Lab: Two-stage logistic -- Race, age, gender, and mortality in 7 US zipcodes

Data: Race, age, gender and mortality data for 5800 individuals living in seven zip codes in the U.S. (zipcode_mortality.dta)

Variables:

- **race**: 0--White; 1--Black.
- **sex**: 0--Male; 1--Female.
- **agecat**: 1--[65,70); 2--[70,75); 3--[75,80); 4--[80,85); 5--[85,+)
- **zip**: 1-7--Indicators of 7 distinct zip codes
- **y**: 0--Alive; 1--Dead.

Goal: Assess the association between mortality and demographic characteristics through a two-level logistic regression model, accounting for clustering within zipcodes.

Exploratory Data Analysis:
Quickly examine the categorical variable summaries.

```
. tab agecat
     agecat |      Freq.     Percent        Cum.
------------+-----------------------------------
    [65,70) |      1,388       23.93       23.93
    [70,75) |      1,416       24.41       48.34
    [75,80) |      1,270       21.90       70.24
    [80,85) |        928       16.00       86.24
     [85,+) |        798       13.76      100.00
------------+-----------------------------------
      Total |      5,800      100.00

. tab sex
        sex |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |      2,270       39.14       39.14
     female |      3,530       60.86      100.00
------------+-----------------------------------
      Total |      5,800      100.00
```

```
. tab y race, column
           |         race
         y |     white      black |     Total
-----------+----------------------+----------
     alive |     5,416         82 |     5,498
           |     94.83      92.13 |     94.79
-----------+----------------------+----------
      dead |       295          7 |       302
           |      5.17       7.87 |      5.21
-----------+----------------------+----------
     Total |     5,711         89 |     5,800
           |    100.00     100.00 |    100.00
```

We observe a higher percentage of death for the black population compared with the white population.

```
. tab zip race

           |         race
       zip |     white      black |     Total
-----------+----------------------+----------
         1 |     2,212          5 |     2,217
         2 |     1,444         72 |     1,516
         3 |       251          9 |       260
         4 |        68          0 |        68
         5 |       860          1 |       861
         6 |       699          2 |       701
         7 |       177          0 |       177
-----------+----------------------+----------
     Total |     5,711         89 |     5,800
```

Zip codes 4 and 7 contain no black individuals.

Analysis assuming independence between individuals:

```
. xi:logit y race i.agecat sex
i.agecat        _Iagecat_1-5          (naturally coded; _Iagecat_1 omitted)

Logistic regression                               Number of obs   =       5800
                                                  LR chi2(6)      =     154.38
                                                  Prob > chi2     =     0.0000
Log likelihood = -1109.2719                       Pseudo R2       =     0.0651
```

```
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   .7277619   .4087863     1.78   0.075    -.0734445    1.528968
  _Iagecat_2 |   .3289547    .260447     1.26   0.207     -.181512    .8394215
  _Iagecat_3 |   1.181652   .2326983     5.08   0.000     .7255715    1.637732
  _Iagecat_4 |   1.451145   .2365551     6.13   0.000     .9875056    1.914785
  _Iagecat_5 |   2.142882    .226895     9.44   0.000     1.698176    2.587588
         sex |  -.4305185   .1228225    -3.51   0.000    -.6712462   -.1897909
       _cons |  -3.767649   .2063255   -18.26   0.000     -4.17204   -3.363259
------------------------------------------------------------------------------
```

We have found:

(1) higher (not statistically significant) risk of death for the black population;

(2) higher risk of death as age increases;

(3) lower risk of death for female compared with male.

Additionally adjusting for the interaction between age and gender:

```
. gen intagesex = agecat * sex
. xi:logit y race i.agecat sex i.intagesex

Logistic regression                               Number of obs   =       5800
                                                  LR chi2(10)     =     158.74
                                                  Prob > chi2     =     0.0000
Log likelihood = -1107.0936                       Pseudo R2       =     0.0669
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   .7152554   .4091241     1.75   0.080    -.0866132    1.517124
  _Iagecat_2 |   .3271268   .3863061     0.85   0.397    -.4300192    1.084273
  _Iagecat_3 |   1.507252   .3340617     4.51   0.000     .8525028       2.162
  _Iagecat_4 |   1.738405   .3461203     5.02   0.000     1.060022    2.416789
  _Iagecat_5 |   2.242273   .3494232     6.42   0.000     1.557416    2.927129
         sex |  -.6195738   .2567102    -2.41   0.016    -1.122717   -.1164311
_Iintagese~1 |    .562087   .4729937     1.19   0.235    -.3649635    1.489138
_Iintagese~2 |   .5512787   .4259249     1.29   0.196    -.2835187    1.386076
_Iintagese~3 |  -.1098395   .3550985    -0.31   0.757    -.8058197    .5861407
_Iintagese~5 |   .3123178   .3450935     0.91   0.365     -.364053    .9886887
       _cons |  -3.951343   .2919392   -13.53   0.000    -4.523534   -3.379153
------------------------------------------------------------------------------
```

Even after adjusting for the interaction between age and gender we find a similar association between race and mortality risks.

In this preliminary model, we find that black individuals have a log-odds of mortality that is 0.72 (-0.09, 1.52) that of white individuals when adjusting for age, sex, and the interaction between age and sex. The log-odds scale is not very interpretable, so we transform our results to odds ratios. Adjusting for age, sex, and their interaction, blacks have an odds of mortality that is exp(0.72) = 2.05 times that of whites with a 95% confidence interval for the odds ratio from exp(-0.09) = 0.91 to exp(1.52) = 4.57.

Stata will present the output on the odds scale when you use the command `logistic`.

```
. xi:logistic y race i.agecat sex i.intagesex i.agecat
Logistic regression                               Number of obs   =       5800
LR chi2(10)     =     158.74                      Prob > chi2     =     0.0000
Log likelihood = -1107.0936                       Pseudo R2       =     0.0669
------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   2.044709   .8365397     1.75   0.080     .9170317    4.559094
_Iagecat_2   |   1.386977   .5357977     0.85   0.397     .6504966    2.957288
_Iagecat_3   |   4.514307   1.508057     4.51   0.000      2.34551    8.688501
_Iagecat_4   |   5.688264   1.968824     5.02   0.000     2.886434     11.2098
_Iagecat_5   |   9.414702   3.289715     6.42   0.000     4.746539    18.67395
         sex |   .5381737   .1381547    -2.41   0.016     .3253946    .8900914
_Iintagese~1 |    1.75433    .829787     1.19   0.235      .694222     4.43327
_Iintagese~2 |   1.735471   .7391802     1.29   0.196      .753129    3.999128
_Iintagese~3 |    .895978   .3181604    -0.31   0.757     .4467216     1.79704
_Iintagese~5 |   1.366589    .471601     0.91   0.365     .6948544    2.687708
------------------------------------------------------------------------------
```

Random intercept model to account for clustering of individuals within zip code:

The model we will fit to examine the association between demographics and mortality, accounting for the correlation among measurements on individuals from the same zip code, may be written as:

$$\text{logit}\{\text{P}(y_{ij}=1)\} = \beta_0 + \beta_1 race_{ij} + \beta_2 agecat_{2ij} + \beta_3 agecat_{3ij} + \beta_4 agecat_{4ij} + \beta_5 agecat_{5ij} + \beta_6 sex_{ij} + U_i$$

where *i* indexes zipcode and *j* indexes individual. $U_i \sim N(0, \sigma_u^2)$ is a random intercept for zipcode.

We cannot use "xi:" with gllamm, so we need to type all the indicator variables for age

category in the glamm command. We will keep the same age and interaction indicator variables for consistent comparison with independence model results.

```
. gllamm y race _Iagecat_2 _Iagecat_3 _Iagecat_4 _Iagecat_5 sex, i(zip) l(logit)
f(binom) adapt

gllamm model
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   .7118934   .4721737     1.51   0.132      -.21355    1.637337
  _Iagecat_2 |   .3282704   .2606332     1.26   0.208    -.1825612     .839102
  _Iagecat_3 |   1.180596   .2332157     5.06   0.000     .7235016    1.637691
  _Iagecat_4 |   1.450593    .236706     6.13   0.000     .9866581    1.914529
  _Iagecat_5 |   2.142917   .2269695     9.44   0.000     1.698065    2.587768
         sex |  -.4316824   .1240596    -3.48   0.001    -.6748346   -.1885301
       _cons |  -3.767305   .2079067   -18.12   0.000    -4.174794   -3.359815
------------------------------------------------------------------------------
Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (zip)
    var(1): .00273478 (.04761232)
------------------------------------------------------------------------------
```

The variance among zipcodes in the average log-odds of death is estimated to be 0.003. There appears to be little or no variability between zipcodes, indicating that the random intercept may be unnecessary.

Prediction

Generate average predicted probability of death for black and white populations:

(1) Predict fixed effects

```
. predict fit1,xb
(xb will be stored in fit1)
```

(2) Calculate average probability from fixed effects (i.e., when zip code random effect is zero)

```
. gen phat1 = exp(fit1)/( 1 + exp(fit1) )
. table race, c(mean phat1)
-------------------------
     race | mean(phat1)
----------+--------------
    white |      .0516184
    black |      .0775048
-------------------------
```

Generate predicted probability of death for black and white populations for each zip code:
(1) Empirical Bayes predictions for the random intercept using `gllapred`:

```
. gllapred ebfit1, linpred
```

(2) Calculate zip code-specific probabilities by adding random effects to average probabilities:

```
. gen ebphat1 = exp(ebfit1)/( 1 + exp(ebfit1) )

. table zip race, c(mean ebphat1)
------------------------------
          |        race
      zip |    white     black
----------+-------------------
        1 | .0520663  .0379699
        2 | .0548089  .0818265
        3 | .0480184  .0746415
        4 |  .043442
        5 | .0525316  .0437937
        6 | .0466228  .0980782
        7 | .0431963
------------------------------
```

Fit random intercept model with interaction between age and gender:

```
.  gllamm  y  race  _Iagecat_2  _Iagecat_3  _Iagecat_4  _Iagecat_5  sex  _Iintagesex_1
_Iintagesex_2 _Iintagesex_3 _Iintagesex_5, i(zip) l(logit) f(binom) adapt

gllamm model
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   .6984531   .4660362     1.50   0.134     -.214961    1.611867
  _Iagecat_2 |   .3265898   .3863807     0.85   0.398    -.4307024    1.083882
  _Iagecat_3 |    1.50638   .3342846     4.51   0.000     .8511939    2.161566
  _Iagecat_4 |   1.738268   .3461883     5.02   0.000     1.059752    2.416785
  _Iagecat_5 |   2.243047   .3497381     6.41   0.000     1.557573    2.928521
         sex |  -.6209422   .2573449    -2.41   0.016    -1.125329   -.1165554
_Iintagese~1 |   .5629581   .4731686     1.19   0.234    -.3644353    1.490351
_Iintagese~2 |   .5518311   .4260114     1.30   0.195    -.2831359    1.386798
_Iintagese~3 |  -.1094562   .3551666    -0.31   0.758    -.8055699    .5866575
_Iintagese~5 |   .3118824   .3452144     0.90   0.366    -.3647254    .9884902
       _cons |  -3.951249   .2930427   -13.48   0.000    -4.525602   -3.376896
```

```
------------------------------------------------------------------------------
Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (zip)
    var(1): .00290965 (.04488606)
------------------------------------------------------------------------------
```

Except for the slightly inflated s.e. of the race coefficient, there is not much difference in the random effect models results compared with ordinary logistic regression results. This is not unexpected because the variance of the zip code random intercept is very small (.00290386), indicating ignorable correlation between individuals within the same zip code.

Prediction

Generate average predicted probability of death for black and white populations:

```
. predict fit2,xb
. gen phat2 = exp(fit2)/( 1 + exp(fit2) )
. sort race
. table race, c(mean phat2)
------------------------
     race | mean(phat2)
----------+-------------
    white |     .0516209
    black |     .0774457
------------------------
```

Generate predicted probability of death for black and white populations for each zip code:

```
. gllapred ebfit2, linpred
. gen ebphat2 = exp(ebfit2)/( 1 + exp(ebfit2) )
. table zip race, c(mean ebphat2)
------------------------------
          |        race
      zip |    white     black
----------+-------------------
        1 | .0521326   .038434
        2 | .0547642   .0818445
        3 | .0482556   .0731838
        4 | .0433313
        5 | .052397    .0361831
        6 | .0466139   .1072525
        7 | .0430694
------------------------------
```

We can see that although on average the black population has higher mortality risk than the white population, however, for each single zip code, it is not necessarily true.

Random Intercept and Slope Model

(Before using "gllamm", define equations for variables that needs random effects)

To save computing time, use the model without interaction between age and gender. **(The following model takes 2 hours running time!!)**

Add random slope to race coefficient:

```
. gen byte one = 1
. eq cons : one
. eq race : race
```

(nrf specifies the number of variables that need random effects, eqs give the corresponding variables.)

```
. gllamm y race _Iagecat_2 _Iagecat_3 _Iagecat_4 _Iagecat_5 sex, i(zip) l(logit)
f(binom) adapt nrf(2) eqs(race cons)
gllamm model
------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        race |   .7146109   .4871685     1.47   0.142    -.2402218    1.669444
  _Iagecat_2 |   .3283549   .2606398     1.26   0.208    -.1824897    .8391995
  _Iagecat_3 |   1.180679   .2332115     5.06   0.000     .7235933    1.637765
  _Iagecat_4 |   1.450654   .2367171     6.13   0.000     .9866972    1.914611
  _Iagecat_5 |   2.142967   .2269858     9.44   0.000     1.698083    2.587851
         sex |  -.4316525   .1240186    -3.48   0.001    -.6747244   -.1885805
       _cons |   -3.76738   .2079092   -18.12   0.000    -4.174874   -3.359885
------------------------------------------------------------------------------
Variances and covariances of random effects
------------------------------------------------------------------------------
***level 2 (zip)
    var(1): .00004335 (.00587423)
    cov(2,1): -.00034093 (.02326116) cor(2,1): -.99985372
    var(2): .0026823 (.04663141)
```

Prediction:

```
. predict fit3,xb
(xb will be stored in fit3)
. gen phat3 = exp(fit3)/( 1 + exp(fit3) )
. table race, c(mean phat3)
--------------------------
     race |    mean(phat3)
----------+---------------
    white |       .0516185
    black |       .0776903
--------------------------

. gllapred ebfit3, linpred
. gen ebphat3 = exp(ebfit3)/( 1 + exp(ebfit3) )

. table zip race, c(mean ebphat3)
------------------------------
          |        race
      zip |    white     black
----------+-------------------
        1 | .0521413  .0380786
        2 | .0538195  .0816576
        3 | .0466652  .0743084
        4 | .0435704
        5 | .0542231  .0441361
        6 | .0468852   .098454
        7 | .0434171
------------------------------
```

There is not much difference in the results when random slope of race is further included into the model.

Comparison of zip code-specific predicted probabilities from the above three random effects models:

```
. table zip race, c(mean ebphat1 mean ebphat2 mean ebphat3)
------------------------------
          |        race
      zip |    white     black
----------+-------------------
        1 | .0520663  .0379699
          | .0521326   .038434
          | .0521413  .0380786
          |
        2 | .0548089  .0818265
          | .0547642  .0818445
          | .0538195  .0816576
          |
        3 | .0480184  .0746415
          | .0482556  .0731838
          | .0466652  .0743084
          |
        4 |  .043442
          | .0433313
          | .0435704
          |
        5 | .0525316  .0437937
          |  .052397  .0361831
          | .0542231  .0441361
          |
        6 | .0466228  .0980782
          | .0466139  .1072525
          | .0468852   .098454
          |
        7 | .0431963
          | .0430694
          | .0434171
------------------------------
```