2019 First year Exam – Methods

Statistics
210-211-212
June 24, 2019
9:00 to 12:00

Instructions:

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.

DO NOT WRITE ON BOTH SIDES

Label each sheet in the upper right hand corner:

1) with your student id number

2) with the problem number

3) and the page number for that problem.

For example, the labeling on a page might be:

ID#912346378

Problem 3, page 2

METHODS 210-211-212 (2019) Problem 1

An education researcher is exploring the effectiveness of using a supplementary text or a study skills workshop to improve student performance in an introductory statistics class. One hundred students are randomly assigned into four groups with 25 students assigned to each of four treatment groups corresponding to the additional resources they are given (none, supplement, skills, both). The final exam score is the response variable. A summary of the results of the study are provided here:

group	n	mean	s.d.
none	25	82.0	22.3
supplement	25	77.5	26.5
skills	25	84.8	24.5
both (skills and supplement)	25	100.8	33.3

(a) Based on these data what is the estimated main effect of the supplementary text on final exam scores? What is the estimated main effect of the skills workshop on final exam scores?

(b) Compute the analysis of variance test statistic for testing the hypothesis that the four treatment groups have equal population means. You can assume the four treatment groups have equal population variances. Tell how you would assess the significance of the test statistic.

(c) Assuming the data in each group follow a normal distribution with the same variance, derive the likelihood ratio test statistic for testing the hypothesis that the four groups have equal population means. How does it relate to the test statistic you computed in part (b).

(d) The analysis of variance test is often described as an unfocused or omnibus test. Explain.

(e) Suppose the researcher had an 'a priori' hypothesis that neither the supplementary text alone nor the study skills workshop alone would be sufficient to improve student performance. She believed that you need both additional resources to have any benefit. Propose and carry out an analysis that addresses the researcher's hypothesis.

(f) The researcher also has available a pre-test score for each student. A summary of the pre-test scores for each group is provided in the table below. Do these data support the researcher's claim to have randomly assigned treatments to students? Explain. (No formal statistical test is required.)

group	n	mean	s.d.
none	25	40.8	11.3
supplement	25	39.4	7.8
skills	25	39.7	10.1
both (skills and supplement)	25	39.3	10.3

(g) The researcher's supervisor suggests that a more appropriate analysis should incorporate the pre-test scores. The researcher argues that because it was a randomized study there is no need to incorporate the pre-test scores. Comment on the arguments of the researcher and her supervisor.

(h) The supervisor is the supervisor so the researcher went ahead and carried out the analysis incorporating pre-test score. The results are summarized below. Was it helpful to the researcher to incorporate the pre-test scores in the analysis? You can refer to any of the data provided here to support your answer.

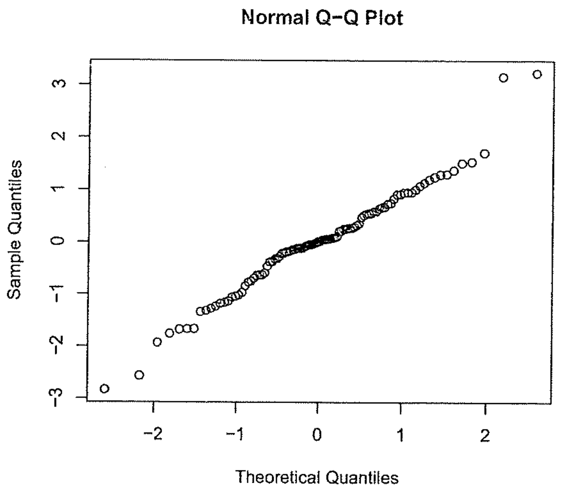
```
Call:
lm(formula = final ~ pretest + as.factor(group), data = education)

Residuals:
    Min      1Q  Median      3Q     Max
-51.846 -11.149   0.229  10.684  59.751
```

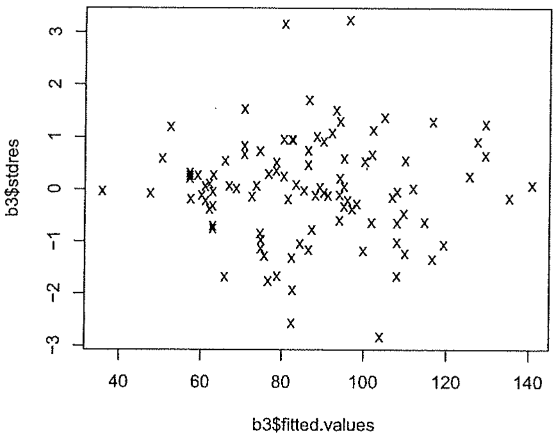
Coefficients:

	Estimate	Std. Error	t value	$Pr(>\|t\|)$
(Intercept)	2.6297	8.7520	0.300	0.764478
pretest	1.9444	0.1935	10.048	$< 2e-16$
I(supplement)	-1.8357	5.3484	-0.343	0.732195
I(skills)	4.9799	5.3460	0.932	0.353947
I(supplement & skills)	21.6777	5.3496	4.052	0.000104

```
Residual standard error: 18.89 on 95 degrees of freedom
Multiple R-squared:  0.5633,Adjusted R-squared:  0.5449
F-statistic: 30.64 on 4 and 95 DF,  p-value: 2.247e-16
```

(i) A normal probability plot and a plot of the standardized residuals versus fitted values are provided below. Comment on whether these plots support the ususal linear regression assumptions.





METHODS 210-211-211 (2019) Problem 2

A linear model is used to assess the effect of unpredictability of maternal behavior on the development of a child's capacity for self-control at 5 years of age. Based on the analysis of an early-life (when the child was 1 year old) video interaction of mother and child a measure of unpredictability called the entropy is computed. Entropy is the primary predictor of interest with larger values indicating greater unpredictability. Greater unpredictability is expected to have a negative impact on child self-control. A number of other factors are known to impact the development of self-control including the gender of the child, family income, maternal age, maternal education, and maternal mood (depression/anxiety score). A linear regression model is fit to all of the variables (all except Gender were standardized). A summary of the regression results are provided here:

```
Call:
lm(formula = SelfControl ~ EntRat + Sex + Income + MatEdYrs +
    MatDepAnxAvg + MatAgeatDeliv, data = newdat2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.48070 -0.58339 -0.04035  0.66483  2.54094
```

```
   Coefficients:
                    Estimate   Std. Error
    (Intercept)      0.20150     0.06873
    Entropy         -0.10548     0.04852
    Gender          -0.42028     0.09255
    Income          -0.05075     0.06226
    MaternalEduc     0.13427     0.06003
    MaternalMood    -0.26790     0.05409
    MaternalAge      0.09130     0.05603
```

```
Residual standard error: 0.902 on 394 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1687,Adjusted R-squared:  0.156
F-statistic: 13.32 on 6 and 394 DF,  p-value: 9.235e-14
```

(a) Interpret the estimated coefficent for the predictor of interest (Entropy).

(b) A test of the hypothesis that the coefficient of Entropy is zero yields a p-value of .03. Carefully explain the meaning of the p-value in this context.

(c) Provide and interpret a 95% confidence interval for the coefficent of Entropy. Note the sample size is quite large.

(d) The correlation of income and child self-control is positive ($r = 0.2$) and highly significant. The coefficient of income in the regression is negative. Explain how this can happen.

(e) Income and maternal education are assumed to have similar effects on child self-control. We wish to obtain a 95% confidence interval for the difference in their coefficients, $\beta_{income} - \beta_{educ}$. Do you have the information that you need to compute this confidence interval? If yes, compute the confidence interval. If no, describe the additional information that would be needed.

(f) Gender clearly has a significant impact on child self-control. Gender is coded as 1 for males and 0 for females; thus the regression results suggest that males have lower self-control scores than females. It is of interest to assess whether the effect of entropy and the other predictors may differ according to the gender of the child. Thus another model is fit that includes interactions of gender with the other predictors. Results are provided below.

(i) Explain how the interpretation of the coefficient of entropy changes in this model relative to the model fit earlier.

(ii) Is there evidence that the effect of the predictors differs according to the gender of the child? Compute an appropriate test statistic to address this question and identify its reference distribution.

```
Call:
lm(formula = SelfControl ~ EntRat + Sex + Income + MatEdYrs +
    MatDepAnxAvg + MatAgeatDeliv + Sex * EntRat + Sex * Income +
    Sex * MatEdYrs + Sex * MatDepAnxAvg + Sex * MatAgeatDeliv,
    data = newdat2)

Residuals:
     Min       1Q   Median       3Q      Max
-2.35872 -0.60188 -0.02828  0.60704  2.62749
```

		Estimate	Std. Error
	(Intercept)	0.21991	0.07123
	Entropy	-0.02667	0.07411
	Gender	-0.43522	0.09409
	Income	0.04961	0.09128
	MaternalEduc	0.03307	0.09059
Coefficients:	MaternalMood	-0.29057	0.07280
	MaernalAge	0.24354	0.08816
	GenderEntropy	-0.13742	0.09765
	GenderIncome	-0.18694	0.12705
	GenderMaternalEduc	0.15168	0.12063
	GenderMaternalMood	0.01920	0.11394
	GenderMaternalAge	-0.25353	0.11394

```
Residual standard error: 0.8949 on 389 degrees of freedom
  (3 observations deleted due to missingness)
Multiple R-squared:  0.1921,Adjusted R-squared:  0.1693
F-statistic: 8.408 on 11 and 389 DF,  p-value: 2.565e-13
```

(g) Regardless of your answer to the previous part, let's focus on the initial regression (without interactions) for this final section of the problem. The dataset contains a number of instances of siblings. Reviewers of the research suggest that this may lead to a violation of the assumption of independent errors in the regression analysis.

(i) How would the lack of independent errors impact the interpretation of the regression results?

(ii) Describe how you can use the residuals from the regression analysis to assess whether this is an important issue.

(iii) Assume we determine that there is a violation of the assumption of independent errors. Briefly describe how you might improve the model to address this issue.

METHODS 210-211-212 (2019), Problem 3

An epidemiologic study is carried out to investigate factors affecting pre-school asthma, a serious and sometimes life-threatening respiratory condition. The covariate of chief interest is E_i, which equals 1 if the child i was exposed to smoking at the time of examination, and 0 otherwise. The response variable is pre-school asthma, Y_i, coded as $Y_i = 0$ if absent; $Y_i = 1$ if mild-to-moderate and $Y_i = 2$ if severe. Observed data are provided in the following table:

E_i	$Y_i = 0$	$Y_i = 1$	$Y_i = 2$	total
0	213	61	84	358
1	14	11	17	42
Total	227	72	101	400

(a) The following model is assumed,

$$\log\left\{\frac{p_{ij}}{p_{i0}}\right\} = \beta_{0j} + \beta_{1j}E_i,$$

for $j = 1, 2$, where $p_{ij} = P(Y_i = j|E_i)$.

Estimate the model parameters based on the data provided above.

In fact, data are available on several characteristics of each child, with the child-specific information summarized by a 6×1 covariate vector $\mathbf{X}_i$. The model in this case is given by

$$\log\left\{\frac{p_{ij}}{p_{i0}}\right\} = \beta_{0j} + \boldsymbol{\beta}_{1j}'\mathbf{X}_i, \qquad (1)$$

where $\boldsymbol{\beta}_{1j}$ is a 6×1 parameter vector. This model is the basis of the remaining parts in this question.

(b) Suppose that $X_{i3} = LBW_i$, an indicator variable taking the value 1 if child i was of low birth weight and 0 otherwise. For $j = 1, 2$, provide an interpretation for β_{1j3} (or some suitable transform thereof), the 3rd element of $\boldsymbol{\beta}_{1j}$.

(c) One of the junior statisticians wonders if the model could be simplified by using a single common regression parameter for $j = 1, 2$. Describe how you could test this hypothesis, $H_0 : \boldsymbol{\beta}_{11} = \boldsymbol{\beta}_{12}$; provide the test statistic and reference distribution. Note: only describe what you would do WITHOUT actually carrying out the test using the above data.

(d) The project team holds one last meeting before submitting a manuscript based mainly on results from model (1). At this meeting, the lead investigator mentions one final 'detail': the data set used for analysis, including the estimation of model (1), was arrived at through "case-control" sampling, where all asthmatic children (i.e., $Y_i = 1$ or $Y_i = 2$) were selected, as well as a random sample of 'controls' ($Y_i = 0$). After learning the number of controls in the original study population, the data analyst assigned to the project *updates* the parameter estimators for model (1) using this additional information (data not shown here).

Should the *original* and *updated* parameter estimates of (β_{0j}, β_{1j}) be equal? If not, how should they differ?

(e) Suppose that $X_{i4} = AGE_i$ in months, and you are not sure if including only the linear term of AGE variable in model (1) is appropriate. How would you conduct the model diagnostics?

END OF QUESTION (3)

METHODS 210-211-212 (2019), Problem 4

Annual cognitive testing is often performed in elderly individuals to assess potential changes in cognition and progression of dementia (a group of symptoms associated with a decline in memory). One test that is performed is the letter fluency test, where participants are asked to state as many words starting with the letter 'A' that they can think of in a 60 second time period. Longitudinal letter fluency data were obtained on $N = 10,900$ individuals from the National Alzheimer's Disease Coordinating Center. Data were collected annually starting from each participant's first visit to the center. The frequency of visits by participant are provided in Appendix 3. Our goal is to estimate and test changes in letter fluency by cognitive status. Data are available on three groups: Normal ($N = 6,583$), Mildly Cognitively Impaired (MCI; $N = 2,723$), and Alzeimer's Disease (AD; $N = 1,594$). The count response of letters for participant i at the first (baseline) visit is denoted Y_{i0}. The remaining observed counts are then denoted $Y_{i1}, \ldots, Y_{in_i}$, $i = 1, \ldots, 10900$.

To begin, we will consider only the change in response by diagnostic groups between the baseline and first followup visit. Consider the following two mean models:

$$\log(\mu_{ij}) = \beta_0 + \beta_1 I_{MCI,i} + \beta_2 I_{AD,i} + \beta_3 I_{[j=1],i} + \beta_4 I_{MCI,i} \times I_{[j=1],i} + \beta_5 I_{AD,i} \times I_{[j=1],i}, j = 0, 1 \quad (1)$$

$$\log(\mu_{ij}) = \gamma_0 + \gamma_1 I_{[j=1],i} + \gamma_2 I_{MCI,i} \times I_{[j=1],i} + \gamma_3 I_{AD,i} \times I_{[j=1],i}, \quad j = 0, 1 \quad (2)$$

where $\text{E}[Y_{ij}] = \mu_{ij}$, $I_{MCI,i}$ and $I_{AD,i}$ are indicators that subject i is MCI or AD, respectively, and $I_{[j=1],i}$ is an indicator of whether the response for subject i was taken at the followup visit. You may assume that adjustment for confounding is not an issue.

(a) Appendix 3 gives GEE regression model estimates for each of the above models assuming a Poisson mean-variance relationship and an independence working correlation structure. Provide precise interpretations of the following model parameters in terms that are understandable by a statistical layman:

i. $\widehat{\beta}_0, \widehat{\beta}_2, \widehat{\beta}_3, \widehat{\beta}_5$

ii. $\widehat{\gamma}_1, \widehat{\gamma}_2$

(b) In the context of the study design, which of the two mean models (1 or 2) is more appropriate for addressing the scientific question of interest. Briefly justify your answer and state which parameters (or combination of parameters) are of primary interest.

(c) For the model that you selected in part (b), provide an asymptotically valid Wald statistic for testing the null hypothesis that there is no change in the mean number of letters reported by Normal, MCI, and AD participants. Your answer should be left unevaluated and written notationally (matrix notation is perfect) and you should state the approximate distribution of your test statistic under the null hypothesis. (Note: I have not provided you with any covariance matrix estimates for the estimated model parameters, so you should simply define the covariance matrix notationally in your answer, if needed.)

(d) For the model that you selected in part (b), parameter estimation was carried out via maximum likelihood and assuming that $Y_{ij} \sim_{\text{ind}} F(\mu_{ij}) \equiv \text{Poisson}(\mu_{ij})$, with $\mu_{ij} = \exp\{\eta_{ij}\}$ where η_{ij} is the linear predictor in your chosen model. Thus we obtained parameter estimates by solving the score equation:

$$\mathcal{U}^F(\vec{\theta}) = \vec{0},$$

where $\vec{\theta}$ is either $\vec{\beta}$ or $\vec{\gamma}$, depending upon which model you selected in (b). Provide the form of $\mathcal{U}^F(\vec{\theta})$. (Note: You need not derive it. Just write it down and define any new notation you added.)

(e) Let $\mathcal{I}^F_{k,l}$ denote the $k-l$ element of Fisher's expected information based upon distribution F (i.e. $\mathcal{I}^F_{k,l} = \mathrm{E}_F\left[\frac{\partial}{\partial \beta_l}\mathcal{U}_k(\vec{\theta})\right]$). Suppose that in reality $Y_{ij} \sim G$ with $\mathrm{E}_G[Y_{ij}] = \mu_{ij}$. Show that

$$\mathrm{E}_G\left[\frac{\partial}{\partial \beta_l}\mathcal{U}_k(\vec{\theta})\right] = \mathrm{E}_F\left[\frac{\partial}{\partial \beta_l}\mathcal{U}_k(\vec{\theta})\right] = \mathcal{I}^F_{k,l}.$$

You may assume any regularity conditions needed.

(f) Again suppose that in reality $Y_{ij} \sim G$ with $\mathrm{E}_G[Y_{ij}] = \mu_{ij}$. Write down the asymptotic distribution of $\widehat{\vec{\theta}}$, the estimate obtained from solving $\mathcal{U}^F(\vec{\theta}) = \vec{0}$.

For the remainder of the problem, we will consider all repeated observations of the test scores for each participant over multiple years. Let Y_{ij} denote the observed test score for participant i at followup year j, $j = 0, \ldots, n_i - 1$. Our interest is in the first-order rate of change (slope) of test scores over time by diagnostic group.

(g) It is hypothesized that the rate of change in test scores is the same by diagnostic group for the first three years, then differs by group after that. Write down an appropriate mean model to address this scientific hypothesis (denote your model parameters by $\beta_0, \beta_1, \ldots, \beta_p$). You may assume that adjustment for confounding is not an issue.

(h) Suppose you used a GEE model assuming a Poisson mean-variance relationship and exchangeable working correlation structure to estimate the mean model parameters in your answer to (g). Let $\widehat{\vec{\beta}}$ denote the resulting estimator and let $\widehat{\Sigma}$ denote the empirical robust variance estimator for $\widehat{\vec{\beta}}$.

i. Write the null and alternative hypothesis for testing whether the slope of MCI participants differs from that of Normals after year 3.

ii. Write down an appropriate test statistic for testing the above null hypothesis (matrix notation is perfect), along with the corresponding approximate reference difference of the statistic under the null hypothesis.

iii. Write the null and alternative hypothesis for testing whe+her the slope of'AD participants differs from 0 at any time during followup (before or after 3 years).

iv. Write down an appropriate test statistic for testing the above null hypothesis (matrix notation is perfect), along with the corresponding approximate reference difference of the statistic under the null hypothesis.

APPENDIX 3

```
##
##### Frequency of subjects by number of visits completed
##
> table(table(letter$id))

   3    4    5    6    7    8    9   10   11   12
3301 2280 1464 1237  939  717  544  264  124   30

##
##### Number of subjects by diagnostic group
##
> table( u.letter$dx )

   Normal    MCI    AD
     6583   2723  1594

##
##### Fit of Model 1
##
> summary(fit.1)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logarithm
 Variance to Mean Relation: Poisson
 Correlation Structure:     Independent

Coefficients:
                           Estimate Naive S.E.   Naive z Robust S.E. Robust z
(Intercept)                2.611813  0.0040753 640.89002   0.0034835 749.7740
I_MCI                     -0.344130  0.0085587 -40.20831   0.0090294 -38.1122
I_AD                      -0.785323  0.0129240 -60.76448   0.0161798 -48.5371
I_j1                       0.035222  0.0057998   6.07301   0.0032500  10.8376
I_MCI:I_j1                 0.011157  0.0121217   0.92044   0.0095041   1.1739
I_AD:I_j1                 -0.112870  0.0180655  -6.24784   0.0152408  -7.4058

Estimated Scale Parameter:  1.4895
Number of Iterations:  1

Working Correlation
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```

```
##
##### Fit of Model 2
##
> summary(fit.2)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logarithm
 Variance to Mean Relation: Poisson
 Correlation Structure:     Independent

Coefficients:
                           Estimate Naive S.E. Naive z Robust S.E. Robust z
(Intercept)                 2.44659  0.0037420 653.814   0.0040001  611.637
I_j1                        0.20044  0.0058443  34.298   0.0041906   47.832
I_MCI:I_j1                 -0.33297  0.0093382 -35.657   0.0091032  -36.577
I_AD:I_j1                  -0.89819  0.0137317 -65.410   0.0161921  -55.471

Estimated Scale Parameter:  1.7627
Number of Iterations:  1

Working Correlation
     [,1] [,2]
[1,]    1    0
[2,]    0    1
```