Biostatistics 140.656, 2018-19
Lab 3 SOLUTION

Topics:

- Logistic regression models for individual outcomes vs. aggregated outcomes
- Interpretation of parameters from logistic random intercept models
- Interpretation of parameters from marginal logistic regression models

Learning Objectives:

Students who successfully complete this lab will be able to:

- Interpret parameters for level-1 and level-2 covariates within a logistic random intercept model
- Interpret parameters for level-1 and level-2 covariates within a marginal logistic regression model
- Describe and implement a logistic regression model for aggregated level-1 data

Scientific Background:

In 2003, the Maryland General Assembly enacted the Public Charter School Act. The act created Maryland's first public charter school program "to establish an alternative means within the existing public school system in order to provide innovative learning opportunities and creative educational approaches to improve the education of students." Charter schools in Maryland are public schools where admission is based on a lottery that is open to all families within the specific county or Baltimore City. Since the passage of the Public Charter School Act in 2003, roughly 35 charter schools have opened and are operating in Maryland; the vast majority of charter schools are in Baltimore City.

There is on-going debate about whether charter schools add value above what students receive at traditional schools; i.e. the child's neighborhood public school.

In Homework 2, you will evaluate whether students in charter schools perform better academically than students in traditional schools within Baltimore City.

The Maryland School Assessment (MSA) is an annual assessment program that tests students in grades 3 through 8 in reading and mathematics (http://reportcard.msde.maryland.gov/). The MSA program ranks student performance with respect to meeting expectations on an ordinal scale: Level 1: did not yet meet expectations, Level 2: partially met expectations, Level 3: approached expectations, Level 4: met expectations, and Level 5: exceeded expectations.

We will consider a good outcome for student performance a "pass" as defined by *Level 3* through *Level 5*.

The data has two levels: i for school (i = 1, .., 121) and j for student within school i (j = 1, …, n_i). The outcome is Y_{ij}, the student level indicator of "pass" on the mathematics MSA. The primary exposure variables of interest are grade level (X_{ij}=3,4,5, a student-level variable) and school type (Z_i =0,1 for traditional (0) vs. charter (1) school).

NOTE: The indicator for charter vs. traditional school was generated by us using data from Baltimore City Schools and the Maryland Association of Public Charter Schools.

NOTE: There are more than 121 elementary schools in Baltimore City; we have excluded a few schools that did not report MSA data for academic year 2017-2018 or whom we could not identify as either a traditional vs. charter school.

In this lab session, you will become familiar with the data structure and focus on comparing the proportion of students who "pass" across the grade levels (ignoring school type) and then separately among charter vs. traditional schools (ignoring grade levels).

Lab exercises:

PART I: Idealized data: I have created a dataset that provides (Y_{ij}, X_{ij}, Z_i) for the 121 elementary schools included in the analysis. Download "MSA2017_individual.csv".

1. Open the dataset and do the following:
    a. Confirm that there are 121 schools in the dataset

```
-------------------------------------------------------------------------------------------------
school_number
-------------------------------------------------------------------------------------------------
type:  numeric (int)

                range:  [4,384]                     units:  1
         unique values:  121                    missing .:  0/17,795

                 mean:   170.468
              std. dev:   106.431

          percentiles:        10%       25%       50%       75%       90%
                               27        64       207       242       327
```

    b. Confirm that there are 24 charter schools

```
bys school_number: gen within_school_counter = _n

tab charter if within_school_counter==1

    charter |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         97       80.17       80.17
          1 |         24       19.83      100.00
------------+-----------------------------------
      Total |        121      100.00
```

    c. Compute the number of students who completed the mathematics MSA for each grade level within each school. For each grade level, summarize the number of students taking the mathematics MSA across the schools

```
bys school_number grade: gen num_students = _N

bys school_number grade: gen within_grade_counter = _n

bys grade: summ num_students if within_grade_counter==1
```



```
-------------------------------------------------------------------------------------------------------
-> grade = Grade 3

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
num_students |        118    50.84746    24.19318         14        134

-------------------------------------------------------------------------------------------------------
-> grade = Grade 4

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
num_students |        120    50.83333    23.84943         13        127

-------------------------------------------------------------------------------------------------------
-> grade = Grade 5

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
num_students |        121    47.06612    23.10654         11        118
```

d. Compute the proportion of students who "passed" the mathematics MSA for each grade level within each school. For each grade level, summarize the proportion who passed across the schools

```
bys school_number grade: egen prop_pass = mean(pass)

bys grade: summ prop_pass if within_grade_counter==1

-------------------------------------------------------------------------------------------------------
-> grade = Grade 3

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   prop_pass |        118    .3848063    .2034303          0   .9310345

-------------------------------------------------------------------------------------------------------
-> grade = Grade 4

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   prop_pass |        120    .3169156    .1993105          0   .9038461

-------------------------------------------------------------------------------------------------------
-> grade = Grade 5

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
   prop_pass |        121    .3383254    .2125012          0          1
```

2. Fit the model below and answer the questions that follow. NOTE: We will assume no contextual effect of grade level.

$$Y_{ij} \sim Bernoulli(p_{ij}), where\ p_{ij} = \Pr(Y_{ij} = 1 | b_{0i}, X_{ij})$$
$$Logit\left(\Pr(Y_{ij} = 1 | b_{0i}, X_{ij})\right) = \beta_0 + b_{0i} + \beta_1 I(X_{ij} = 4) + \beta_2 I(X_{ij} = 5)$$
$$b_{0i} \sim N(0, \tau^2)$$

HINT: Stata users, you can use meqrlogit as follows:

```
gen grade4 = grade=="Grade 4"
gen grade5 = grade=="Grade 5"
meqrlogit pass grade4 grade5 || school_number:
```

HINT: R users, you can use glmer within the lme4 package as follows:

```
library(lme4)
data$grade4 = data$grade=="Grade 4"
data$grade5 = data$grade=="Grade 5"
fit = glmer(pass~grade4+grade5+(1|school_number),data=data,family="binomial",nAGQ = 7)
summary(fit)
```

```
Mixed-effects logistic regression               Number of obs     =     17,795
Group variable: school_number                   Number of groups  =        121

                                                Obs per group:
                                                              min =         13
                                                              avg =      147.1
                                                              max =        372

Integration points =   7                        Wald chi2(2)      =      67.69
Log likelihood = -10449.188                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grade4 |  -.3378649   .0412615    -8.19   0.000    -.4187361   -.2569938
      grade5 |  -.1945319   .0416384    -4.67   0.000    -.2761417    -.112922
       _cons |   -.564165    .088361    -6.38   0.000    -.7373494   -.3909806
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school_num~r: Identity       |
                  var(_cons) |   .8273499   .1160835      .6284333    1.089229
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 2338.27     Prob >= chibar2 = 0.0000
```

a. Interpret the value of $\exp(\beta_0)$.

The estimate of $\exp(\beta_0)$ is exp(-0.564165) = .5688349.
The odds of passing the mathematics MSA for 3rd graders in the average school (i.e. set b_{0i}=0).

b. Interpret the value of $\exp(\beta_1)$.

The estimate of $\exp(\beta_1)$ is exp(-0.3378649) = .7132916

For any given school, the odds of passing the mathematics MSA for a 4th grader are roughly 29% lower than the odds of passing for a 3rd grader.

c. Provide an estimate of the intraclass correlation coefficient, i.e. $Corr(Y_{ij}, Y_{ik})$. Provide an interpretation of this statistic within the context of the problem.

The intraclass correlation coefficient can be computed as: $\frac{.8273499}{.8273499+\frac{\pi^2}{3}} = 0.20$

This can be interpreted as the correlation between the binary indicator of passing the mathematics MSA for any two students from the same school.

OR

The proportion of the total variation in the binary indicator of passing the mathematics MSA that is explained by heterogeneity between schools.

d. From this model, you can compute the proportion of 5th graders who pass the mathematics MSA for each school. Compute an interval that contains the proportion of 5th graders who pass the mathematics MSA for 95% of all schools.

For the average school (b_{0i}=0), the estimated log odds of passing the mathematics MSA is $\hat{\beta}_0+\hat{\beta}_2$ = -.564165 + (-.1945319) = -.7586969.

An interval that contains the log odds of passing the mathematics MSA for 5th graders will be given by: -.7586969 + / - 1.96 x sqrt(.8273499), which is -2.54 to 1.02.

Converting this interval to the probability/proportion scale:

$\frac{\exp(-2.54)}{1+\exp(-2.54)}$ to $\frac{\exp(1.02)}{1+\exp(1.02)}$

0.07 to 0.73

e. Evaluate the estimation procedure by refitting the model using a smaller/larger number of integration points. In Stata, the default number of integration points for meqrlogit is 7. If you used the glmer command provided above, I set the number of integration points to 7. Set the number of integration points to 4 and 14 and make a conclusion about the stability/convergence of the results of your model.

The key model results based on 4 integration points:

```
------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grade4 |  -.3378649   .0412615    -8.19   0.000     -.418736   -.2569938
      grade5 |   -.194532   .0416384    -4.67   0.000    -.2761418   -.1129221
       _cons |  -.5641665   .0883597    -6.38   0.000    -.7373484   -.3909847
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school_num~r: Identity       |
                  var(_cons) |   .8273202   .1160757      .6284161    1.089181
------------------------------------------------------------------------------
```

The key model results based on 14 integration points:

```
------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grade4 |  -.3378649   .0412615    -8.19   0.000    -.4187361   -.2569938
      grade5 |  -.1945319   .0416384    -4.67   0.000    -.2761417    -.112922
       _cons |   -.564165   .0883611    -6.38   0.000    -.7373495   -.3909804
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school_num~r: Identity       |
                  var(_cons) |   .8273518   .1160841      .6284342    1.089233
------------------------------------------------------------------------------
```

The results of the model fit based on 7 integration points are essentially the same as the results based on 4 and 14 integration points. Given the stability in the results, we are confident in our results.

3. Fit the following model below and answer the questions that follow.

$$Y_{ij} \sim Bernoulli(p_{ij}), where\ p_{ij} = \Pr(Y_{ij} = 1 | a_{0i}, Z_i)$$
$$Logit\left(\Pr\left(Y_{ij} = 1 | a_{0i}, Z_i\right)\right) = \alpha_0 + a_{0i} + \alpha_1 Z_i$$
$$a_{0i} \sim N(0, \sigma^2)$$

Using Stata to fit the model:

```
meqrlogit pass charter || school_number:

Mixed-effects logistic regression               Number of obs     =     17,795
Group variable: school_number                   Number of groups  =        121

                                                Obs per group:
                                                              min =         13
                                                              avg =      147.1
                                                              max =        372

Integration points =   7                        Wald chi2(1)      =       2.42
Log likelihood = -10481.922                     Prob > chi2       =     0.1195

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     charter |   .3285578   .2110304     1.56   0.119    -.0850542    .7421697
       _cons |  -.8045252    .093923    -8.57   0.000    -.9886109   -.6204394
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school_num~r: Identity       |
                  var(_cons) |   .8042139   .1129891      .6106338    1.059162
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 2323.06     Prob >= chibar2 = 0.0000
```

a. Interpret the value of $\exp(\alpha_0)$.

The estimated value of exp(α_0) is exp(-.8045252) = .4473003.

For the average school traditional school (i.e. $a_{0i} = 0\ and\ Z_i = 0$), the odds of passing the mathematics MSA is 0.4473.

b. Interpret the value of exp(α_1).

The estimated value of exp(α_1) is exp(.3285578) = 1.388963.

Comparing two schools with the same random effect (i.e. a_{0i}) but differ in charter school status, the odds of a student from the charter school passing the mathematics MSA is 39 percent greater than the odds of a student from the traditional school passing.

c. Provide an interval that contains the proportion of students who pass the mathematics MSA for 95% of the charter schools.

First, we compute the 95% interval for the log odds of passing among students from charter schools: $\alpha_0 + \alpha_1$ +/- 1.96 σ, which is -2.23 to 1.28

Then converting the interval to the probability scale: 9.7% to 78.2%

d. As an alternative to the logistic random intercept model, fit the following marginal model. Compare the estimate of exp(γ_1) to exp(α_1). Describe the reason for any differences.

$$Y_{ij} \sim Bernoulli(p_{ij}), where\ p_{ij} = \Pr(Y_{ij} = 1 | Z_i)$$
$$Logit(\Pr(Y_{ij} = 1|Z_i)) = \gamma_0 + \gamma_1 Z_i$$
$$Corr(Y_{ij}, Y_{ik}) = \rho$$

HINT: In Stata, this could be accomplished by:

```
xtset school_number
xtgee pass charter, family(binomial) corr(exch)
```

HINT: In R, this can be accomplished by:

```
library(geepack)
fit = geeglm(pass~charter,family="binomial",corstr="exchangeable",data=data,
id=data$school_number)
```



```
xtgee pass charter, family(binomial) corr(exch)

GEE population-averaged model                   Number of obs     =     17,795
Group variable:               school_num~r      Number of groups  =        121
Link:                                logit      Obs per group:
Family:                         binomial                    min =         13
Correlation:                  exchangeable                  avg =      147.1
                                                            max =        372
                                                Wald chi2(1)      =       2.19
Scale parameter:                         1      Prob > chi2       =     0.1393

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     charter |   .2916692   .1973071     1.48   0.139    -.0950456    .6783839
       _cons |  -.6970228   .0903111    -7.72   0.000    -.8740293   -.5200163
------------------------------------------------------------------------------
```

The estimate for $\exp(\gamma_1)$ is exp(0.2916692) = 1.33866
The estimate for $\exp(\alpha_1)$ is exp(.3285578) = 1.388963
We expect that the estimate for parameter associated with the random intercept model would be more positive.

The marginal parameter can be interpreted as: The odds of passing the mathematics MSA for a student attending a Baltimore City charter school is roughly 34 percent greater than the odds of passing for a student attending a traditional school in Baltimore City.

Part 2: Actual data: The MSA program makes the results of MSA testing for each public school in Maryland publicly available on its website. However, the MSA program aggregates data by grade level and school, in other words, instead of a dataset with a row per child per grade level per school, i.e. (Y_{ij}, X_{ij}), the public is provided with a row of data per grade level per school.

Recall, Y_{ij} is the student level indicator of "pass" on the mathematics MSA for student j from school *i*, i = 1, .., 121 and j = 1, …, n_i. The primary level-1 exposure variable is grade level, X_{ij}=3,4,5.

For each school *i* and grade level *k* = 3, 4 and 5 respectively,

$N_{ik} = \sum_{j=1,n_i} I(X_{ij} = k)$ is the number of students in grade *k* from school *i* who took the mathematics MSA

$T_{ik} = \sum_{j=1,n_i} Y_{ij} \times I(X_{ij} = k)$ is the number of students who "pass" the mathematics MSA in grade *k* from school *i*

Lastly, define $G_{ik} = k$, a grade variable with values 3, 4 and 5.

Here is a listing of the selected variables from the first 7 rows of available data. Notice that not all schools will have scores available for all grade levels. In the dataset, G_{ik} is

"grade", N_{ik} is "tested_count", T_{ik} is "pass" and Z_i is "charter" (the school level charter school indicator).

```
list school_number school_name tested_count grade pass charter in 1/7
```

	school~r	school_name	grade	tested~t	pass	charter
1.	314	SharpLeadenhall Elementary	Grade 5	11	1	0
2.	314	SharpLeadenhall Elementary	Grade 4	13	1	0
3.	371	Lillie May Carroll Jackson School	Grade 5	13	3	1
4.	322	New Song Academy	Grade 5	14	5	0
5.	89	Rognel Heights ElementaryMiddle	Grade 3	14	3	0
6.	322	New Song Academy	Grade 4	15	6	0
7.	379	Roots and Branches School	Grade 5	15	1	1

Given that we have this aggregated school-level and grade-level data, how should you fit the models?

Without losing any information, the model for the aggregated data can be expressed as:

$$T_{ik} \sim Binomial(N_{ik}, p_{ik}), where\ p_{ik} = \Pr(Y_{ij} = 1 | b_{0i}, X_{ij} = k)$$
$$Logit(\Pr(Y_{ij} = 1 | b_{0i}, X_{ij} = k)) = \beta_0 + b_{0i} + \beta_1 I(X_{ij} = 4) + \beta_2 I(X_{ij} = 5)$$
$$b_{0i} \sim N(0, \tau^2)$$

So, when we are fitting the models, we will need to specify both the number of students who "pass" (this is our outcome) and the number of students who took the test for a given grade level (this is "tested_count" in the dataset).

Download the "HW2 MSA 2018.csv" and fit the model above using the commands provided below. Confirm your results are the same using the individual level data (see Question 2) and the aggregated data.

Stata users:

```
gen grade4 = grade==”Grade 4”
gen grade5 = grade==”Grade 5”
meqrlogit pass grade4 grade5 || school_num: , binomial(tested_count)
```

R users:

```
data$grade4 = ifelse(data$Grade==”Grade 4”,1,0)
data$grade5 = ifelse(data$Grade==”Grade 5”,1,0)
fit = glmer(cbind(pass,Tested_Count-pass)
~grade4+grade5+(1|School_Number),data=data,family="binomial",nAGQ=7)
summary(fit)
```


You can compare the results below to your model output in Question 2 from Part I.

```
meqrlogit pass grade4 grade5 || school_number: , binomial(tested_count)


Mixed-effects logistic regression               Number of obs     =        359
Binomial variable: tested_count
Group variable: school_number                   Number of groups  =        121

                                                Obs per group:
                                                              min =          1
                                                              avg =        3.0
                                                              max =          3

Integration points =   7                        Wald chi2(2)      =      67.69
Log likelihood =  -1251.758                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
        pass |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      grade4 |  -.3378649   .0412615    -8.19   0.000     -.418736   -.2569938
      grade5 |  -.1945318   .0416384    -4.67   0.000    -.2761417    -.112922
       _cons |   -.564165    .088361    -6.38   0.000    -.7373493   -.3909807
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
school_num~r: Identity       |
                  var(_cons) |   .8273485   .1160833      .6284322    1.089227
------------------------------------------------------------------------------
LR test vs. logistic model: chibar2(01) = 2338.27     Prob >= chibar2 = 0.0000
```