Lab 3: Two levels Poisson models
(taken from *Multilevel and Longitudinal Modeling Using Stata*, p. 376-390)

Goal: To see if a major health-care reform which took place in 1997 in Germany was a success in decreasing the number of doctor visits.

Data: A subset of the German Socio-Economic Panel data comprised of women working full time in the 1996 panel wave preceding the reform and the 1998 panel wave following the reform will be considered. The dataset is called drvisits.dta, in which there are the following variables:

- **Id**: person identifier (i)
- **Numvisits**: self-reported number of visits to a doctor during the three months prior to the interview (y_{ij})
- **Reform**: dummy variable for interview being during the year after the reform versus the year before the reform (x_{2ij})
- **Age**: age in years (x_{3ij})
- **Educ**: education in years (x_{4ij})
- **Married**: dummy variable for being married (x_{5ij})
- **Badh**: dummy variable for self-reported current health being classified as 'very poor' or 'poor' (versus 'very good', 'good' or 'fair') (x_{6ij})
- **Loginc**: logarithm of household income (x_{7ij})

Also note that there are only two levels in this dataset: i denotes woman and j denotes interview. We do not have any visit-level observations, we only have women's reports of number of doctor visits prior to each interview, which will be our outcome.

Exploratory Data Analysis:
We need to learn about the structure of the data. Is everyone interviewed both before and after the reform, or are some people only interviewed once? Note that we can think of the reform variable as our time variable, since it indicates before and after reform.

```
. xtdes if numvisit<., i(id) t(reform)

      id:  3, 4, ..., 9189                                   n =       1518
  reform:  0, 1, ..., 1                                      T =          2
           Delta(reform) = 1 unit
           Span(reform)  = 2 periods
           (id*reform uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       1         1         2       2       2

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      709     46.71   46.71 |  11
      418     27.54   74.24 |  .1
      391     25.76  100.00 |  1.
 ---------------------------+---------
     1518    100.00         |  XX
```

There is a total of 1,518 women included in this data set. Of these women, 709 were interviewed both before and after reform, 418 were interviewed only after the reform, and 391 were interviewed only before the reform.

Single-level Poisson Model:
First of all, we consider conventional Poisson regression for the number of doctor visits, written as: $\log(E(Y_{ij}))=\beta_1+\beta_2 x_{2ij}+\beta_3 x_{3ij}+\beta_4 x_{4ij}+\beta_5 x_{5ij}+\beta_6 x_{6ij}+\beta_7 x_{7ij}$. In Stata, this model can be fitted using either `poisson` or `glm` command. The estimates from these two commands are identical and displayed as follows.

```
. poisson numvisit reform age educ married badh loginc summer, irr

Poisson regression                                Number of obs   =       2227
                                                  LR chi2(7)      =    1429.00
                                                  Prob > chi2     =     0.0000
Log likelihood = -5942.6924                       Pseudo R2       =     0.1073

------------------------------------------------------------------------------
    numvisit |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      reform |   .8689523   .0230968    -5.28   0.000     .8248423    .9154212
         age |   1.004371   .0013088     3.35   0.001     1.001809    1.006939
        educ |   .9894036   .0059465    -1.77   0.076      .977817    1.001127
     married |   1.042542    .029055     1.49   0.135     .9871229    1.101073
        badh |   3.105111   .0941052    37.39   0.000     2.926039    3.295142
      loginc |   1.160559   .0418632     4.13   0.000     1.081342     1.24558
      summer |   1.010269   .0408237     0.25   0.800     .9333421    1.093536
------------------------------------------------------------------------------
```

The above model assumes that the repeated doctor visit counts from the same person are independent (given the covariates), which we know is likely untrue. Thus the standard errors from this model are not trustworthy. Also, note that we did not need to include an offset in this model, since doctor visits were counted for the same interval, namely, 3 months, for all subjects at both time points. The estimated incident-rate ratio for the reform variable is 0.87, implying a population average 13% reduction in the number of doctor visits per month between 1996 and 1998 for given covariate values.

To handle the problem of overdispersion (the variance is larger than the expectation conditioned on the covariates), we may use the quasi-likelihood method. In the quasi-likelihood approach, we do not specify a statistical model (exact parametric distribution), but instead we merely specify the expectation and the variance of the counts. Specifically we specify $E(Y_{ij})$ as given above, and $Var(Y_{ij}) = \varphi E(Y_{ij})$, where φ is the overdispersion parameter. In Stata, we can use the `glm` command with the scale option to obtain maximum quasi-likelihood estimates.

```
. glm numvisit reform age educ married badh loginc summer, family(poisson)
link(log) eform scale(x2)

Generalized linear models                          No. of obs      =      2227
Optimization     : ML                              Residual df     =      2219
                                                   Scale parameter =         1
```

```
Deviance         =  7419.853221              (1/df) Deviance =  3.343782
Pearson          =  9688.740471              (1/df) Pearson  =  4.366264

Variance function: V(u) = u                  [Poisson]
Link function    : g(u) = ln(u)              [Log]
                                             AIC             =  5.344133
Log likelihood   =  -5942.69244              BIC             =  -9685.11

------------------------------------------------------------------------------
             |                 OIM
    numvisit |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      reform |   .8689523   .0482622    -2.53   0.011     .7793268    .9688851
         age |   1.004371   .0027347     1.60   0.109     .9990249    1.009745
        educ |   .9894036   .0124256    -0.85   0.396     .9653472    1.014059
     married |   1.042542   .0607123     0.72   0.474     .9300882    1.168593
        badh |   3.105111   .1966385    17.89   0.000     2.742665    3.515454
      loginc |   1.160559   .0874758     1.98   0.048     1.001173     1.34532
      summer |   1.010269   .0853037     0.12   0.904     .8561784    1.192091
------------------------------------------------------------------------------
(Standard errors scaled using square root of Pearson X2-based dispersion)
```

Here, the estimated regression coefficients are identical to the ones obtained from the previous model. However, the estimated standard errors are larger. We see from the output next to (1/df) Pearson that the overdispersion parameter is estimated as 4.366264. Comparing this value to 1 (the value when there is no overdispersion and the poisson assumption is met), we see that the data is overdispersed and should not be modeled as a Poisson distribution. The estimated standard errors from quasi-likelihood are then $\sqrt{4.366}$ =2.09 times as large as those from maximum likelihood.

Another method for handling overdispersion is via a random intercept model. All random intercept models induce overdispersion (check out lecture 11 for connection between the conditional and marginal variances), but we can include a random intercept even in a single level model on the level-1 units to model the overdispersion. The model and its implied marginal mean and variance are exactly the same as those for two-level models but the difference is that the random intercept varies between the level-1 units and hence does not produce any dependence among groups of observations. The model can be written as: $\log(E(Y_{ij}|\zeta_{ij}^{(1)}))=\beta_1+\beta_2 x_{2\text{ij}}+...+\beta_7 x_{7\text{ij}}+\zeta_{ij}^{(1)}$, $\zeta_{ij}^{(1)}|\underline{x_{ij}} \sim N(0,\tau^2)$. The (1) superscript denotes that the random intercept varies at level 1. In order to fit this model, we generate an identifier obs for the level-1 observations, and then specify obs as the clustering variable in `xtpoisson`.

```
. gen obs=_n
. xtpoisson numvisit reform age educ married badh loginc summer, i(obs) normal
> irr

Random-effects Poisson regression              Number of obs      =      2227
Group variable: obs                            Number of groups   =      2227

Random effects u_i ~ Gaussian                  Obs per group: min =         1
                                                              avg =       1.0
                                                              max =         1
```

```
                                                  Wald chi2(7)      =    272.60
Log likelihood  = -4546.8881                      Prob > chi2       =    0.0000

------------------------------------------------------------------------------
    numvisit |        IRR   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      reform |    .881623   .0466248    -2.38   0.017     .7948166      .97791
         age |   1.002419   .0026053     0.93   0.353     .9973256    1.007538
        educ |   1.005101   .0117969     0.43   0.665     .9822433     1.02849
     married |   1.084023   .0602281     1.45   0.146     .9721784    1.208735
        badh |   3.203538   .2429967    15.35   0.000     2.760985    3.717027
      loginc |   1.151836   .0840599     1.94   0.053     .9983224    1.328956
      summer |   .9576537   .0786351    -0.53   0.598     .8152943    1.124871
-------------+----------------------------------------------------------------
    /lnsig2u |  -.1143904   .0548408    -2.09   0.037    -.2218765   -.0069043
-------------+----------------------------------------------------------------
     sigma_u |   .9444097   .0258961                       .894994    .9965538
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01) =  2791.61 Pr>=chibar2 = 0.000
```

We see from the estimated standard deviation of the level-1 random intercept of 0.94 and the highly significant likelihood-ratio test that there is evidence for overdispersion. Also note that all coefficients except the intercept have population-average interpretations.

Two-level Poisson Model:

To account for the non-independence between observations from the same person, we may instead include a random intercept in the Poisson model at level-2. This model is given by: $\log(E(Y_{ij}|\zeta_{1j}))=\beta_1+\beta_2 x_{2ij}+...+\zeta_{1i}$, $\zeta_{1i}\sim N(0,\tau^2)$. As before, the parameters have both person-specific and population average interpretations. Here, the random intercept model can be obtained using `gllamm`.

```
. gllamm numvisit reform age educ married badh loginc summer, family(poisson)
link(log) i(id) eform adapt

number of level 1 units = 2227
number of level 2 units = 1518

gllamm model

log likelihood = -4643.3427

------------------------------------------------------------------------------
    numvisit |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      reform |   .9547481   .0310831    -1.42   0.155     .8957293    1.017656
         age |   1.006002   .0028266     2.13   0.033     1.000477    1.011557
        educ |   1.008646   .0127702     0.68   0.497     .9839247    1.033988
     married |   1.077896    .059554     1.36   0.175     .9672696    1.201174
        badh |   2.466857     .15192    14.66   0.000     2.186367     2.78333
      loginc |   1.097486   .0746825     1.37   0.172     .9604523    1.254071
      summer |   .8673159   .0562616    -2.19   0.028     .7637672    .9849033
------------------------------------------------------------------------------

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

```
***level 2 (id)

    var(1): .81691979 (.04972777)
------------------------------------------------------------------------------
```

The number of visits for a person at the two occasions are specified as conditionally independent given the random intercept. Interpretations of a few coefficients are given:

- On average, each woman's reporting of doctor visits decreased after the reform by 4.5%, holding all other factors constant during that time period.
- A one year increase in age within a woman was associated with a 0.6% increase in reported visits, holding reform status and all other predictors constant.
- A one year increase in education level within a woman was associated with a 0.8% increase in the number of reported visits, holding reform status and all other predictors constant.
- Women who are married on average report 7.8% more visits than unmarried women, holding all other factors constant.

As we have already discussed, we would expect that including a random-intercept at level-2 has, at least to some degree, addressed the problem of overdispersion. However, the model uses a single parameter to induce both overdispersion for the level-1 units and dependence among the level-1 units in the same cluster. Sometimes there may be additional overdispersion at level 1 not accounted for by the random effect at level 2. For instance, in the health-care reform data, there may be unobserved heterogeneity between occasions within persons because medical problems can lead to several extra doctor visits within the same 3-month period. After conditioning on the person-level random effect, the counts at the occasions are then overdispersed. The simplest approach to handling overdispersion at level 1 in a two-level random-intercept Poisson model is to use the sandwich estimator for the standard errors.

```
.gllamm, robust eform

Robust standard errors
------------------------------------------------------------------------------
    numvisit |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      reform |   .9547481   .0503036    -0.88   0.379     .8610748    1.058612
         age |   1.006002   .0031322     1.92   0.055     .9998817     1.01216
        educ |   1.008646   .0127823     0.68   0.497     .9839016    1.034012
     married |   1.077896   .0708484     1.14   0.254     .9476075    1.226097
        badh |   2.466857   .2880487     7.73   0.000     1.962236    3.101249
      loginc |   1.097486   .0956035     1.07   0.286     .9252297    1.301812
      summer |   .8673159   .0722128    -1.71   0.087     .7367263    1.021053
------------------------------------------------------------------------------

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

***level 2 (id)

    var(1): .81691979 (.0523264)
------------------------------------------------------------------------------
```

We see that the robust confidence intervals are somewhat wider than those using model-based standard errors.

Random-coefficient Poisson Model:
Now, we introduce an additional person-level random coefficient for reform. In this model, the effect of the health-care reform is different across persons. The addition of the random coefficient on reform, means that the reform fixed effect no longer has a population average interpretation. The intercept and reform effect both have person-specific interpretations, while all other coefficients may be interpreted as either population average or person-specific. The model is given:

$$\log(E(Y_{ij}|\zeta_{1i}^{(2)},\zeta_{2i}^{(2)}))=\beta_1+(\beta_2+\zeta_{2i}^{(2)})x_{2ij}+\beta_3 x_{3ij}+\beta_4 x_{4ij}+\beta_5 x_{5ij}+\beta_6 x_{6ij}+\beta_7 x_{7ij}+\zeta_{1i}^{(2)},$$

$$\begin{pmatrix}\zeta_{1i}^{(2)}\\ \zeta_{2i}^{(2)}\end{pmatrix}\sim N\left(\begin{pmatrix}0\\0\end{pmatrix},\begin{bmatrix}\psi_{11} & \psi_{12}\\ \psi_{12} & \psi_{22}\end{bmatrix}\right).$$

We can fit this random coefficient Poisson model in `gllamm` using the estimates from the random intercept model as starting values.

```
. matrix a=e(b)
. eq rc: reform

. gllamm numvisit reform age educ married badh loginc summer, ///
>   family(poisson) link(log) i(id) nrf(2) eqs(ri rc) from(a) eform adapt

number of level 1 units = 2227
number of level 2 units = 1518

Condition Number = 812.85816

gllamm model

log likelihood = -4513.8005

------------------------------------------------------------------------------
    numvisit |     exp(b)   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      reform |   .9023139    .048376    -1.92   0.055     .8123103     1.00229
         age |   1.003457   .0028304     1.22   0.221     .9979246     1.00902
        educ |   1.008889   .0128058     0.70   0.486     .9841001    1.034303
     married |   1.086858   .0640872     1.41   0.158     .9682361    1.220013
        badh |    3.02813   .2322063    14.45   0.000     2.605564    3.519226
      loginc |   1.135641   .0866071     1.67   0.095     .9779708     1.31873
      summer |   .9140484   .0741615    -1.11   0.268     .7796627    1.071597
------------------------------------------------------------------------------

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

***level 2 (id)

    var(1): .90914639 (.06767415)
    cov(2,1): -.43462173 (.07121034) cor(2,1): -.49034779

    var(2): .86413303 (.10415938)
------------------------------------------------------------------------------
```

Note that the estimated incidence-rate ratio for reform now implies an average 10% reduction in the expected number of visits per year for a given person and is nearly significant at the 5% level.