2017 First Year Exam - Methods
Statistics 210-202-203
June 26, 2016
9:00 to 12:00

- There are 4 questions on the examination, each with multiple parts. Select any 3 of them to solve.

- Your solutions to each of the 3 problems you solve should be written on separate sheets of paper. Label *each sheet* with your student identification number, the problem number, and the page for that problem written in the upper right hand corner. For example, the labeling on a page might be:

ID# 912345678
Problem 3, Page 2

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

1. A nationwide study was conducted with the aim of improving patient satisfaction at urban hospitals. A total of n hospitals were enrolled in this study between 2011 – 2013. Each hospital received a score at the start of the study in 2011 and at the end of the study in 2013. At the start of the study in 2011, all the hospitals received satisfaction scores between 70 – 75 which is considered to be good or average. The hospital-specific score was the average of all the scores recorded from each patient immediately upon discharge through a phone or a face-to-face interview conducted by a research company. The lowest possible score given by a patient is 0 (highly dissatisfied) and the highest possible score is 100 (highly satisfied).

   As noted above, all the hospitals have comparable satisfaction scores at the start of the study. As part of the effort to improve satisfaction, the hospitals were randomly assigned to participate in one of two workshops focused on improving communication by nurses and technicians. The approach of the first workshop was heavily based on practical exercises while the style of the second workshop was online self-training. The variables used in this study are defined as follows: Y_i is the patient satisfaction score for hospital i at the end of the study in 2013; W_i is the workshop training to which hospital i was assigned (here W_i which is either 1 or 2) and x_i is the average number of nurses assigned to 5 patients.

   (a) Formulate a model for patient satisfaction where the mean score takes into account potential different outcomes for the two workshops and potential interaction between the workshop method and the average number of nurses per 5 patients. Write the model in the form $Y_i = \mu_i + \epsilon_i$. Specify the mean component μ_i and the random component ϵ_i. For this problem it will be considered valid to consider the distribution of the scores to be approximately Gaussian, assuming standard classical assumptions for the linear regression model.

   (b) Write the model in matrix notation. Make sure that you specify the components of all vectors and matrices used in the model.

   (c) Explain how you could conduct a test for no interaction between the workshop method and the average number of nurses per 5 patients using the concept of nested models, where the full model $\mathcal{M}$ contains the main effects of training type and the average number of nurses per 5 patients and the interaction of the two main effects, and the reduced model $\mathcal{M}_0$ does not contain the interaction effect. A complete answer should include the null and alternative hypotheses; the linear models for $\mathcal{M}$ and $\mathcal{M}_0$; the test statistic; the distribution of the test statistic under the the null hypothesis; and the rejection region.

   (d) Define $\delta(a)$ to be the difference in the expectation of the distributions of satisfaction scores for the two training types where the average number of nurses per 5 patients is a. Under the no-interaction model, suppose that a 95% confidence interval for $\delta(a)$ (assuming equal probability tails) is given by (L_a, U_a). Now denote a 95% confidence interval (again assuming equal probability tails) for the difference $\delta(b)$, where $a < b$, to be (L_b, U_b). Which of the following is true? Explain.

      i. $U_a - L_a < U_b - L_b$
      ii. $U_a - L_a = U_b - L_b$
      iii. $U_a = U_b$ and $L_a = L_b$
      iv. No conclusion due to incomplete information.

   (e) Following the above notation, consider the following two parameterizations of the mean function μ_i:

$$\begin{aligned} \mu_i^A &= \alpha_0 + \alpha_1 W_{1i} + \alpha_2 x_i + \alpha_{12} W_{1i} x_i \\ \mu_i^B &= \beta_0 + \delta_0 W_{1i} + (\beta_1 + \delta_1 W_{1i}) x_i \end{aligned}$$

   where $W_{1i} = 1$ if the i-th hospital adopted the first workshop and $W_{1i} = 0$ if it adopted the second workshop.

i. Using the first parameterization μ_i^A, derive the mean function for the hospital population that used the first workshop and then derive the mean function for the hospital population that used the second workshop.

ii. Derive the mean functions from (i) using the second parameterization μ_i^B.

iii. Show that these two parameterizations are equivalent, i.e., there is a one-to-one function between the slopes and intercepts in the two parameterizations.

2. To study predictors that are associated with elevated fasting blood glucose Y_i (in mg/dl units) among African-American seniors, a geriatric doctor recorded the body mass index (BMI) x_{1i} (in kg/m^2), total daily average calorie intake x_{2i} (averaged over the previous 6 months) and gender. This dataset consists of $n = 27$ subjects randomly selected from the male senior African-American population and $n = 27$ from the female senior African-American population. Gender is encoded through the indicator variables G_{1i} for the male group and G_{2i} for the female group. We will assume that the distribution of the morning blood glucose, Y_i, for any level of BMI and total daily average calorie intake and for both males and females to be Gaussian. Consider the model

$$Y_i = (\beta_0 + \delta_0 G_{2i}) + (\beta_1 + \delta_1 G_{2i}) x_{1i} + (\beta_2 + \delta_2 G_{2i}) x_{2i} + \epsilon_i, \quad i = 1, \ldots, 54 \tag{1}$$

where the ϵ_i's are iid $N(0, \sigma^2)$. Moreover, the male subjects are indexed by $i = 1, \ldots, 27$ and the females by $i = 28, \ldots, 54$.

(a) Denote the response vector to be $\mathbf{Y} = [Y_1, \ldots, Y_{54}]'$; the error vector to be $\epsilon = [\epsilon_1, \ldots, \epsilon_{54}]'$; and the parameter vector to be $\underline{\beta} = [\beta_0, \beta_1, \beta_2, \delta_0, \delta_1, \delta_2]'$. Let's formulate the regression model in matrix notation to be

$$\mathbf{Y} = \mathbf{X}\underline{\beta} + \epsilon.$$

Give the elements of the design matrix $\mathbf{X}$.

(b) Denote the least squares estimator of $\underline{\beta}$ to be $\widehat{\underline{\beta}}$; the vector of predicted values to be $\widehat{\underline{Y}} = \mathbf{X}\widehat{\underline{\beta}}$ and the residuals to be $\underline{R} = \underline{Y} - \widehat{\underline{Y}}$. Suppose that the squared norm of the observed residual vector is $\underline{R}'\underline{R} = 50$. Give an unbiased estimate of the error variance σ^2.

(c) In the following questions, you will need to perform statistical inference. You are given the following calculations from the data:

$$\widehat{\underline{\beta}} = \begin{pmatrix} 100.00 & 1.00 & .01 & 10 & 0.10 & 0.01 \end{pmatrix}'$$

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{pmatrix} 1.0 & 0.10 & 0.001 & 0.02 & 0.01 & 0.01 \\ 0.10 & 0.10 & 0.5 & 0.2 & 0.01 & 0.01 \\ 0.001 & 0.5 & 0.001 & 0.1 & 0 & 0 \\ 0.02 & 0.02 & 0.01 & 1.0 & 0 & 0 \\ 0.01 & 0.01 & 0 & 0 & 1.0 & 0.01 \\ 0.01 & 0.01 & 0 & 0 & 0.01 & 1.0 \end{pmatrix}.$$

i. Consider only the population with total daily calorie intake of 2000 cal. Use ANOVA to test the null hypothesis that the male and female regression lines are parallel across BMI. That is, test the null hypothesis that there is no interaction between gender and BMI for the population with total daily calorie intake of 2000 cal.

ii. Using the calculations above, give a 95% confidence interval for the difference in the expected morning fasting blood glucose between male vs female African-Americans for the subpopulation with daily calorie intake of 2000 cal and BMI of 30 kg/m^2. Note: (i.) a complete answer should include an expression of the true unknown value in terms of the parameter vector $\underline{\beta}$; (ii.) you should specify the percentiles used in calculating the confidence interval; (iii.) you do not need to carry out any matrix calculations.

(d) Let $\underline{Q} = [1, \ldots, 1, 2, \ldots, 2]$ be a vector whose first 27 elements are all 1's and whose last 27 elements are all 2's. Show that the vector of residuals $\underline{R}$ and the vector $\underline{Q}$ are orthogonal, i.e., $\sum_{i=1}^{54} R_i Q_i = 0$.

3. In a country where there is perceived increased incidence of crimes and lawlessness a leader with a motive of increasing control of power can exploit the situation by declaring martial law. In a recent survey, each of $N = 1000$ survey respondents responded with either $Y_i = 1$ if they support the imposition of martial law and $Y_i = 0$ otherwise. The following variables were also collected for each participant: E_i which takes a value of 0 if he/she does not have a college degree and 1 if he/she does; G_i which takes on a value of 1 for a female participant and 0 otherwise; over-all satisfaction with the current economy x_i with a range of $0 - 100$ (0 if extremely dissatisfied and 100 if highly satisfied).

(a) Questions on preliminary analyses of the data.

i. In a preliminary analysis of the data, about 70% of the women without a college degree are against the imposition of martial law; and also about 70% of the women with a college degree are against the imposition of martial law. Does this indicate that there is no interaction between gender and education? Given a brief explanation (in 2-3 sentences) in terms of the log odds. You may provide sketches/plots in your explanation.

ii. Further preliminary analysis gives the following results: Among those without a college degree, 70% of the women and 75% of the men oppose martial law. However, among those with a college degree, 70% of the women and only 65% of the men oppose martial law. (a.) If there are about the same number of men and women in the sample, would this suggest that there is a difference between genders? Explain in 2-3 brief sentences. (b.) Does this suggest the presence of an interaction between gender and education?

(b) Questions on formulating a model.

i. Formulate a model for the response Y_i that takes into account (a.) the main effects of gender, education and over-all satisfaction with the economy and (b.) the two-way interaction effects between (i.) gender and education and (ii.) education and over-all satisfaction with the economy. A complete answer should include the conditional distribution of the response and an expression of the expectation of the response (or some transform of the expectation).

ii. Based on your model above express the probability that a randomly selected person with the following attributes: female, college graduate with a level of satisfaction of 80 would support the imposition of martial law.

iii. Compare the probability of supporting martial law for males vs. females among college graduates with a (low) level of satisfaction of 10. You may use the odds ratio or the log odds ratio.

iv. Compare the probability of supporting martial law for males vs. females among college graduates with a (high) level of satisfaction of 90. You may use the odds ratio or the log odds ratio.

v. Did you expect the two odds ratio (or log odds ratio) above to be identical? Explain.

(c) Questions related to a specific model. Suppose that we fit the model

$$Y_i|E_i, x_i \sim \text{Bernoulli where } \Pr(Y_i = 1|E_i, x_i) = \frac{\exp(\beta_0 + \delta_0 E_i + \beta_1 x_i + \delta_1 E_i x_i)}{1 + \exp(\beta_0 + \delta_0 E_i + \beta_1 x_i + \delta_1 E_i x_i)}.$$

Define the parameter vector $\beta = [\beta_0, \beta_1, \delta_0, \delta_1]$; $\widehat{\gamma}$ to be the maximum likelihood estimate of the parameter γ. Suppose that the following quantities were computed:

$$[\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\delta}_0, \widehat{\delta}_1] = [1, -0.01, -2, -0.01] \tag{2}$$

$$\text{cov}\widehat{\beta} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \frac{1}{4} \times 10^{-4} & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & \frac{1}{25} \times 10^{-4} \end{pmatrix} \tag{3}$$

i. Give the estimated probability curves for the college graduate population and the non-college graduate population. Write your answer as: "The estimated probability curves for the male and female populations are, respectively,"

$$\widehat{\Pr}(Y_i = 1|E_i = 0, x_i) = \ldots$$
$$\widehat{\Pr}(Y_i = 1|E_i = 1, x_i) = \ldots$$

ii. Derive a 95% confidence interval for the probability of supporting martial law for the college graduate population with low level of satisfaction of $x = 10$.

iii. Derive a 95% confidence interval for the odds ratio of supporting martial for the college graduate population with a level of satisfaction of $x = 10$ vs the college graduate population with a level of satisfaction of $x = 90$.

iv. Conduct a formal test for the interaction between education and satisfaction.

v. Let θ be the value of satisfaction where the probability of supporting martial law among non-college graduates is $\frac{2}{3}$.

A. Find a point estimate for θ.

B. Derive an asymptotic 95% confidence interval for θ.

4. Emerging evidence is showing the importance of an individual microbiome in a variety of diseases. The human microbiome describes the collection of microbes that live on and inside each individual. The number of microbes outnumber our "human" cells 10 to 1. In particular, the *diversity* of microbes has been linked to several human diseases: for example, low diversity in the gut has been linked to obesity and inflammatory bowel disease.

(a) A summary statistic called *alpha diversity* is usually computed to provide information on the distribution of the microbes sequenced in an individual at a given time. This is a continous measurement, with values that range usually between 1 and 4.5. The following dataset contains information on the alpha diversity measurements from gut microbiome in a study of 100 individuals assigned to two different treatments (50 in Treatment A and 50 in Treatment B), collected each week over a period of 12 weeks.

i. Refer to the linear mixed effects model **mod1** in the Appendix. Write the mathematical form of the assumed model in matrix form (the assumed model, not the fitted model). Clearly define any variables used, and write out the elements of each vector or matrix in the model. Identify which terms in the model are fixed, and which are random. State all model assumptions.

ii. Write a sentence interpreting the estimated standard deviation of the random intercepts in the context of the problem.

iii. Provide an interpretation of each estimated (fixed effect) coefficient in **mod1**. Based on the estimated coefficients, determine if there is enough evidence of a treatment effect over time. Motivate your answer.

iv. Write the estimated marginal variance-covariance matrix.

v. Describe a likelihood ratio test to determine if a random slope should be added to the model. Clearly specify the hypotheses being tested, the estimation method and the relevant test statistic. You may refer to model **mod1.s** in the Appendix, if necessary.

vi. Recent studies using molecular methods have also indicated clear differences in the composition of gut microbiota among young individuals, adults and the elderly. Consider the age group as a variable with 3 categories. Describe a likelihood ratio test to determine if the microbiome diversity is associated to the different age groups in the study considered here. Clearly specify the hypotheses being tested, the estimation method and the relevant test statistic. You may refer to **mod1.age** in the Appendix if necessary.

(b) As an alternative to computing a summary statistic (which summarizes measurements from all microbiome species into one single value), microbiologists are often interested in investigating how a particular microbial species changes over time. Since microbes are observed from sequencing experiments, this generally corresponds to tracking how the observed counts of that particular species change from start to end of a treatment.

i. Propose a marginal model for studying how a species' microbial counts change over treatment and time in the population. Clearly define the mean model, the conditional covariance function, and the assumed model for the within-subject association among the responses.

ii. Now consider **mod2** in the Appendix. Write a sentence interpreting the coefficient of the interaction term for this model.

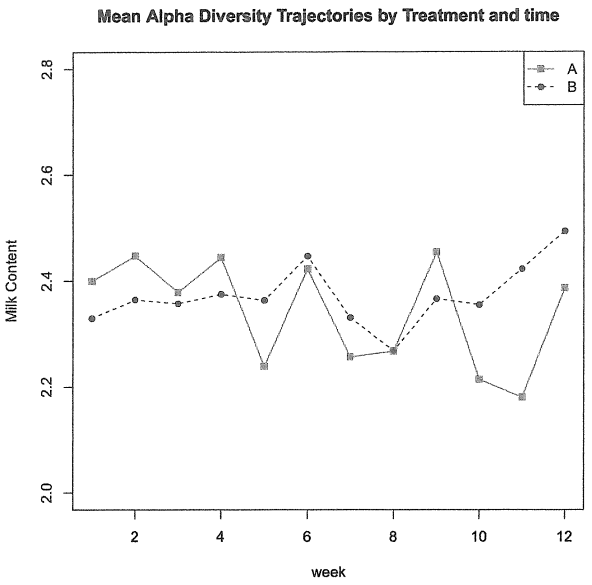
iii. In fitting model **mod2**, we considered a compound symmetry working correlation structure. Alternatively, one could have considered a different working correlation structure, e.g. an autoregressive (AR) structure. As a result, (1) would the estimation of the coefficients in the marginal model change? (2) If yes, why would they change? (3) If yes, would you expect the change to be substantial? Motivate your answer.

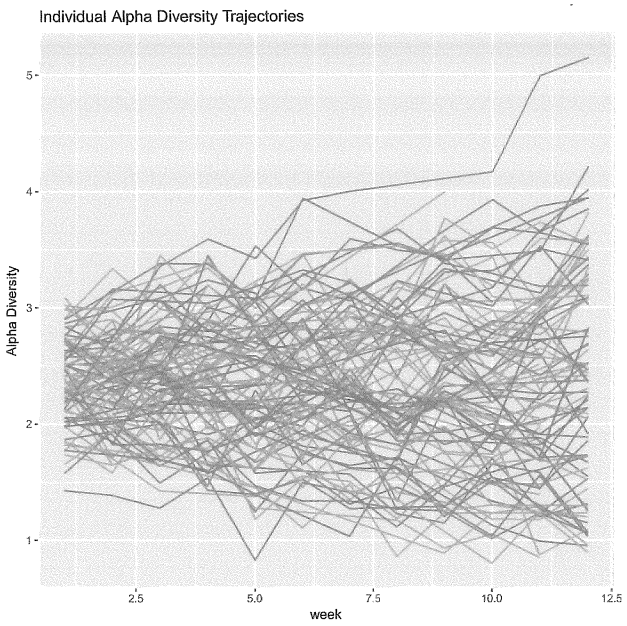
(c) The dataset contains quite a few missing data. The investigators of this study hypothesize that the reason might be age-dependent: young individuals might be less prone to collect the necessary stool samples for the prolonged time required by the study.

i. Determine if the missing data mechanism specifies missingness completely at random (MCAR), missingness at random (MAR) or not missingness at random (NMAR). Motivate your answer.

ii. Based on your answer above, discuss whether and under what circumstances model **mod2** can be used to obtain valid inferences under the assumed missing data mechanism. Motivate your answer, commenting in particular on the required first-order and second-order assumptions.

Appendix





mod1

```
mod1<-lme(resp~week*treat, random=~1|ID,data=data1)
summary(mod1)
## Linear mixed-effects model fit by REML
##  Data: data1
##        AIC      BIC    logLik
##   1055.481 1084.658 -521.7405
##
## Random effects:
##  Formula: ~1 | ID
##         (Intercept)  Residual
## StdDev:   0.5604631 0.3482212
##
## Fixed effects: resp ~ week * treat
##                  Value  Std.Error  DF   t-value p-value
## (Intercept)  2.4089603 0.08531689 858 28.235444  0.0000
## week        -0.0105992 0.00460557 858 -2.301381  0.0216
## treatB      -0.1035720 0.12223457  98 -0.847322  0.3989
## week:treatB  0.0230325 0.00659442 858  3.492718  0.0005
##  Correlation:
##             (Intr) week   treatB
## week        -0.345
## treatB      -0.698  0.241
## week:treatB  0.241 -0.698 -0.352
##
## Number of Observations: 960
## Number of Groups: 100
```

mod1.s

```
##        Model df       AIC       BIC    logLik
## mod1       1  6 1055.4810 1084.6575 -521.7405
## mod1.s     2  8  649.4011  688.3032 -316.7006
```

mod1.age

```
##            Model df      AIC      BIC    logLik   Test  L.Ratio p-value
## mod1.age       1  6 1031.278 1060.479 -509.6388
## mod1.age.1     2  8 1033.892 1072.827 -508.9459 1 vs 2 1.385793  0.5001
```

mod2

```
summary(mod3)
##
## Call:
## geeglm(formula = species ~ treat * week, family = **, data = data2,
##     id = ID, corstr = "exchangeable")
##
##  Coefficients:
##             Estimate  Std.err    Wald Pr(>|W|)
## (Intercept)  7.83088  0.82789  89.470   <2e-16 ***
## treatB      -1.61574  0.91544   3.115   0.0776 .
## week         0.81809  0.03501 546.156   <2e-16 ***
## treatB:week -0.41197  0.03558 134.078   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation: Structure = exchangeable  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha   0.3682  0.2335
## Number of clusters:   100   Maximum cluster size: 12
```