Biostatistics 140.656, 2018-19
Lab 4 Solution (Partial)

Topics:

- Mixed logistic regression models for hierarchical data with more than 2 levels
- Interpretation of parameters from mixed logistic regression model with multiple random intercepts
- Interpretation of parameters from mixed logistic regression model with random slopes

Learning Objectives:

Students who successfully complete this lab will be able to:
- Explore data with 3 levels of nesting
- Interpret coefficients for a level 1 exposure in a mixed logistic model with random intercepts defined at level 2 and 3
- Interpret results from a mixed logistic model including multiple random intercept and random slope

Scientific Background:

You will be analyzing data from a survey conducted in Guatemala in 1987. The survey identified a nationally representative sample of 5160 mothers, between 15 and 44 years of age, with the primary purpose of understanding factors that could affect the immunization status of children who were born in the previous 5 years and alive at the time of the interview.

The data available represent 2,158 children (level 1) aged 1 – 4 years of age from 1595 mothers (level 2) from within 161 communities (level 3). The 2,158 children received at least one immunization. The outcome of interest is whether the child received the full set of immunizations.

Starting in 1986, the government of Guatemala undertook a series of campaigns to immunize the population against major childhood illnesses. The immunization campaign visited most of the country and often located children in their own households. The full set of immunizations at the time of the campaign included three doses of DPT vaccine (against diphtheria, whooping cough, and tetanus), three doses of polio vaccine, one dose of BCG (antituberculosis) and one dose of the measles vaccine.

The survey conducted in 1987 offers an opportunity to evaluate the likelihood of children receiving the full set of immunization during both absence (pre-1986) vs. presence (1986-7) of the campaign (i.e. a natural experiment). An important variable is whether the child was at least 2 years old at the time of the survey/interview, in which case the child was eligible to receive all immunizations during the campaign. If this variable is associated with immunization status, there is some indication that the campaign worked.

In your final exam, you will be conducting a series of analyses to address questions that are relevant to government health officials both from the prospective of improving the immunization coverage rate but also to explore factors related to differences in the odds of full immunization that could provide insight into how to improve the design and implementation of future campaigns.

In this lab, you will:

1. Summarize the three-level nested structure of the data
2. Fit and interpret a random intercept only logistic regression model within this three-level nested data example
3. Estimate the effect of the immunization campaign within a random effects logistic regression model
4. Estimate the heterogeneity of the effect of the immunization campaign across the Guatemalan communities and determine if community characteristics explain observed heterogeneity.

<u>Data:</u>

The data set is called **guatemala.csv,** which can be downloaded directly from our website. The dataset contains children *i* nested in mother *j* nested in community *k*. It contains the following subset of variables.

Level 1 (children)

- immun: dummy variable for child being immunized, the response variable.
- kid2p: child at least 2 years old at the time of the interview (indicator for exposure to the campaign or not)

Level 2 (mother)

- mom: identifier for mother
- Ind: Indigenous Ethnicity (indigenous vs. not)

Level 3 (community)

- cluster: identifier for communities
- rural: dummy variable for community being rural
- pcInd: percent of indigenous mothers in the community (ranges from 0 to 1)

Brief EDA of the hierarchical clustering:

How many communities are in the study, how many mothers and how many children?

```
. codebook cluster mom kid

-------------------------------------------------------------------------------------------
cluster                                                                         (unlabeled)
-------------------------------------------------------------------------------------------

                 type:  numeric (int)

                range:  [1,240]                    units:  1
        unique values:  161                    missing .:  0/2,158

                 mean:   145.858
             std. dev:   59.3406

          percentiles:        10%       25%       50%       75%       90%
                               63        94       148       202       226

-------------------------------------------------------------------------------------------
mom                                                                             (unlabeled)
-------------------------------------------------------------------------------------------

                 type:  numeric (int)

                range:  [2,2782]                   units:  1
        unique values:  1,595                  missing .:  0/2,158

                 mean:   1502.63
             std. dev:   751.435

          percentiles:        10%       25%       50%       75%       90%
                              498       859    1471.5      2208      2571

-------------------------------------------------------------------------------------------
kid                                                                             (unlabeled)
-------------------------------------------------------------------------------------------

                 type:  numeric (int)

                range:  [2,4627]                   units:  1
        unique values:  2,158                  missing .:  0/2,158

                 mean:   2445.98
             std. dev:   1267.71

          percentiles:        10%       25%       50%       75%       90%
                              775      1346    2373.5      3636      4255
```

We have 161 communities; 1595 mothers and 2158 children.

How many mothers in each community?

```
bys cluster mom: gen kid_counter = _n
bys cluster: egen junk = count(mom) if kid_counter==1
bys cluster: egen num_moms = min(junk)
drop junk
bys cluster: gen cluster_counter = _n
summ num_moms if cluster_counter==1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
num_moms |       161    9.906832    6.022667          1         37
```

The average number of mothers per community was roughly 10 with a range from 1 to 37.

How many children in each community?

```
. bys cluster: egen cluster_kids = count(cluster)
. summ cluster_kids if cluster_counter==1


    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
cluster_kids |       161    13.40373    8.750271          1         55
```

The number of children per community ranges from 1 to 55. Average number of children per community is 13.

How many children per mother?

```
. tab num_kids if kid_counter==1

   num_kids |      Freq.     Percent        Cum.
------------+-----------------------------------
          1 |      1,063       66.65       66.65
          2 |        501       31.41       98.06
          3 |         31        1.94      100.00
------------+-----------------------------------
      Total |      1,595      100.00
```

The number of children per mother ranges from 1 to 3 with 67% of the mothers contributing data from a single child.

<u>Brief EDA of the primary outcome</u>

What is the sample prevalence of immunization?

```
. summ immun


    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       immun |     2,158    .4467099    .4972673          0          1
```

What is the sample prevalence of immunization by study period (during or post campaign)?

```
-----------------------------------------------------------------------
-> kid2p = 0

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       immun |       492    .2845528    .4516604          0          1

-----------------------------------------------------------------------
-> kid2p = 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       immun |     1,666    .4945978    .5001209          0          1
```

NOTE: I will be providing you with all the necessary regression output relating to the mixed logistic regression models I want you to consider. Most of these regression models take a bit of time to run. To save you this time, I will provide the output and your goal is to focus on the interpretation.

1. Three-level random intercepts only model

To start we will ignore all the covariates in the model and simply decompose the variation in log odds of receiving the full course of immunizations into that attributable to differences across mothers within a given community and across communities.

Let Y_{ijk} be the indicator for completing the full course of immunizations for kid *k* from mom *j* within community *i.* Then the random intercept only model is given by:

$$log\left(\frac{Pr(y_{ijk}=1)}{1-Pr(y_{ijk}=1)}\right)=\beta_0+b_i+b_{ij},\ \ b_i\sim N(0,\sigma^2), b_{ij}\sim N(0,\tau^2), Cov\left(b_i,b_{ij}\right)=0$$

In the model above, we allow each child to have his/her own log odds of receiving the full course of immunizations which depends on their mother and community membership.

The fit of this model is given below:

```
. meqrlogit immun || cluster: || mom: , intp(12)


Mixed-effects logistic regression               Number of obs     =      2,158

-----------------------------------------------------------------------------
                |     No. of       Observations per Group       Integration
 Group Variable |     Groups    Minimum    Average    Maximum      Points
----------------+------------------------------------------------------------
        cluster |        161          1       13.4         55          12
            mom |      1,595          1        1.4          3          12
-----------------------------------------------------------------------------

                                                Wald chi2(0)      =          .
Log likelihood = -1396.5023                     Prob > chi2       =          .

-------------------------------------------------------------------------------
       immun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+-----------------------------------------------------------------
       _cons |  -.3733179   .1301502    -2.87   0.004    -.6284075   -.1182283
-------------------------------------------------------------------------------

-------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+-------------------------------------------------
cluster: Identity            |
                  var(_cons) |   1.329794   .3547759      .7883008    2.243244
-----------------------------+-------------------------------------------------
mom: Identity                |
                  var(_cons) |    4.14361   .9381624      2.658623    6.458044
-------------------------------------------------------------------------------
LR test vs. logistic model: chi2(2) = 174.06              Prob > chi2 = 0.0000
```

a. Interpret the estimated value of β_0 and $\exp(\beta_0)/[1+\exp(\beta_0)]$.

The estimated value of β_0 is -0.37; this represents the log odds of receiving the full course of immunizations for children of the average mom (i.e. b_{ij}=0) within the average community (i.e. $b_i = 0$).

The estimate value of $\exp(\beta_0)/[1+\exp(\beta_0)]$ is approximately 41%; this represents the probability of receiving the full course of immunizations for children of the average mom within the average community.

b. Create an interval that contains the probability of a child receiving the full course of immunizations among children of 95% of the moms from the "average" community.

The "average" community is defined by setting $b_i = 0$.

Within any community *i*, the random intercept defined at the mom level allows for the log odds of receiving the full course of immunizations to vary from mom to mom.

Therefore, our calculation would be:

$\beta_0 +/- 1.96 \times \tau$ → -0.37 +/- 1.96 * sqrt(4.14) → -4.36 to 3.62

Converting this to an interval representing the probability of receiving the full course of immunizations for 95% of the moms from the "average" community yields 0.013 to 0.974.

c. Create an interval that contains the probability of a child receiving the full course of immunizations among children of 95% of the moms in Guatemala.

Here we are considering both sources of variation, across community and across moms within a community.

Therefore, our calculation would be

$\beta_0 +/- 1.96 \times \sqrt{\sigma^2+\tau^2}$ → -0.37 +/- 1.96 * sqrt(4.14+1.31) → -4.95 to 4.21

Converting this to an interval representing the probability of receiving the full course of immunizations for 95% of the moms in Guatemala yields 0.007 to 0.985.

d. Intra-class correlation coefficients:

Applying the latent variable formulation of the mixed logistic model below implies that $Var(Y_{ijk}) = \sigma^2 + \tau^2 + \frac{\pi^2}{3}$.

This is derived by the following:
$Y_{ijk} = 1 \rightarrow {y^*}_{ijk} = \beta_0 + b_i + b_{ij} + \varepsilon_{ijk} > 0,$
$b_i \sim N(0, \sigma^2), b_{ij} \sim N(0, \tau^2), \varepsilon_{ijk} \sim \textit{independent Logistic}, Var(\varepsilon_{ijk}) = \frac{\pi^2}{3},$
and lastly b_i, b_{ij} *and* ε_{ijk} independent.

We may want to compute various intraclass correlation coefficients within this random intercept only model.

- What is the correlation between the binary indicator of receiving the full course of immunizations for two children from the same mother, i.e. Corr(Y_{ijk}, Y_{ijm}) (i.e. the intraclass correlation for moms)

HINT: $\frac{Cov(Y_{ijk}, Y_{ijm})}{\sigma^2+\tau^2+\frac{\pi^2}{3}} = \frac{Cov(b_i+b_{ij}, b_i+b_{ij})}{\sigma^2+\tau^2+\frac{\pi^2}{3}} = \frac{\sigma^2+\tau^2}{\sigma^2+\tau^2+\frac{\pi^2}{3}} = \frac{4.14+1.33}{4.14+1.33+3.29} = 0.62$

- What is the correlation between the binary indicator of receiving the full course of immunizations for two children from different mothers within the same community, i.e. Corr(Y_{ijk}, Y_{imn}) (i.e. the intraclass correlation for community)?

$$\frac{Cov(Y_{ijk}, Y_{imn})}{\sigma^2+\tau^2+\frac{\pi^2}{3}} = \frac{Cov(b_i+b_{ij}, b_i+b_{im})}{\sigma^2+\tau^2+\frac{\pi^2}{3}} = \frac{\sigma^2}{\sigma^2+\tau^2+\frac{\pi^2}{3}} = \frac{1.33}{4.14+1.33+3.29} = 0.15$$

2. Three-level random intercept model with level-1 covariate

Now consider adding the primary covariate (*kid2p*) to the model. Recall, this is the indicator for whether the child was exposed to the immunization campaign or not.

$$log\left(\frac{Pr(y_{ijk}=1)}{1-Pr(y_{ijk}=1)}\right) = \beta_0 + b_i + b_{ij} + \beta_1 kid2p_{ijk},\ \ b_i \sim N(0, \sigma^2), b_{ij} \sim N(0, \tau^2), Cov(b_i, b_{ij}) = 0$$

In the model above, we allow each child to have his/her own log odds of receiving the full course of immunizations which depends on their mother and community membership; in addition, we assume that the effect of the immunization campaign is to change the log odds of receiving the full course of immunizations by the same factor (β_1) for each child.

The fit of this model is provided below:

```
. meqrlogit immun kid2p || cluster: || mom: , intp(12)


Mixed-effects logistic regression               Number of obs     =      2,158

-----------------------------------------------------------------------------
                |     No. of       Observations per Group       Integration
 Group Variable |     Groups    Minimum    Average    Maximum      Points
----------------+------------------------------------------------------------
        cluster |        161          1       13.4         55          12
```

```
           mom |      1,595          1        1.4          3          12
-----------------------------------------------------------------------------

                                                Wald chi2(1)      =      61.26
Log likelihood = -1353.5579                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       immun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   1.668822   .2132227     7.83   0.000     1.250913    2.086731
       _cons |  -1.722079   .2331868    -7.38   0.000    -2.179117   -1.265041
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cluster: Identity            |
                  var(_cons) |   1.592901   .4322593      .9358395     2.71129
-----------------------------+------------------------------------------------
mom: Identity                |
                  var(_cons) |   5.229659   1.183445      3.356224    8.148842
------------------------------------------------------------------------------
LR test vs. logistic model: chi2(2) = 189.90              Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.


. meqrlogit, or

Mixed-effects logistic regression               Number of obs     =      2,158

-----------------------------------------------------------------------------
                |     No. of       Observations per Group       Integration
 Group Variable |     Groups    Minimum    Average    Maximum      Points
----------------+------------------------------------------------------------
        cluster |        161          1       13.4         55          12
            mom |      1,595          1        1.4          3          12
-----------------------------------------------------------------------------

                                                Wald chi2(1)      =      61.26
Log likelihood = -1353.5579                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       immun | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   5.305913   1.131341     7.83   0.000     3.493531    8.058526
       _cons |   .1786943   .0416691    -7.38   0.000     .1131414    .2822276
------------------------------------------------------------------------------
Note: _cons estimates baseline odds (conditional on zero random effects).

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cluster: Identity            |
                  var(_cons) |   1.592901   .4322593      .9358395     2.71129
-----------------------------+------------------------------------------------
mom: Identity                |
                  var(_cons) |   5.229659   1.183445      3.356224    8.148842
------------------------------------------------------------------------------
LR test vs. logistic model: chi2(2) = 189.90              Prob > chi2 = 0.0000
```

a. Interpret the estimated value of $\exp(\beta_1)$.

For a given mom within a community in Guatemala, the odds of her completing the full course of immunizations for a child during the immunization campaign is 5.30 times the odds of completing the full course of immunizations for a child prior to the immunization campaign (95% CI: 3.49 to 8.06).

b. Describe an approach to estimate (ignore inference for now) the population level / marginal association between the odds of receiving the full course of immunizations and the immunization campaign.

We could fit a marginal model that would take the form:

$$log\left(\frac{\Pr(y_{ijk} = 1)}{1 - \Pr(y_{ijk} = 1)}\right) = \alpha_0 + \alpha_1 kid2p_{ijk},$$

$Corr(Y_{ijk}, Y_{ijm}) = \rho_2, Corr(Y_{ijk}, Y_{imn}) = \rho_1$, i.e. children within moms are exchangeable and moms within communities are exchangeable.

3. Three-level random intercepts plus random slope model

NOTE: No solution will be provided below as these questions or similar will be part of your final exam.

If you are a member of the health department in Guatemala, you may be interested in understanding if there was heterogeneity in the effect of the campaign across the communities.

Quantifying the heterogeneity may be of central interest as well as identifying communities where the campaign was more or less effective could lead to subsequent targeted changes in implementation of future campaigns.

Extend the model from part 2 to include a community level random slope for *kid2p.*

$$log\left(\frac{\Pr(y_{ijk}=1)}{1-\Pr(y_{ijk}=1)}\right) = \beta_0 + b_{0i} + b_{0ij} + (\beta_1 + b_{1i})kid2p_{ijk},$$

$$b_{0i} \sim N(0, \sigma_0{}^2), b_{1i} \sim N(0, \sigma_1{}^2), Cov(b_{0i}, b_{1i}) = \tau_{01},$$
$$b_{0ij} \sim N(0, \tau^2),$$
$$Cov(b_{0i}, b_{0ij}) = 0, Cov(b_{1i}, b_{0ij}) = 0$$

The fit of this model is presented below:

```
meqrlogit immun kid2p || cluster: kid2p, cov(uns) || mom: , intp(12)


Mixed-effects logistic regression               Number of obs     =      2,158

----------------------------------------------------------------------------
                |     No. of       Observations per Group       Integration
 Group Variable |     Groups    Minimum    Average    Maximum      Points
----------------+-----------------------------------------------------------
        cluster |        161          1       13.4         55          12
            mom |      1,595          1        1.4          3          12
----------------------------------------------------------------------------

                                                Wald chi2(1)      =      42.50
Log likelihood = -1348.8518                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       immun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   1.933637   .2966054     6.52   0.000     1.352302    2.514973
       _cons |  -1.986354   .3108885    -6.39   0.000    -2.595684   -1.377024
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cluster: Unstructured        |
                 var(kid2p) |   1.978795   1.064116      .6896979    5.677311
                 var(_cons) |   3.671978   1.461849      1.682786    8.012563
           cov(kid2p,_cons) |  -1.981638    1.14151     -4.218957    .2556801
-----------------------------+------------------------------------------------
mom: Identity                |
```

```
                  var(_cons) |   5.889154     1.3969      3.699551     9.37469
------------------------------------------------------------------------------
LR test vs. logistic model: chi2(4) = 199.32              Prob > chi2 = 0.0000


. meqrlogit, or

Mixed-effects logistic regression               Number of obs     =      2,158

----------------------------------------------------------------------------
                |     No. of       Observations per Group       Integration
 Group Variable |     Groups    Minimum    Average    Maximum      Points
----------------+-----------------------------------------------------------
        cluster |        161          1       13.4         55          12
            mom |      1,595          1        1.4          3          12
----------------------------------------------------------------------------

                                                Wald chi2(1)      =      42.50
Log likelihood = -1348.8518                     Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       immun | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   6.914616   2.050912     6.52   0.000     3.866314    12.36628
       _cons |   .1371947   .0426523    -6.39   0.000     .0745948    .2523284
------------------------------------------------------------------------------
Note: _cons estimates baseline odds (conditional on zero random effects).

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
cluster: Unstructured        |
                  var(kid2p) |   1.978795   1.064116      .6896979    5.677311
                  var(_cons) |   3.671978   1.461849      1.682786    8.012563
            cov(kid2p,_cons) |  -1.981638    1.14151     -4.218957    .2556801
-----------------------------+------------------------------------------------
mom: Identity                |
                  var(_cons) |   5.889154     1.3969      3.699551     9.37469
------------------------------------------------------------------------------
LR test vs. logistic model: chi2(4) = 199.32              Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.
```

a. Interpret the estimated value of $\exp(\beta_1)$.

b. Provide an interval that contains the effect of the campaign for 95% of the communities in Guatemala.

c. Estimate the $Corr(b_{0i}, b_{1i})$ and interpret this value within the context of the problem