Lab 10: Three-level logistic, Guatemala Data

Goal: learn how to implement three-level logistic models

The objective of this study is to identify important family- and community-level factors that affect whether Guatemalan children are immunized. A nationally representative sample of 5160 mothers, between 15 and 44 years old were interviewed.

Data: The data set used is called **guatemala.dta,** which can be downloaded directly from our website. The dataset comprises children *i* nested in mothers *j* nested in communities *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.

Level 2 (mothers)
-mom: identifier for mother
-Ethnicity (dummy variables with 'Latino' as reference category)
indNoSpa: mother is indigenous, not Spanish speaking
indSpa: mother is indigenous, Spanish speaking
-Mother's eduation (dummy variables with 'no education' as reference category)
monEdPri: mother has primary education
monEdSec: mother has secondary education
-Husband's education (dummy variables with 'no education' as reference category)
husEdPri: husband has primary education
husEdSec: husband has secondary education
husEdDK: husband's education is not known

Level 3 (communities)
-cluster: identifier for communities
-rural: dummy variable for community being rural
-pcInd81: percentage of population that was indigenous in 1981

Brief EDA:
How many communities are in the study and how many children per community?

```
. codebook cluster

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

                  type:  numeric (float)

                 range:  [1,240]                      units:  1
          unique values:  161                     missing .:  0/2159

                  mean:   145.814
```

```
              std. dev:   59.3619

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

We have 161 communities.

It appears that the minimum number of children per community is 1 and the maximum is 55.

What is the overall proportion of children in the study who have been immunized?

```
. summ immun

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       immun |      2159    .446503    .497245          0          1
```

The first model: three-level random intercept model

We use indices *i*, *j*, *k* for children, mothers and communities, respectively. The binary response Y_{ijk} may be modeled by a generalized linear mixed model with linear predictor.

$$\log\left(\frac{p(y_{ijk}=1)}{1-p(y_{ijk}=1)}\right)=\eta_{ijk}$$

$$\eta_{ijk}=\beta_0+\beta_1 kid2p_{ijk}+\beta_2 indNoSpa_{jk}+\ldots+\beta_{10}pcInd81_k+U_{jk}+U_k$$

Here U_{jk} is the random intercept for mom *j* in cluster *k*. U_k is the random intercept for cluster *k*. The random intercepts are assumed to be independently normally distributed.

The Stata command is:

```
gllamm immun kid2p indNoSpa indSpa  momEdPri momEdSec husEdPri husEdSec
husEdDK rural pcInd81, family(binomial) link(logit) i(mom cluster)
nip(5)
```

The `i(mom cluster)` part of the gllamm command specifies the hierarchical structure of the data with the lowest levels (finest clusters) specified first and the higher levels specified next. We used only 5 quadrature points because estimation would otherwise be quite slow for this sample (as is, it takes less than 5 minutes on my computer). With as few as 5 points, adaptive quadrature is sometimes unstable, so we have used ordinary quadrature by omitting the adapt option.

The results are

```
number of level 1 units = 2159
number of level 2 units = 1595
number of level 3 units = 161

Condition Number = 10.125573

gllamm model

log likelihood = -1328.0727
```

```
------------------------------------------------------------------------------
       immun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |    1.712282   .2139083     8.00   0.000     1.293029    2.131535
    indNoSpa |   -.2992919   .4837166    -0.62   0.536    -1.247359    .6487752
      indSpa |   -.2178983    .361165    -0.60   0.546    -.9257687    .4899721
    momEdPri |    .3789442   .2154968     1.76   0.079    -.0434219    .8013102
    momEdSec |    .3836724   .4605474     0.83   0.405    -.5189838    1.286329
    husEdPri |    .4934885   .2244022     2.20   0.028     .0536682    .9333087
    husEdSec |    .4466857   .4008267     1.11   0.265    -.3389202    1.232291
     husEdDK |   -.0079424   .3485074    -0.02   0.982    -.6910043    .6751195
       rural |   -.8642705    .300585    -2.88   0.004    -1.453406   -.2751347
     pcInd81 |    -1.17417   .4953426    -2.37   0.018    -2.145023   -.2033158
       _cons |   -1.054729   .4085557    -2.58   0.010    -1.855484   -.2539746
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (mom)

    var(1): 5.427267 (1.318504)

***level 3 (cluster)

    var(1): 1.1338842 (.37262627)
```

We now increase the number of quadrature points to the default of 8 per dimension and use adaptive quadrature to obtain more accurate results. We use the previous estimates as stating values (took about 10 minutes):

```
matrix a=e(b)

gllamm immun kid2p indNoSpa indSpa  momEdPri momEdSec husEdPri husEdSec
husEdDK rural pcInd81, family(binomial) link(logit) i(mom cluster)
from(a) adapt

gllamm, eform

estimates store model1
```

We get

```
number of level 1 units = 2159
number of level 2 units = 1595
number of level 3 units = 161

Condition Number = 9.6662017

gllamm model

log likelihood = -1328.4911

------------------------------------------------------------------------------
       immun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |    1.711931   .2148514     7.97   0.000      1.29083    2.133032
    indNoSpa |    -.300227   .4770976    -0.63   0.529    -1.235321    .6348671
      indSpa |   -.1580678   .3565839    -0.44   0.658    -.8569595    .5408238
    momEdPri |    .3840292   .2167929     1.77   0.076    -.0408771    .8089355
    momEdSec |    .3615277   .4732679     0.76   0.445    -.5660604    1.289116
    husEdPri |    .4988082   .2271986     2.20   0.028     .0535071    .9441094
```

```
    husEdSec |    .438249   .4039136     1.09   0.278    -.3534071    1.229905
     husEdDK |  -.0091359   .3514171    -0.03   0.979    -.6979009     .679629
       rural |  -.8941843   .2994106    -2.99   0.003    -1.481018   -.3073503
     pcInd81 |  -1.155453   .4936293    -2.34   0.019    -2.122949   -.1879575
       _cons |  -1.025186   .4056784    -2.53   0.012    -1.820301   -.2300704
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (mom)

    var(1): 5.1730109 (1.176587)

***level 3 (cluster)

    var(1): 1.028134 (.31704835)
------------------------------------------------------------------------------


. gllamm, eform

number of level 1 units = 2159
number of level 2 units = 1595
number of level 3 units = 161

Condition Number = 9.6662017

gllamm model

log likelihood = -1328.4911

------------------------------------------------------------------------------
       immun |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |    5.53965   1.190201     7.97   0.000     3.635804    8.440421
    indNoSpa |   .7406501   .3533624    -0.63   0.529     .2907414    1.886771
      indSpa |   .8537919   .3044485    -0.44   0.658     .4244507    1.717421
    momEdPri |   1.468188   .3182928     1.77   0.076     .9599471    2.245516
    momEdSec |   1.435521   .6793859     0.76   0.445     .5677577    3.629576
    husEdPri |   1.646758   .3741411     2.20   0.028     1.054964    2.570523
    husEdSec |   1.549991   .6260624     1.09   0.278     .7022912    3.420905
     husEdDK |   .9909057   .3482212    -0.03   0.979     .4976288    1.973146
       rural |    .408941   .1224413    -2.99   0.003      .227406    .7353929
     pcInd81 |   .3149148   .1554512    -2.34   0.019     .1196782    .8286499
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (mom)

    var(1): 5.1730109 (1.176587)

***level 3 (cluster)

    var(1): 1.028134 (.31704835)
------------------------------------------------------------------------------
```

The second model: three-level random intercept model with a subset of the covariates

This model only has kid2p and the community level variables in the fixed part. Then, the binary response Y_{ijk} can again be modeled by a generalized linear mixed model with linear predictor.

$$\log\left(\frac{p(y_{ijk}=1)}{1-p(y_{ijk}=1)}\right)=\eta_{ijk}$$

$$\eta_{ijk}=\beta_0+\beta_1 kid2p_{ijk}+\beta_2 rural_k+\beta_3 pcInd81_k+U_{jk}+U_k$$

Here U_{jk} is the random intercept for mom *j* in community *k*. U_k is the random intercept for community *k*. The random intercepts are assumed to be independently normally distributed. (analogous to the three-level random intercept model for continuous outcomes)

The Stata commands are:

```
matrix a=e(b)
gllamm immun kid2p rural pcInd81, family(binomial) link(logit) i(mom
cluster) from(a) skip adapt
gllamm, eform

estimates store model2

number of level 1 units = 2159
number of level 2 units = 1595
number of level 3 units = 161

Condition Number = 5.4765463

gllamm model

log likelihood = -1335.0434

------------------------------------------------------------------------------
       immun |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   1.684492   .2143569     7.86   0.000      1.26436    2.104624
       rural |  -1.069097   .2852915    -3.75   0.000    -1.628258   -.5099358
     pcInd81 |  -1.665784   .3583539    -4.65   0.000    -2.368145    -.963423
       _cons |  -.2142503   .3072693    -0.70   0.486    -.8164871    .3879864
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (mom)

    var(1): 5.2514807 (1.2012076)

***level 3 (cluster)

    var(1): 1.0428961 (.31659208)
------------------------------------------------------------------------------


. gllamm, eform

number of level 1 units = 2159
number of level 2 units = 1595
number of level 3 units = 161

Condition Number = 5.4765463

gllamm model
```

```
log likelihood = -1335.0434

------------------------------------------------------------------------------
       immun |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   5.389713   1.155322     7.86   0.000     3.540827    8.204017
       rural |   .3433184   .0979458    -3.75   0.000     .1962712    .6005342
     pcInd81 |   .1890424   .0677441    -4.65   0.000     .0936543    .3815845
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (mom)

    var(1): 5.2514807 (1.2012076)

***level 3 (cluster)

    var(1): 1.0428961 (.31659208)
------------------------------------------------------------------------------
```

The estimate of the odds ratio for kid2p has not changed considerable compared with the estimate for model 1, suggesting that discarding the level 2 (mother-level) covariates does not dramatically effect the estimate.

The third model: random coefficients

The binary response Y_{ijk} may be modeled by a generalized linear mixed model with linear predictor.

$$\log\left(\frac{p(y_{ijk}=1)}{1-p(y_{ijk}=1)}\right)=\eta_{ijk}$$

$$\eta_{ijk}=\beta_0+(\beta_1+U_{k1})kid2p_{ijk}+\beta_2 rural_k+\beta_3 pcInd81_k+U_{jk}+U_{k0}$$

Here U_{jk} is the random intercept for mom *j* in cluster *k*. U_{k0} is the random intercept for cluster *k*, U_{k1} is the random slope for cluster *k* on kid2p. The random intercept U_{jk} is assumed to be independently normally distributed. The random intercept U_{k0} and the random slope U_{k1} are multivariate normally distributed.

The Stata commands are:

```
gen cons=1
eq inter: cons
eq slope: kid2p
matrix a = e(b)
matrix a = (a,.2,0)
gllamm immun kid2p rural pcInd81, family(binomial) link(logit) i(mom
cluster) nrf(1 2) eqs(inter inter slope) nip(8 4 4) from(a) copy adapt
eform

estimates store model3
```

Results:

```
number of level 1 units = 2159
number of level 2 units = 1595
number of level 3 units = 161

Condition Number = 7.1034028

gllamm model

log likelihood = -1330.8167

------------------------------------------------------------------------------
       immun |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   6.714491   1.961258     6.52   0.000     3.787763    11.90264
       rural |   .3289608   .0988563    -3.70   0.000     .1825361    .5928426
     pcInd81 |    .176146   .0667231    -4.58   0.000     .0838383    .3700865
------------------------------------------------------------------------------


Variances and covariances of random effects
------------------------------------------------------------------------------

***level 2 (mom)

    var(1): 5.8122902 (1.3924712)

***level 3 (cluster)

    var(1): 2.4200713 (1.0954187)
    cov(1,2): -1.5220874 (.94879562) cor(1,2): -.72917723

    var(2): 1.8004656 (.98859075)
------------------------------------------------------------------------------


. estimates store model3

.
. lrtest model2 model3
(log-likelihoods of null models cannot be compared)

likelihood-ratio test                                 LR chi2(2)  =      8.45
(Assumption: model2 nested in model3)                 Prob > chi2 =    0.0146
```

The small p-value here suggests that we should include the random slope on kid2p.

Coefficient Interpretations

β_0: The log odds of immunization for a child who is less than 2 years old of a *typical mother* in a *typical community* at baseline (non-rural community with zero% indigenous population in 1981).

β_1: The log odds ratio for immunization comparing a child being at least 2 years old to a child less than 2 years old of a *typical mother* in a *specific community*, controlling for the community being rural and the % of indigenous population in the community in 1981.

β_2: The log odds ratio for immunization comparing a child from a rural community versus a child from a non-rural community of a *specific mother* from a *specific*

community, controlling for the % of indigenous population in the community in 1981.

β_3: The log odds ratio for immunization of a child associated with a 1% increase in the % of indigenous population in the community in 1981 of a *specific mother* from a *specific community*, controlling for the community being rural or not.

U_{k0}: The difference in the log odds of immunization for a child who is less than 2 years old of a *typical mother* at baseline (non-rural community with zero % indigenous population in 1981) comparing a *specific community* to a *typical community.*

U_{jk}: The *mother-specific* random deviation of log odds of immunization for a child who is less than 2 years old of a *typical community* at baseline (non-rural community with zero % indigenous population in 1981).

U_{k1}: The *mother-specific* random deviation of log odds ratio of immunization comparing a child who is less than 2 years old to a child who is greater than 2 years old of a *specific community* controlling for rural community and the % of indigenous population in 1981).

Cross-level interaction.

To reduce analysis complexity, we'll focus on the 2-stage multi-level model first.
Level 1: children (denoted by *i*)
Level 2: community (denoted by *k*).

Model 1: What is the effect of $kid2p_{ik}$ accounting for the between-community heterogeneity?

$$\log\left(\frac{p(y_{ik}=1)}{1-p(y_{ik}=1)}\right)=\eta_{ik}$$

$$\eta_{ik}=\beta_{0k}+\beta_{1k}kid2p_{ik}$$

$$\beta_{0k}=\beta_0+U_{k0}$$

$$\beta_{1k}=\beta_1+U_{k1}$$

β_{0k}: community-specific intercept, i.e., baseline log odds of being immunized (<2y)

β_{1k}: community-specific slope of $kid2p_{ik}$, i.e., log OR being immunized comparing >=2y versus <2y.

The equivalent 1-line writing of η_{ijk} is:

$$\eta_{ik}=\beta_0+\beta_1 kid2p_{ik}+U_{k0}+U_{k1}kid2p_{ik}$$

β_0: overall intercept (fixed effects)
β_1: main effect of $kid2p_{ik}$ (fixed effects)

```
. eq inter: cons

. eq slope: kid2p

. gllamm immun kid2p, family(binomial) link(logit) i(cluster) nrf(2) eqs(inter
slope) nip(4 4) adapt eform

number of level 1 units = 2159
number of level 2 units = 161

gllamm model

------------------------------------------------------------------------------
       immun |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   3.073634   .4969815     6.94   0.000     2.238823    4.219728
------------------------------------------------------------------------------

Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (cluster)

    var(1): 1.2882633 (.47966448)
    cov(2,1): -.65561142 (.39690843) cor(2,1): -.71194885

    var(2): .65824989 (.36732232)
------------------------------------------------------------------------------
```

Model 2: Does community-level covariates explain the between-community heterogeneity in the baseline log odds of being immunized?

$$\log\left(\frac{p(y_{ik}=1)}{1-p(y_{ik}=1)}\right)=\eta_{ik}$$
$$\eta_{ik}=\beta_{0k}+\beta_{1k}kid2p_{ik}$$
$$\beta_{0k}=\beta_0+\beta_2 rural_k+\beta_3 pcInd81_k+U_{k0}$$
$$\beta_{1k}=\beta_1+U_{k1}$$

The equivalent 2-stage writing of η_{ijk} is:

$$\eta_{ik}=\beta_0+(\beta_1+U_{k1})kid2p_{ik}+\beta_2 rural_k+\beta_3 pcInd81_k+U_{k0}$$

β_{0k}, β_{1k}, β_0, β_1: Same as above.
β_2: main effect of $rural_k$ (fixed effects)
β_3: main effect of $pcInd81_k$ (fixed effects)

```
. gen cons=1
. eq inter: cons
. eq slope: kid2p

. gllamm immun kid2p rural pcInd81, family(binomial) link(logit) i(cluster)
nrf(2) eqs(inter slope) nip(4 4) adapt eform
```

```
number of level 1 units = 2159
number of level 2 units = 161

Condition Number = 7.1205404

gllamm model

------------------------------------------------------------------------------
       immun |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   2.984958   .4724544     6.91   0.000     2.188826    4.070662
       rural |   .5294077   .0867878    -3.88   0.000     .3839278    .7300136
     pcInd81 |   .3842638   .0782185    -4.70   0.000      .257848    .5726576
------------------------------------------------------------------------------
Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (cluster)

    var(1): .85945899 (.36518027)
    cov(2,1): -.4942948 (.33061796) cor(2,1): -.68798101

    var(2): .60061203 (.34310316)
------------------------------------------------------------------------------
```

The variance of the random intercept decrease, indicating that the community-level covariates $rural_k$ and $pcInd81_k$ explain the between-community variability in baseline log odd of being immunized. The statistical significance of the main effects of $rural_k$ and $pcInd81_k$ also suggests this conclusion.

Model 3:Does community-level covariates explain the between-community heterogeneity in both the baseline log odds of being immunized and the log OR being immunized comparing >=2y versus <2y?

$$\log\left(\frac{p(y_{ik}=1)}{1-p(y_{ik}=1)}\right)=\eta_{ik}$$

$$\eta_{ik}=\beta_{0k}+\beta_{1k}kid2p_{ik}$$

$$\beta_{0k}=\beta_0+\beta_2 rural_k+\beta_3 pcInd81_k+U_{k0}$$

$$\beta_{1k}=\beta_1+\beta_4 rural_k+\beta_5 pcInd81_k+U_{k1}$$

The equivalent 2-stage writing of η_{ijk} is:

$$\eta_{ik}=\beta_0+\beta_2 rural_k+\beta_3 pcInd81_k+U_{k0}+(\beta_1+\beta_4 rural_k+\beta_5 pcInd81_k+U_{k1})kid2p_{ik}$$

$$\eta_{ik}=\beta_0+\beta_1 kid2p_{ik}+\beta_2 rural_k+\beta_3 pcInd81_k+\beta_4 rural_k*kid2p_{ik}+\beta_5 pcInd81_k*kid2p_{ik}+U_{k0}+U_{k1}*kid2p_{ik}$$

β_{0k}, β_{1k}, β_0, β_1, β_2, β_3: Same as above.

β_4 :cross-level interaction between $rural_k$ and $kid2p_{ik}$ (fixed effects)

β_5 :cross-level interaction between $pcInd81_k$ and $kid2p_{ik}$ (fixed effects)

```
. gen int_2p_ru = kid2p * rural
. gen int_2p_pc = kid2p * pcInd81
. eq inter: cons
```

```
. eq slope: kid2p

. gllamm immun kid2p rural pcInd81 int_2p_ru int_2p_pc, family(binomial)
link(logit) i(cluster) nrf(2) eqs(inter slope) nip(4 4) adapt eform
gllamm model

------------------------------------------------------------------------------
       immun |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       kid2p |   2.311586   .7539445     2.57   0.010     1.219784    4.380635
       rural |   .5115291   .1639525    -2.09   0.036     .2729278    .9587223
     pcInd81 |   .2402431   .0980197    -3.50   0.000     .1079839     .534494
   int_2p_ru |   1.045638   .3464314     0.13   0.893     .5462218    2.001676
   int_2p_pc |   1.755981    .727406     1.36   0.174     .7796755     3.95481
------------------------------------------------------------------------------
Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (cluster)

    var(1): .95682725 (.39271689)
    cov(2,1): -.56582712 (.34798787) cor(2,1): -.72621719

    var(2): .63445517 (.34985769)
------------------------------------------------------------------------------
```

The variance of the random slope remains approximately the same, indicating that the community-level covariates $rural_k$ and $pcInd81_k$ do not explain the between-community variability in the log OR being immunized comparing >=2y versus <2y. This can be also inferred from the non-statistically significant (cross-level) interaction between $kid2p_{ik}$ and the community-level variables $rural_k$ and $pcInd81_k$.

Coefficient Interpretations

β_0: The log odds of immunization for a child who is less than 2 years old of a *typical community* at baseline (non-rural community with zero % indigenous population in 1981).

β_1: The log odds ratio of immunization comparing a child being at least 2 years old to a child less than 2 years old in a *typical* non-rural community with zero % of indigenous population in 1981.

β_2: For a *specific* community, the log odds ratio of immunization of a child associated with the community being rural or not, controlling for the % of indigenous population in 1981 and whether the child is at least 2 years old or not.

β_3: For a *specific* community, the log odds ratio of immunization of a child associated with a 1% increase in the % of indigenous population in 1981, controlling for the community being rural or not and whether the child is at least 2 years old or not.

β_4: For a *specific* community, the change in log odds ratio of immunization comparing a child being at least 2 years old to a child less than 2 years old

associated with the community being rural or not, controlling for the % of indigenous population in 1981.

β_5: For a *specific* community, the change in log odds ratio of immunization comparing a child being at least 2 years old to a child less than 2 years old associated with a 1% increase in the % of indigenous population in 1981, controlling for the community being rural or not.