? What are the implications of your answer when deciding whether or not to include lower level terms when modeling interactions?

(g) Now consider this model:

$$Y_i = \zeta_0 + \zeta_1 \texttt{age.c}_i + \zeta_2 I_{\texttt{Aortic},i} + \zeta_3 \ I_{\texttt{female},i} + \zeta_4 \texttt{age.c}_i \times I_{\texttt{female},i} + \epsilon_i, \quad i = 1, \ldots, 156, \tag{4}$$

Based on the output below, carry out a hypothesis test to test the null hypothesis that the association between mean surgical time and age does not vary by sex. Your answer should clearly state your null and alternative hypothesis in terms of the model parameters, your resulting test statistic, the level of your test, and the conclusion you draw.

```
> fit4 <- lm( surgtime ~ age.c + I(procedure=="Aortic") + female + age.c:female, data=surgery )
> summary(fit4)

Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                    45.371     24.982    1.82   0.0713 .
age.c                           1.172      0.422    ----   ----
I(procedure == "Aortic")TRUE   30.459     15.297    1.99   0.0483 *
female                         31.094     30.035    ----   ----
age.c:female                   -0.435      0.512    ----   ----
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Multiple R-squared:  0.131,Adjusted R-squared:  0.107
F-statistic: 5.67 on 4 and 151 DF,  p-value: 0.000281
```

(h) From the above, what is the estimated association between mean surgical time and age among male patients? Among female patients?

(i) Below is the model root-MSE and a partially redacted estimate of Var$[\widehat{\vec{\zeta}}]$. Using this and the model estimates for `fit4`, provide a 95% CI for the association between mean surgical time and age among females.

```
> summary(fit4)$sigma
[1] 35.28

> vcov(fit4)
            (Intercept)     age "Aortic"   female   age:female
(Intercept)     624.120 -10.348  -25.722 -630.026     10.47414
age             -10.348   -----    0.368   10.433     -0.17999
"Aortic"        -25.722   0.368  234.002   79.454     -1.51185
female         -630.026  10.433   79.453   ------    -15.05835
age:female       10.474  -0.179   -1.511  -15.058      0.26167
```

(j) Based upon the above model output, provide a 95% prediction interval for the surgical time of a randomly sampled female patient aged 70 years that is going to undergo an aortic surgical procedure.

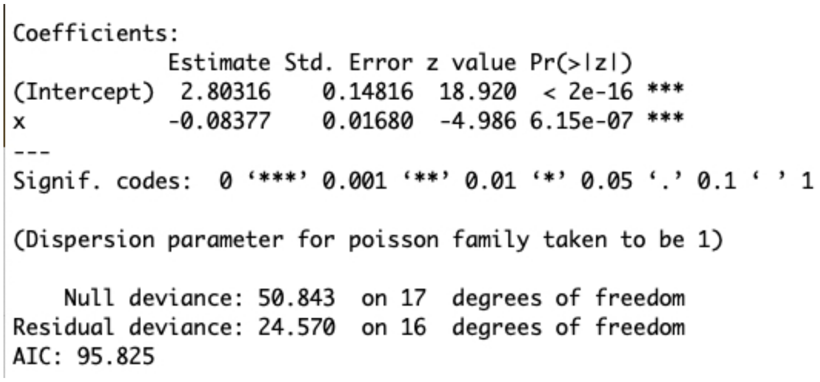
(k) Finally, you are asked to perform residual diagnostics to determine if age should be entered into the model as a linear term. Precisely define what residual you will consider, what plot you would produce to assess this, and give an example of the shape of the plot that would indicate that the linear term for age is acceptable.

END OF QUESTION (2)

3. Consider the following data: random households in Oakland, California, were selected and a single person was interviewed in each house. The respondents were asked a number of questions relating to a stressful event and were asked to say whether they had experienced the event, and when this occurred. Data were extracted on the 147 respondents who had exactly one event. The question of interest here is whether individuals can recollect events. Let Y_i denote the number of individuals who had a stressful event in month i, $i = 1, \cdots, 18$.

Month recalled *i*	1	2	3	4	5	6	7	8	9
No. of respondents Y_i	15	11	14	17	5	11	10	4	8
Month recalled *i*	10	11	12	13	14	15	16	17	18
No. of respondents Y_i	10	7	9	11	3	6	1	1	4

The R output given by the `glm` function used on the data above (x=Month recalled) is:



```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.80316    0.14816  18.920  < 2e-16 ***
x           -0.08377    0.01680  -4.986 6.15e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 50.843  on 17  degrees of freedom
Residual deviance: 24.570  on 16  degrees of freedom
AIC: 95.825
```

a) Write the equation of the model used to analyze the data.
b) Provide an interpretation of the parameters in the systematic component of the model.
c) Derive the form of the log-likelihood function.
d) Derive the form of the score function.
e) Derive the form of Fisher’s (expected) information matrix, evaluate it using the data given and confirm the standard errors reported in the output.
f) Derive the form of the log-likelihood ratio test statistic to test $H_0: \beta = 0$ vs. $H_1: \beta \neq 0$. Compute the value of the log-likelihood ratio test statistic using the data given and confirm the value of the test statistic with the one that you can derive using the R output.

END OF QUESTION (3)

4. A clinical trial was conducted to investigate the effectiveness of two new treatments for patients with onychomycosis, a disease of fungal infection of the nails with signs of nail discoloration, thickening, and brittle. If successfully treated, patients would grow new, clear and healthy nail. At Baseline, the study randomized patients to receive a 12-week treatment of either Drug A, Drug B, or a placebo. Clear nail length (mm) of one infected toenail was measured at Baseline, Week 6 and Week 12.

Let Y_{ij} represent the clear nail length at the j-th time (j = 1 for Baseline (Week 0), = 2 for Week 6, and = 3 for Week 12) for the i-th subject (i = 1, ..., N). Denote μ_{ij} = $E\left(Y_{ij}\right)$.

(A) Parametric Curves analysis

1) Treating Time as a continuous covariate,
   i. Write down a linear regression model statement for longitudinal data that includes main effects, Trt (Drug A, Drug B, and Placebo) and Time (Weeks 0, 6, and 12), and their interaction Trt-by-Time. Make Placebo as the referenced group for Trt.
   ii. Clearly define all covariates in the model.
   iii. Specify all appropriate model assumptions, including the distributional properties and an unstructured covariance assumption.

2) Give interpretation of each regression coefficient in the model:

3) Test the joint null hypothesis that changes in the mean clear nail length are not different among treatments by using the Wald test statistic $W^2 = \left(L\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{L}\widehat{\text{Cov}(\hat{\beta})}\boldsymbol{L}'\right)^{-1}\left(L\widehat{\boldsymbol{\beta}}\right)$.
   i. Write down the null hypothesis H_0.for the test.
   ii. Provide the $\boldsymbol{L}$ matrix.
   iii. How many degrees of freedom are there for the Wald test?

4) Compare the treatment effect in growing clear nail between Drug A and Drug B by using the Wald test statistic $W^2 = \left(L\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{L}\widehat{\text{Cov}(\hat{\beta})}\boldsymbol{L}'\right)^{-1}\left(L\widehat{\boldsymbol{\beta}}\right)$.
   i. Write down the null hypothesis H_0.for the test.
   ii. Provide the $\boldsymbol{L}$ matrix.
   iii. How many degrees of freedom are there for the Wald test?

(B) Generalized Estimating Equations (GEE) analysis

In the onychomycosis study, define a binary response of treatment success based on the clear nail length (mm) as follows:

$$R_{ij} = \begin{cases} 1, if\ Y_{ij} - Y_{i1} \geq 2\ mm \\ 0, if\ Y_{ij} - Y_{i1} < 2\ mm \end{cases}, for\ j = 2, 3$$

For the following questions, we will drop Drug B from consideration and the categorical Time variable has 2 occasions (2 and 3). Denote π_{ij} = $E\left(R_{ij}\right) = Prob\left(R_{ij} = 1\right)$.

1) Treating Time as a categorical covariate,

i. Write down a GEE model statement with logit link $\log\left\{\frac{Prob(R_{ij}=1)}{Prob(R_{ij}=0)}\right\}$ for longitudinal data that includes main effects, Trt (Drug A and Placebo) and Time (2, and 3), and their interaction Trt-by-Time. Make Placebo and Week 6 as the referenced groups for Trt and Time, respectively.
ii. Clearly define all covariates in the model.
iii. Specify all appropriate model assumptions, including the distributional properties and Working Covariance parameters.

2) Give interpretation of each regression coefficient in the model.

3) Test the null hypothesis that Drug A has no effect on changes in the odds of successful response by using the Wald test statistic $W^2 = (L\widehat{\boldsymbol{\beta}})' \left(\boldsymbol{L}\widehat{\mathrm{Cov}(\hat{\beta})}\boldsymbol{L}'\right)^{-1} (L\widehat{\boldsymbol{\beta}})$.
i. Write down the null hypothesis H_0.for the test.
ii. Provide the $\boldsymbol{L}$ matrix.
iii. How many degrees of freedom are there for the Wald test?

4) Give a point estimate and a 95% confidence interval for the odds ratio in probability of successful response between Drug A and Placebo at Week 12.

END OF QUESTION (4)