First Year Qualifying Exam

Methods 210, 211, 212

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** (grams)	Body Weight (kilograms)	Gestation Period (days)	Litter Size
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:

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

   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, $i = 1, \ldots, 156$, 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 $\widehat{\text{Var}}[\vec{\zeta}]$. 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. The following table contains data from a study to evaluate the risk factors (including Seat belt use, Gender, and Location) for car crash injuries. The response categories are (1) not injured, (2) injured but not transported by emergency medical services, (3) injured and transported by emergency medical services but not hospitalized, (4) injured and hospitalized and survived, (5) injured and died.

			Response				
Gender	Location	Seat belt	1	2	3	4	5
Female	Urban	No	7287	175	720	91	10
		Yes	11587	126	577	48	8
	Rural	No	3246	73	710	159	31
		Yes	6134	94	564	82	17
Male	Urban	No	10381	136	566	96	14
		Yes	10969	83	259	37	1
	Rural	No	6123	141	710	188	45
		Yes	6693	74	353	74	12

Let Y denote the response category; X denote gender, with

$$X = \begin{cases} 0 & \text{for Male,} \\ 1 & \text{for Female;} \end{cases}$$

Z denote the location, with

$$Z = \begin{cases} 0 & \text{for Urban,} \\ 1 & \text{for Rural;} \end{cases}$$

and U denote the seat belt use, with

$$U = \begin{cases} 0 & \text{for Wearing seat belt,} \\ 1 & \text{for No seat belt.} \end{cases}$$

A proportional odds model (cumulative logits) was fit to the data:

$$\text{logit}[\Pr(Y \leq j)] = \alpha_j + \beta_1 * X + \beta_2 * Z + \beta_3 * U + \beta_4 * Z * U.$$

Parameter estimates and standard errors as output from the `R` function `vglm` are given in the table below.

Coefficients:	Value	Std. Error
(Intercept):1	3.307	0.0351
(Intercept):2	3.482	0.0356
(Intercept):3	5.349	0.0470
(Intercept):4	7.256	0.0914
X	-0.546	0.0272
U	-0.760	0.0394
Z	-0.699	0.0424
Z*U	-0.124	0.0548

The correlation matrix of the estimated parameters is

	Intercept1	Intercept2	Intercept3	Intercept4	X	U	Z
Intercept2	0.987						
Intercept3	0.748	0.756					
Intercept4	0.384	0.388	0.510				
X	-0.497	-0.492	-0.380	-0.196			
U	-0.707	-0.699	-0.534	-0.274	0.073		
Z	-0.637	-0.629	-0.480	-0.247	0.026	0.558	
Z*U	0.472	0.465	0.348	0.178	0.021	-0.714	-0.773

(3a) For males in urban areas wearing seat belts, calculate the estimated probability of being injured and hospitalized and survived and provide a 95% CI for it.

(3b) Find the estimated gender effect (female v.s. male) on the cumulative odds ratio, given seat belt use and location. Interpret the point estimate and the corresponding 95% CI.

(3c) Calculate the estimated probability of no injury among females in rural areas who wear seat belts all the time, and calculate a 95% CI for it.

(3d) Find the estimated cumulative odds ratio between response and seat belt use for those in rural locations. Is this estimate different than the estimate from those in urban locations? Why? Justify your answer.

(3e) The key model assumption in the above fitting is the "proportional odds assumption". If one fits the above model allowing all coefficients to change over j, will that be reasonable? Why? How to test the proportional odds assumption?

END OF QUESTION (3)

4. An Internet company is interested in testing the efficacy of two different series of coordinated campaign ads. More specifically, the company is interested in understanding the likability of the campaigns over time. For that purpose, the company decides to conduct an experiment to compare the ratings of the two campaigns by randomly assigning a group of 30 volunteers to see either campaign A (also coded as 0 and "red") or campaign B (also coded as 1 and "blue") for a total of 8 weeks. Each volunteer will see advertisements only from the assigned group of ads for the entire 8 weeks. At the end of each week, the volunteers report their overall ratings on different aspects of the campaign, which are then summarized in a single (continuous) measurement. Figure 1 in the Appendix reports a plot with the individual profiles ("spaghetti plot") as well as the weekly means of the reported ratings from the individuals assigned to the two campaigns.

1. For the following questions, you may refer to Part 1 in the Appendix.

(a) Write the mathematical form of the assumed model (the assumed model, not the fitted model). Clearly state all the modeling assumptions with particular regard to the mean and covariance functions.

(b) Provide *precise* interpretations of each parameter and identify the parameter(s) of primary interest given the goals of the study.

(c) Provide the expression (in symbols) of the estimator $\hat{\beta}$ of the fixed effect coefficients. Discuss if such estimator is unbiased and characterize its asymptotic distribution.

(d) Provide the expression (in symbols) of the marginal variance implied by the model defined in (1.a). After writing down the expression in symbols, write down the estimated value of each variance parameter from the output of Part 1.

(e) Discuss maximum likelihood (ML) versus restricted maximum likelihood (REML) estimators of the variance parameters, in particular their unbiasedness and consistency properties.

2. Now consider the output reported in Part 2 of the Appendix.

(a) Write the mathematical form of the assumed model and highlight the differences with respect to model (1.a).

(b) Consider the test reported at the end of Part 2. Clearly specify the hypotheses being tested and the hypothesis testing approach. Discuss the appropriateness of the testing procedure used in this case.

3. Now refer to Part 3 of the Appendix.

(a) Clearly state all the assumptions of the model, in particular all the assumptions on the mean and variance-covariance structure. Identify the corresponding estimates in the reported output.

(b) Propose a test for assessing if the campaign B has different ratings than the campaign A. Clearly specify (and justify) the hypotheses being tested, the hypothesis testing approach and the relevant test statistic.

(c) Another data analyst suggests that an exchangeable correlation structure should provide inferences approximately equivalent to those of model `mod1`. Comment on this statement.

- Xu et al. (*Statistica Sinica*, 2012) propose a mix-GEE approach, where the working correlation matrix is represented by a combination of a finite number of correlation matrices, say $W_i(\boldsymbol{\alpha}) = \sum_{l=1}^{L} \pi_l W_i^{(l)}(\alpha_l)$, with $\boldsymbol{\alpha} = (\alpha_1, \ldots \alpha_L)$. Such a working correlation matrix is motivated by assuming that the data $\boldsymbol{y}_i = (y_{i1}, \ldots, y_{iT})^\tau$ are from a mixture of L independent components:

$$\boldsymbol{y}_i = z_{i1}\boldsymbol{y}_i^{(1)} + \cdots + z_{iL}\boldsymbol{y}_i^{(L)},$$

where $\boldsymbol{y}_i^{(1)}, \ldots, \boldsymbol{y}_i^{(L)}$ are the L independent components, and the $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iL})^\tau$ are latent indicators that have only one entry equal to one with the rest of its entries equal to zero. Let $\pi_l = \Pr(z_{il} = 1)$, $l = 1, \ldots, L$. Note that $\sum_{l=1}^{L} \pi_l = 1$.

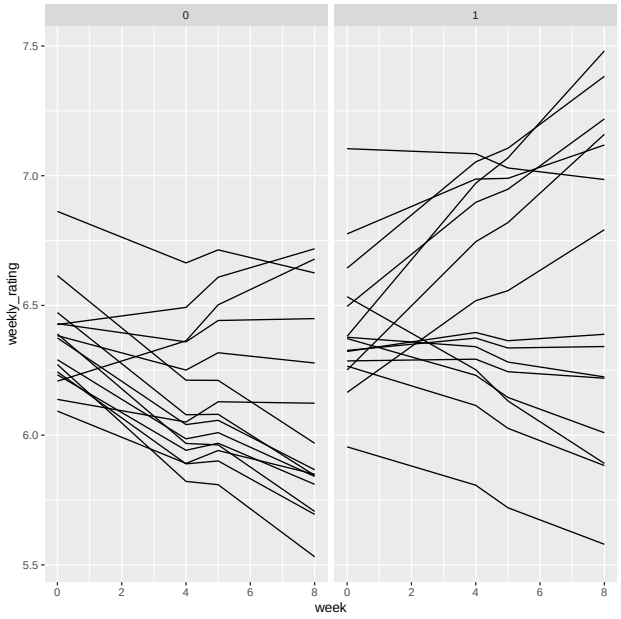
Assume a consistent estimator $\hat{\pi}_l$ and $\hat{\alpha}_l$, and that $W_i(\alpha)$ is positive semi-definite. Further, assume it is possible to approximate $W_i^{-1}(\boldsymbol{\alpha}) \approx \mathcal{H}_i(\boldsymbol{\alpha}) = \sum_l \tilde{\pi}_l \mathcal{H}_i^{(l)}(\boldsymbol{\alpha})$ for some positive semi-definite symmetric matrices $\mathcal{H}_i^{(l)}$ and weights $\tilde{\pi}_l$.

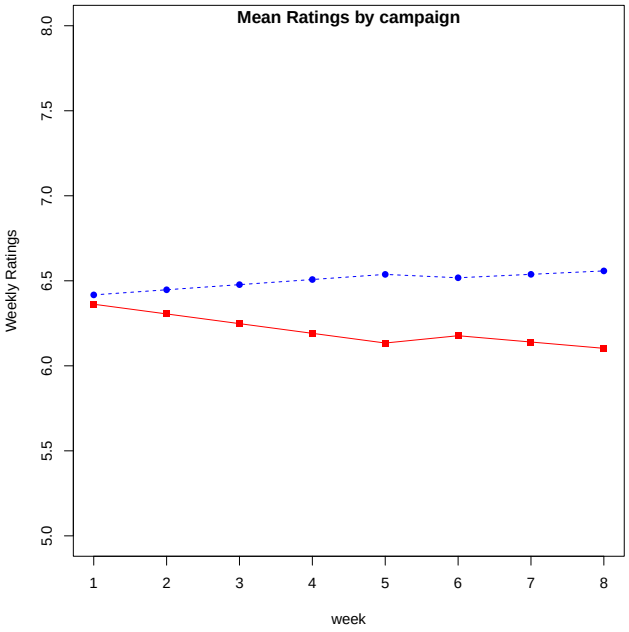
(d) Write down the first-order generalized estimating equation for the marginal model. Each term of the estimating equation should be fully defined.

(e) Provide the expression (in symbols) of the solution $\hat{\beta}$ of the generalized estimating equation using the proposed mixture of working correlation matrices with the assumptions above. Discuss if such estimator is unbiased and characterize its asymptotic distribution.

(f) With reference to the Gauss-Markov for correlated data, discuss under what conditions the estimator $\hat{\beta}$ above has minimal variance among all estimators that are linear in $\boldsymbol{y}$.

Appendix





Part 1

```
mod1=lme(weekly_rating ~ week*Ad_type, data=Ad_data, random = ~ 1 |Subject)
summary(mod1)
## Linear mixed-effects model fit by REML
##   Data: Ad_data
##         AIC       BIC   logLik
##   -97.09029 -75.58931 54.54515
##
## Random effects:
##  Formula: ~1 | Subject
##         (Intercept)  Residual
## StdDev:   0.3240245 0.1553001
##
## Fixed effects:  weekly_rating ~ week * Ad_type
##                  Value  Std.Error  DF  t-value p-value
## (Intercept)   6.326621 0.08721741 238 72.53851  0.0000
## week         -0.033710 0.00517667 238 -6.51183  0.0000
## Ad_type       0.108235 0.12334404  28  0.87751  0.3877
## week:Ad_type  0.052193 0.00732092 238  7.12935  0.0000
##  Correlation:
##              (Intr) week   Ad_typ
## week         -0.237
## Ad_type      -0.707  0.168
## week:Ad_type  0.168 -0.707 -0.237
##
## Standardized Within-Group Residuals:
##          Min           Q1          Med           Q3          Max
## -3.066271202 -0.509423443  0.003976672  0.536337546  3.062035585
##
## Number of Observations: 270
## Number of Groups: 30
```

Part 2

```
mod2=lme(weekly_rating ~ week*Ad_type, data=Ad_data,
    random = ~ week |Subject)
```

```
anova(mod1, mod2)
##      Model df       AIC       BIC   logLik   Test L.Ratio p-value
## mod1     1  6  -97.0903  -75.5893  54.5451
## mod2     2  8 -822.6553 -793.9873 419.3276 1 vs 2 729.565  <.0001
```

Part 3

```
library(geepack)

mod3=geeglm(weekly_rating ~  week*Ad_type,
               family = "gaussian",
               data = Ad_data,
               id = Subject, corstr = "exchangeable")
summary(mod3)
##
## Call:
## geeglm(formula = weekly_rating ~ week * Ad_type, family = "gaussian",
##     data = Ad_data, id = Subject, corstr = "exchangeable")
##
##  Coefficients:
##              Estimate  Std.err      Wald Pr(>|W|)
## (Intercept)   6.32662  0.04839 17095.249   <2e-16 ***
## week         -0.03371  0.01144     8.687   0.0032 **
## Ad_type       0.10824  0.08340     1.684   0.1943
## week:Ad_type  0.05219  0.02039     6.554   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation structure = exchangeable
## Estimated Scale Parameters:
##
##             Estimate Std.err
## (Intercept)   0.1218 0.02252
##   Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.8036 0.04176
## Number of clusters:   30  Maximum cluster size: 9
```

END OF QUESTION (4)