**SURVMETH 687 / SURV 617**

Applications of Statistical Modeling

Fall 2024 Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

# Midterm Examination

**IMPORTANT: This exam is scheduled for two hours (1pm-3pm). Read all questions and descriptions carefully and make sure to answer all parts of each question. Use the backsides of each sheet if necessary. Please contact Dr. Si immediately if you need clarification on any of the problems (845-798-2013 /** [**yajuan@umich.edu**](mailto:yajuan@umich.edu)**).**

**The exam has 10 pages and 14 questions in total.**

We analyze a sample of science exam scores of 1523 students from 73 schools in the U.S. The dataset consists of the following 6 variables:

* school: school identifier
* student: student identifier
* female: indicator of the sex of a student (0: Male; 1: Female)
* type: indicator of the school type (0: Private; 1: Public)
* written: total score on the written report
* result: final evaluation of the science subject (0: Unsatisfactory; 1: Satisfactory)

We would like to examine sexual differences in the exam performance and whether school types contribute to different evaluation results. We address the objectives with the following analyses.

1. **Analysis of total score on the written report via marginal models.**

We fit a marginal linear model with a pre-specified compound symmetry structure in the variance-covariance matrix to account for the correlation between students’ writing scores from the same school. The model includes sex of the student, school type, and their two-way interaction as predictors.

gls(written ~ type + female + type \* female,   
 correlation = corCompSymm(form = ~ 1 | school),  
 data = exam  
 )

The output is shown below.

Generalized least squares fit by REML  
 Model: written ~ type + female + type \* female   
 Data: exam   
 AIC BIC logLik  
 11824.03 11855.99 -5906.016  
   
 Correlation Structure: Compound symmetry  
 Formula: ~1 | school   
 Parameter estimate(s):  
 Rho   
 0.2852153   
   
 Coefficients:  
 Value Std.Error t-value p-value  
 (Intercept) 49.32727 1.0990124 44.88327 0.0000  
 type -0.27810 0.9199879 -0.30229 0.7625  
 female -2.51813 0.8756798 -2.87563 0.0041  
 type:female 0.13391 1.1909722 0.11244 0.9105  
   
 Correlation:   
 (Intr) type female  
 type -0.457   
 female -0.477 0.558   
 type:female 0.355 -0.771 -0.727  
   
 Standardized residuals:  
 Min Q1 Med Q3 Max   
 -3.4947480 -0.8062622 -0.1003404 0.6192233 3.2651981   
   
 Residual standard error: 13.22764   
 Degrees of freedom: 1523 total; 1519 residual

Answer the following Part A questions:

1. What is the estimated correlation between students’ writing scores from the same school? [2 Points]

**Rho   
 0.2852153**

1. Write the elements on the diagonal and off the diagonal of the estimated variance-covariance matrix. [4 points]

**The compound symmetry structure has the same element values on the diagonal (13.22764^2 = 174.97) and the same element values off the diagonal (13.22764^2 \* 0.2852153=49.90)**

1. Interpret the estimated intercept and THREE coefficients in ***practical terms***. [12 Points]

**49.33 -0.27810 \* type -2.51813\* female + 0.13391 \* type: female**

**= 49.33 -0.27810 \* type + (0.13391 \* type -2.51813)\* female**

**= 49.33 + (0.13391 \*female -0.27810) \* type -2.51813\* female**

**Intercept: The average writing score of male students in private schools is 49.33.**

**Coef of type: The average writing score of male students in public schools is 0.28 lower than that of male students in private schools.**

**For male students enrolled in public schools, their average writing score is 0.28 lower than those in private schools.**

**Coef of female: The average writing score of female students in private schools is 2.52 lower than that of male students in private schools.**

**In private schools, the average writing score of female students is 2.52 lower than that of male students.**

**Coef of the interaction of type and female: The sexual disparity of writing scores in public schools is 0.13 larger than that in private schools.**

**For female students enrolled in public schools, their average writing score is 0.14 [-(0.13391 \*female -0.27810)] lower than those in private schools.**

**In public schools, the average writing score of female students is 2.38 [-(0.13391 \* type -2.51813)] lower than that of male students.**

1. **Analysis of total score on the written report via multilevel models.**

To estimate the variance of students’ writing scores between schools, we fit a multilevel model with a random intercept that varies across schools. The model includes sex of the student, school type, and their two-way interaction as predictors.

lmer(  
 written ~ type + female + type \* female + (1| school),   
 data = exam,  
 REML = TRUE  
)

The output is shown below.

Linear mixed model fit by REML ['lmerMod']  
 Formula: written ~ type + female + type \* female + (1 | school)  
 Data: exam  
   
 REML criterion at convergence: 11811  
   
 Scaled residuals:   
 Min 1Q Median 3Q Max   
 -3.1314 -0.6707 -0.0693 0.6574 3.8596   
   
 Random effects:  
 Groups Name Variance Std.Dev.  
 school (Intercept) 49.9 7.064   
 Residual 125.1 11.183   
 Number of obs: 1523, groups: school, 73  
   
 Fixed effects:  
 Estimate Std. Error t value  
 (Intercept) 49.3273 1.0990 44.883  
 type -0.2781 0.9200 -0.302  
 female -2.5181 0.8757 -2.876  
 type:female 0.1339 1.1910 0.112  
   
 Correlation of Fixed Effects:  
 (Intr) type female  
 type -0.457   
 female -0.477 0.558   
 type:female 0.355 -0.771 -0.727

Answer the following Part B questions:

1. Write the model in multilevel form inserting actual estimates of parameters, including estimated variance and covariance terms and assumed distributions. [20 Points]

Assume the total score of the writing report is denoted by for student enrolled in school . Let denote the random intercept varying schools and denote the random error of the scores.

1. What is the estimated intra-school correlation? [2 points]
2. Compare the estimated fixed effects to the coefficient estimates in **Part A**. Describe your findings and provide the justification. [4 Points]

**They are almost identical. The multilevel model for a normal outcome with only random intercepts is equivalent to the marginal model with a compound-symmetry structure.**

1. We fit a new multilevel model with a random intercept and a random coefficient of sex, both of which vary across schools. The model includes sex of the student, school type, and their two-way interaction as predictors.

lmer(  
 written ~ type + female + type \* female + (1 + female | school),   
 data = exam,  
 REML = TRUE  
)

The output is shown below.

Linear mixed model fit by REML ['lmerMod']  
 Formula: written ~ type + female + type \* female + (1 + female | school)  
 Data: exam  
   
 REML criterion at convergence: 11811  
   
 Scaled residuals:   
 Min 1Q Median 3Q Max   
 -3.0958 -0.6876 -0.0621 0.6549 3.8427   
   
 Random effects:  
 Groups Name Variance Std.Dev. Corr   
 school (Intercept) 52.583 7.251   
 female 3.896 1.974 -0.24  
 Residual 124.187 11.144   
 Number of obs: 1523, groups: school, 73  
   
 Fixed effects:  
 Estimate Std. Error t value  
 (Intercept) 49.3638 1.1224 43.981  
 type -0.2864 0.9227 -0.310  
 female -2.5925 0.9225 -2.810  
 type:female 0.1704 1.1924 0.143  
   
 Correlation of Fixed Effects:  
 (Intr) type female  
 type -0.451   
 female -0.509 0.535   
 type:female 0.350 -0.772 -0.692

Shall we include the random coefficient in the model? Explain in detail with the **procedure and computed statistics** to support your decision. [8 points]

**We conduct the approximate likelihood ratio test to compare the model with both random intercepts and coefficients to that with only random intercepts.**

**Step 0: The two models are fit.**

**Step 1: We calculate the test statistic, which is the difference between the two REML criterion values (-2 \* log-likelihood): 11811 - 11811 = 0.**

**Step 2: We use the reference distribution, which is a mixture of two Chi-square distributions.**

**Step 3: Since the test statistics is 0, the p-value will become 1. We cannot reject the null hypothesis that the variance of the random coefficients is 0, and we recommend not including the random coefficients.**

1. **Analysis of the final evaluation result of the science subject via generalized linear mixed effect models.**

To estimate the variation of students’ final evaluation results between schools, we fit a generalized linear mixed effect model with a random intercept that varies across schools. The model includes sex of the student, school type, and their two-way interaction as predictors.

glmer(result ~ type + female + type \* female + (1 | school),  
 family = binomial,  
 data = exam)

The output is shown below.

Generalized linear mixed model fit by maximum likelihood (Laplace  
 Approximation) [glmerMod]  
 Family: binomial ( logit )  
 Formula: result ~ type + female + type \* female + (1 | school)  
 Data: exam  
   
 AIC BIC logLik deviance df.resid   
 1739.4 1766.1 -864.7 1729.4 1518   
   
 Scaled residuals:   
 Min 1Q Median 3Q Max   
 -3.5503 -0.7227 0.3685 0.6240 2.2153   
   
 Random effects:  
 Groups Name Variance Std.Dev.  
 school (Intercept) 1.67 1.292   
 Number of obs: 1523, groups: school, 73  
   
 Fixed effects:  
 Estimate Std. Error z value Pr(>|z|)   
 (Intercept) 0.2665 0.2077 1.283 0.200   
 type -0.1321 0.1869 -0.707 0.480   
 female 0.8336 0.1780 4.683 2.82e-06 \*\*\*  
 type:female 0.1263 0.2486 0.508 0.611   
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 Correlation of Fixed Effects:  
 (Intr) type female  
 type -0.442   
 female -0.470 0.519   
 type:female 0.333 -0.751 -0.689

Answer the following Part C questions:

1. Write the generalized linear mixed effect model inserting actual estimates of parameters, including estimated variance and covariance terms and assumed distributions. [16 Points]

**Assume the result of the science subject is denoted by for student enrolled in school where indicates satisfactory results; otherwise. Let denote the random intercept varying schools.**

1. Interpret the estimated intercept and THREE coefficients in ***practical terms***. [12 Points]

**Intercept: For male students in a given private school, the overall log odds of obtaining satisfactory science reports is 0.27.**

**For male students in a given private school, if the school-level random effect is predicted as 0.13, the probability of obtaining satisfactory science reports is 0.60 [exp(0.27 + 0.13)/(1+ exp(0.27 + 0.13))].**

**Coef of type: For male students in a given school, the odds of obtaining satisfactory science reports would have reduced by 12.2% [1- exp(-0.13)] if the school were switched from being private to public.**

**Coef of female: In a given private school, the odds of obtaining satisfactory science reports for female students are 2.29 [exp(0.83)] times of those for males.**

**Coef of the interaction of type and female:**

**In a given public school, the odds of obtaining satisfactory science reports for female students are 2.61 [exp(0.83 + 0.13 \* type)] times of those for males.**

1. Can we use the analysis results to assess whether overall female students have similar final evaluation performances to those of males? If so, provide the statistical evidence. If not, provide the reasoning. [4 Points]

**No. The inference is restricted to a given school, at the school-specific level. We cannot use it to compare overall female students to male students.**

1. **Analysis of the final evaluation result of the science subject via generalized estimation equation.**

To assess the contribution of school characteristics to students’ final evaluation performance, we fit a generalized estimation equation model with an exchangeable working correlation structure. The model includes sex of the student, school type, and their two-way interaction as predictors.

geeglm(result ~ type + female + type \*female,  
 id = school,  
 waves = student,  
 family = binomial("logit"),  
 data = exam,  
 corstr = "exchangeable")

The output is shown as below.

Call:  
geeglm(formula = result ~ type + female + type \* female, family = binomial("logit"), data = exam, id = school, waves = student,

corstr = "exchangeable")  
   
 Coefficients:  
 Estimate Std.err Wald Pr(>|W|)   
 (Intercept) 0.2077 0.1696 1.500 0.220675   
 type -0.1086 0.1331 0.666 0.414449   
 female 0.6345 0.1673 14.388 0.000149 \*\*\*  
 type:female 0.1019 0.1909 0.285 0.593453   
 ---  
 Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
   
 Correlation structure = exchangeable   
 Estimated Scale Parameters:  
   
 Estimate Std.err  
 (Intercept) 1.017 0.06211  
 Link = identity   
   
 Estimated Correlation Parameters:  
 Estimate Std.err  
 alpha 0.2482 0.04775  
 Number of clusters: 73 Maximum cluster size: 83

Answer the following Part D questions:

1. Describe how you choose different working correlation structures when using GEE. [2 Points]

**We can fit GEE models with different working correlation structures and compare them based on the QIC values. The model with smaller QIC value will be preferred.**

1. Is the following statement TRUE or FASE? Provide justification.
   * When using GEE, the choice of the working correlation structure does not matter for the consistency of the estimated coefficients. [3 Points]

**TRUE. The estimated coefficients with GEE are consistent even when the working correlation structure is misspecified.**

1. Can we use the analysis results to assess whether students enrolled in public schools have reduced sexual disparity in the likelihood of having satisfactory final evaluation performances than those in private schools? If so, provide the statistical evidence. If not, provide the reasoning. [6 Points]

**YES. The GEE results are on the population level, as a marginal model. We can check the coefficient estimates of the interaction term between type and female.**

**They are not significantly different from 0.**

1. Compare the estimated coefficient estimates to those in **Part C**. Describe your findings and provide the justification. [5 Points]

**They are different. The GEE inference is on the population average level, while the generalized mixed effect model gives subject-level inference results.**