2019 First year Exam – Methods

Statistics
210-210B-210C
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 210-210B-210C (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.

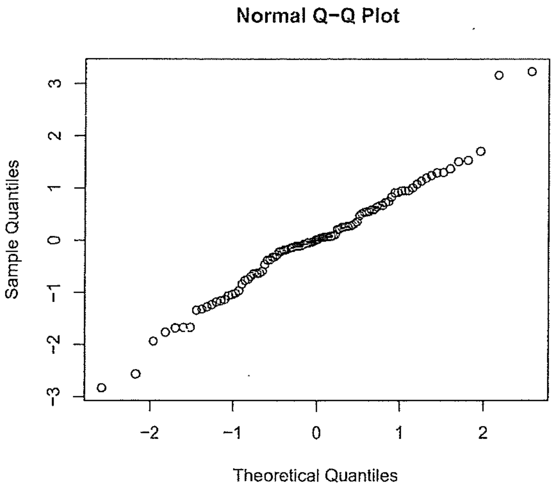
```
Call:
lm(formula = final ~ pretest + as.factor(group), data = education)

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

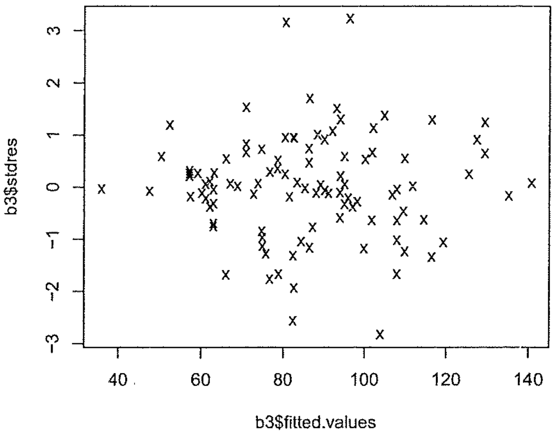
Coefficients:

	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 210-210B-210C (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
Coefficients:	(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
	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 210-210B-210C (2019) Problem 3

Aspirin, or acetylsalicylic acid (ASA), is a commonly used pain reliever for minor aches and pains and to reduce fever. Since the aspirin blocks the action of platelets, reducing clots, it has also been administered to patients immediately after a heart attack to prevent further clot formation and cardiac tissue death. In 2016, the U.S. Preventive Services Task Force (USPSTF) has come up with slightly modified recommendations for the primary prevention of cardiovascular disease using aspirin. Based on their review of the published data, they encourage the use of aspirin in adults aged 50 to 69 years with a high risk of cardiovascular disease. In individuals younger than 50 years of age and older than 70 years of age, the data are insufficient to recommend daily aspirin.

We consider a study investigating the association between heart attacks and the use of aspirin. Since Age is a potential confounder, we consider the following indicator variables:

$$Y = \begin{cases} 1 & \text{if heart attack} \\ 0 & \text{if no heart attack} \end{cases} \qquad \text{Aspirin} = \begin{cases} 1 & \text{if aspirin} \\ 0 & \text{if placebo} \end{cases}$$

$$\text{Age1} = \begin{cases} 1 & \text{if age is } 40-50 \\ 0 & \text{otherwise} \end{cases} \qquad \text{Age2} = \begin{cases} 1 & \text{if age is } > 50 \\ 0 & \text{otherwise} \end{cases}$$

The following table shows the results of fitting logistic regression models for $P(Y=1)$:

Model	Covariates	Estimate $\hat{\beta}$	Standard Error	log-likelihood
1	None	-2.99	0.19	-116.54
2	Aspirin	-0.82	0.41	-114.41
3	Age1	-0.19	0.47	-116.27
	Age2	0.17	0.45	
4	Aspirin	-0.82	0.41	-114.14
	Age1	-0.18	0.47	
	Age2	0.19	0.45	
5	Aspirin	-0.65	0.63	-113.83
	Age1	-0.22	0.59	
	Age2	0.39	0.54	
	(Age1)*Aspirin	0.10	0.97	
	(Age2)*Aspirin	-0.68	1.03	

(a) Test the null hypothesis of constant aspirin effect on the risk of heart actack across age groups (i.e, no interaction between aspirin and age).

(b) Based on the model with additive/main effects for age and aspirin (Model 4):

   i. Calculate the MLE of the odds ratio of aspirin use on heart attack, adjusting for age. Provide a 95% confidence interval for this odds ratio and interpret it in context.

   ii. Perform a Wald test of the null hypothesis that there is no effect of aspirin on the risk of heart attack, controlling for age. What do you conclude?

   iii. Perform a likelihood ratio test of the null hypothesis that there is no effect of age on the risk of heart attack, controlling for aspirin use. State your conclusion.

(c) Evaluate the deviance of each model provided in the table and assess its goodness-of fit. Which models do not provide adequate fit to the data?

(d) Perform model selection using analysis-of-deviance. Make sure you describe all the steps to arrive at your final model.

METHODS 210-210B-210C (2019) Problem 4

An insurance company is interested in assessing the impact of two new insurance plans on the use of medical services over a wide area of the United States. Thus, the company rolls out the two plans in several regions of the country, both urban and rural and records the monthly visits to in-network doctors over a period of 12 months, based on the claims from individuals in each region (id), in order to evaluate the impact of the two insurance plans. Covariates of interest are the type of region (urban or rural), and the seniority of the people in the area, captured by the median age in that area.

(a) For the following question, you may refer to the model mod1 in the Appendix. 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. Identify which terms in the model are fixed, and which are random.

(b) Comment on the use of the offset term: what is the rationale for including the term in the model?

(c) Provide an interpretation of each estimated coefficient in mod1. Based on the estimated coefficients, what would be your recommendation to the insurance company?

(d) Now refer to model mod2. State all the modeling assumptions and discuss the estimation technique used to fit the model.

(e) Discuss if your interpretation of the estimated coefficients differs with respect to model mod1 above, and if so, how.

(f) Now we consider an updated model that takes into consideration the effect of urban or rural regions (see mod2.urban). Rural regions may be characterized by a reduced availability of medical centers. Describe the test used to assess the importance of urban regions and interpret the results in the context of the problem.

(g) Model mod3 considers a different working correlation structure for the serial correlation of the observed measurements. Discuss how the estimated effects are expected to change as a result of the new correlation structure. Motivate your answer.

(h) Based on the output shown, which of the two models characterized by the exchangeable and AR(1) working correlation structures (respectively) would you prefer, and why?

(i) The dataset contains some missing data. The investigators believe the reason is due to a network-related difficulty which sometimes prevents records from being sent from a region to the headquarters of the insurance company. Identify the missing data mechanism and discuss the validity of the modeling approaches in mod1 and mod2 under the identified mechanism. Would your answer change if the missing data were due to a policy that allowed a region not to send a monthly record if the differences with respect to the previous two months were under a prespecified threshold?

Appendix

mod1

```
## Generalized linear mixed model fit by maximum likelihood (Laplace
##   Approximation) [glmerMod]
##  Family: poisson  ( log )
## Formula: NVisits ~ factor(Insurance_type) * month + (1 | id)
##    Data: Insurance.data
##  Offset: log(Area)
##
##      AIC      BIC   logLik deviance df.resid
##    12991    13016    -6490    12981     1195
##
## Scaled residuals:
##     Min      1Q Median      3Q     Max
## -6.137 -0.925 -0.071  0.881  6.537
##
## Random effects:
##  Groups Name        Variance Std.Dev.
##  id     (Intercept) 0.171    0.413
## Number of obs: 1200, groups:  id, 100
##
## Fixed effects:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    2.89173    0.05852   49.42   <2e-16 ***
## factor(Insurance_type)B       -0.13499    0.08277   -1.63    0.103
## month                          0.01580    0.00620    2.55    0.011 *
## factor(Insurance_type)B:month  0.12190    0.00885   13.78   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Correlation of Fixed Effects:
##             (Intr) fc(I_)B month
## fctr(Ins_)B -0.707
## month       -0.049  0.034
## fctr(In_)B:  0.034 -0.050  -0.701
```

mod2

```
##
## Call:
## geeglm(formula = NVisits ~ factor(Insurance_type) * month, family = poisson(link = "log"),
##     data = Insurance.data, offset = log(Area), id = id, corstr = "exchangeable")
##
##  Coefficients:
##                               Estimate Std.err    Wald Pr(>|W|)
## (Intercept)                     2.9790  0.0920 1047.87   <2e-16 ***
## factor(Insurance_type)B        -0.1585  0.1094    2.10   0.1476
## month                           0.0158  0.0255    0.38   0.5361
## factor(Insurance_type)B:month   0.1219  0.0437    7.79   0.0052 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)      118    42.8
##
## Correlation: Structure = exchangeable  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.977  0.0504
## Number of clusters:   100   Maximum cluster size: 12
```

mod2.urban

```
## coefficient of Urban
anova(mod2, mod2.urban, test=FALSE)
```

```
Model 1 NVisits ~ Urban + factor(Insurance_type) * month
Model 2 NVisits ~ factor(Insurance_type) * month
  Df   X2 P(>|Chi|)
1  1 6.59      0.01 *
```

mod3

```
##
## Call:
## geeglm(formula = NVisits ~ factor(Insurance_type) * month, family = poisson(link = "log"),
##     data = Insurance.data, offset = log(Area), id = id, corstr = "ar1")
##
##  Coefficients:
##                               Estimate Std.err    Wald Pr(>|W|)
## (Intercept)                     2.9824  0.0907 1080.16   <2e-16 ***
## factor(Insurance_type)B        -0.1718  0.1086    2.50   0.1139
## month                           0.0154  0.0261    0.35   0.5555
## factor(Insurance_type)B:month   0.1383  0.0433   10.19   0.0014 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Estimated Scale Parameters:
##             Estimate Std.err
## (Intercept)      118    42.6
##
## Correlation: Structure = ar1  Link = identity
##
## Estimated Correlation Parameters:
##       Estimate Std.err
## alpha    0.995 0.00822
## Number of clusters:   100   Maximum cluster size: 12
```

```
##   Model      QIC     QICu   LQLik Trace px dQIC RelQLik   QICwt Cum.Wt
## 1    2 1 -6642080 -6642134 3321071  30.5  4  0.0 1.00000 0.99388  0.994
## 2    3 2 -6642070 -6642122 3321065  30.0  4 10.2 0.00615 0.00612  1.000
```