ADVANCED STATISTICAL METHODS:

PART 2: INTRODUCTION TO MULTILEVEL MODELING IN STATA

Learning objectives:
1. To understand that multilevel modeling is an important regression technique for analyzing clustered data (i.e., patients clustered in hospitals), which is commonly encountered in surgical outcomes studies.
2. To appreciate that multilevel models have many other practical applications, including profiling hospital quality and decomposing hospital-level variation in outcomes.
3. To create multilevel models in STATA and then evaluate the usefulness of a random effects model to determine how much hospital-level variation in outcomes after cardiac surgery is explained by patient risk factors.

MULTILEVEL MODELS IN STATA:

Open the new dataset and summarize the data

For this analysis, we will use a modified version of the Maryland coronary artery bypass surgery dataset used in earlier labs (Maryland.CABG.2001_hospital.dta). This new dataset has hospital-level variables that are necessary for this exercise. We will be creating a multilevel model with 2 levels: 1) patient and 2) hospital. The patients are clustered within 10 hospitals and we will use a hospital identifier to specify this relationship in our laboratory exercise.

Type the command:

```
summarize
```

STATA output:

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>key</td>
<td>4668</td>
<td>2.42e+13</td>
<td>0</td>
<td>2.42e+13</td>
<td>2.42e+13</td>
</tr>
<tr>
<td>age</td>
<td>4668</td>
<td>65.78535</td>
<td>10.73579</td>
<td>16</td>
<td>94</td>
</tr>
<tr>
<td>atype</td>
<td>4650</td>
<td>2.172258</td>
<td>.8960876</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>died</td>
<td>4661</td>
<td>.0281056</td>
<td>.1652922</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>female</td>
<td>4668</td>
<td>.3018423</td>
<td>.4591064</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>los</td>
<td>4668</td>
<td>8.430805</td>
<td>7.654034</td>
<td>0</td>
<td>114</td>
</tr>
<tr>
<td>pay1</td>
<td>4644</td>
<td>1.893196</td>
<td>1.028039</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>pr1</td>
<td>0</td>
<td>0</td>
<td></td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Npr1</td>
<td>4668</td>
<td>3613.051</td>
<td>1.207581</td>
<td>3610</td>
<td>3619</td>
</tr>
<tr>
<td>race</td>
<td>4654</td>
<td>1.569618</td>
<td>1.38387</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>totchg</td>
<td>4668</td>
<td>26447.18</td>
<td>21297.32</td>
<td>4422</td>
<td>355980</td>
</tr>
<tr>
<td>hosp</td>
<td>4668</td>
<td>5.405955</td>
<td>2.582088</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>volume</td>
<td>4668</td>
<td>662.6045</td>
<td>239.8957</td>
<td>1</td>
<td>999</td>
</tr>
</tbody>
</table>
You should notice two new variables, hosp and volume, which represent the hospital number (1 to 10) and the annual hospital volume (range 1 to 999), respectively.

**Exploring the hospital volume mortality relationship**

We will first explore the relationship between hospital volume and mortality in this dataset. To obtain a rough idea of whether volume is important, we will divide hospitals into two groups, high and low volume. Begin by summarizing the volume variable.

Type the command:

```
summarize volume
```

**STATA output:**

```
. sum volume

 Variable |       Obs        Mean    Std. Dev.       Min        Max
---------+--------------------------------------------------------
    volume |      4668    662.6045    239.8957          1        999
---------+--------------------------------------------------------
```

We can then create a new variable, highvol, using the mean as a cutoff and tabulate the results.

Type the commands:

```
gen highvol=1 if volume>662
replace highvol=0 if highvol ~=1
tab highvol
```

**STATA output:**

```
. tab highvol

            highvol |      Freq.     Percent        Cum.       
----------------+----------------+--------------------+---------
           0 |      2,856       61.18       61.18       
           1 |      1,812       38.82      100.00       
----------------+----------------+--------------------+---------
            Total |      4,668      100.00       
```

This output shows that 39% of patients have surgery in high volume hospitals, as defined by a volume above 662 cases.
Next we will determine whether high volume hospitals have lower mortality rates in Maryland.

Type the commands:

\texttt{tab died highvol, chi col}

**STATA output:**

<table>
<thead>
<tr>
<th>Died during hospitalization</th>
<th>highvol</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
<td>1</td>
<td>Total</td>
</tr>
<tr>
<td>Alive</td>
<td>2,763</td>
<td>1,767</td>
<td>4,530</td>
</tr>
<tr>
<td></td>
<td>96.98</td>
<td>97.52</td>
<td>97.19</td>
</tr>
<tr>
<td>Dead</td>
<td>86</td>
<td>45</td>
<td>131</td>
</tr>
<tr>
<td></td>
<td>3.02</td>
<td>2.48</td>
<td>2.81</td>
</tr>
<tr>
<td>Total</td>
<td>2,849</td>
<td>1,812</td>
<td>4,661</td>
</tr>
<tr>
<td></td>
<td>100.00</td>
<td>100.00</td>
<td>100.00</td>
</tr>
</tbody>
</table>

\texttt{Pearson chi2(1) = 1.1613 Pr = 0.281}

The results show that high volume hospitals have lower mortality (2.5% vs. 3.0%) but this result does not reach statistical significance in a chi square test (P=0.281). However, this test of significance is based on a dichotomous volume variable, which is not ideal. When you make a continuous variable into a dichotomous variable you lose information. To test the true significance of volume, we will generate a new variable log(volume), as this is the typical relationship between volume and mortality.

Type the commands:

\texttt{gen logvol=log(volume)}

Now we will evaluate the significance of logvol using simple logistic regression.

Type the commands:

\texttt{logistic died logvol}
In this analysis, hospital volume has a statistically significant relationship to mortality (P=0.044). But this relationship is relatively weak, especially compared to other procedures (e.g., esophagectomy and pancreatectomy), which is consistent with the published literature.

Creating a multilevel model

We will now introduce the commands for creating multilevel logistic regression models in STATA. The basic command is `xtmelogit`. We will first create a model that includes no fixed effects (i.e., no patient characteristics) and a hospital random effect:

Type the commands:

```
xtnelogit died || hosp:
```
In this output, the data element of greatest interest is the standard deviation of the random effect, sd(_cons), which = 0.341. However, we are actually more interested in the hospital-level variance. We can change the command to calculate the variance of the random effect by adding ", var" to the command:

Type the commands:

```
xtmelogit died || hosp:, var
```

STATA output:

Mixed-effects logistic regression               Number of obs      =      4661
Group variable: hosp                            Number of groups   =        10
Obs per group: min =         1
               avg =     466.1
               max =       999
Integration points =   7                        Wald chi2(0)       =         .
```

```
                 died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-----------------+----------------------------------
          _cons |  -3.554464   .1514625   -23.47   0.000    -3.851325   -3.257603
```

```
                     Random-effects Parameters |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------+-----------------------------------------
        hosp: Identity |                     var(_cons) |   .1165482   .1090072     .0186371    .7288397
```

```
LR test vs. logistic regression: chibar2(01) =     3.72  Prob>=chibar2 = 0.0269
```

The variance of the random effect is 0.1165. In the following analyses, we will be evaluating how much the hospital-level variance declines when adding additional variables, such as patient risk factors and hospital volume.

Patient risk factors and hospital level variance

First, we will create a patient risk score that combines all important risk factors into a single number. To do this, we will create logistic regression model including all relevant patient variables. Since many of the patient factors are categorical (not dichotomous or continuous) we will use the “xi:” modif

Type the commands:

```
xi: logistic died age i.atype female i.pay1 i.race
```
This output shows that age, admission type, payer status, and race are all important risk factors for mortality after coronary artery bypass surgery. We can generate a risk score based on these variables using the “predict” command. However, we will use the “xb” option (log(odds of death) rather than the default, which provides a predicted probability. The reason for this modification is that probability has a sigmoidal relationship to mortality, whereas the log(odds) has a linear relationship and is normally distributed. The linear predictor will therefore be a better score for inclusion in the multilevel model.

Type the commands:

```
predict ptrisk, xb
sum ptrisk
hist ptrisk
```
Note the normal distribution of the risk score. (To compare these results to the predicted probability you can repeat the previous steps without the “xb” option).

Next we will construct a multilevel model with the “ptrisk” score included as an independent variable:

Type the commands:

```
xtneglogit died ptrisk || hosp:, var
```

**STATA output:**

```
Mixed-effects logistic regression
Number of obs      =      4564
Number of groups   =        10
Obs per group: min =         1
               avg =     456.4
               max =       996
Integration points =   7                        Wald chi2(1)       =     53.11
Log likelihood =   -563.76721                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
  died |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
ptrisk |   .9900761   .1358526     7.29   0.000     .7238099    1.256342
   _cons |  -.0338985   .4688941    -0.07   0.942    -.952914     .885117
------------------------------------------------------------------------------

Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+-----------------------------------------------
hosp: Identity               |                                
            var(_cons) |   .0635111   .0801197      .0053586    .7527412
------------------------------------------------------------------------------
LR test vs. logistic regression: chibar2(01) =    1.18 Prob>=chibar2 = 0.3185
```
In these results, the hospital level variance has decreased from 0.1165482 to 0.0635111 after the addition of the patient risk score. In other words, a large fraction of hospital level variance (difference in mortality across hospitals) can be explained by patient characteristics. We can calculate the exact proportion in STATA using “display” and a simple formula:

Type the commands:

```
display (.1165482 - .0635111)/.1165482
```

STATA output:

```
. display (.1165482 - .0635111)/.1165482
   0.4550658
```

This means that 46% of hospital variation in mortality rates can be explained by patient characteristics.

**Exercise:** You can now use this same method to determine the proportion of hospital level variation in mortality explained by hospital volume (using the log(volume) variable).

**Note:** We use a single patient risk score because including all of the variables in the multilevel regression makes the model much more complex and it may fail to find a solution. (Note: multilevel logistic models have no closed form mathematical solution so the solution is an approximation and they are prone to failure when a complex set of input variables is used).

**Other useful knowledge:** In the command xtmelogit “me” stands for mixed effects modeling. “Mixed effects” simply means that variables can be modeled as fixed or random components. Before you get too confused, fixed effects are nothing new – standard regression models are all fixed effects models. The only new capability is the use of random effects, which is why these models are often referred to as “random effect models”. Multilevel linear models can be created using the same syntax using the command xtmixed.