Biostat 209 Lab

Repeated Measures 3

As before – please work through the lab and lab quiz. Feel free to ask question via audio and/or chat. Happy to field questions about lab, lecture, homework or projects.

Lab Summary

The purpose of this lab is to get practice using the command `xtgee`. Download the GAbabies.dta dataset from the course web site and start a log file. Recall that this dataset follows successive birthweights of infants to mothers (each of whom had five children) from vital statistics in Georgia. We are going to be interested in whether birthweight increases with birth order and the mothers' age. The variables in the dataset are:

1) Mother's ID number (`momid`)
2) Birth order (`birthord`)
3) Mother's age at birth of infant (`momage`)
4) Mother's age at birth of first infant (`initage`)
5) Change in mother's age from first infant to current infant (`timesnc`)
6) Birthweight of infant (bweight)
7) Change in birthweight of infant from first infant to current infant (delwght)
8) Whether the birthweight is under 3000g or not (lowbirth).

Linear regression with clustered data

Let's start with a simple regression analysis using the predictors birth order and age at first birth.

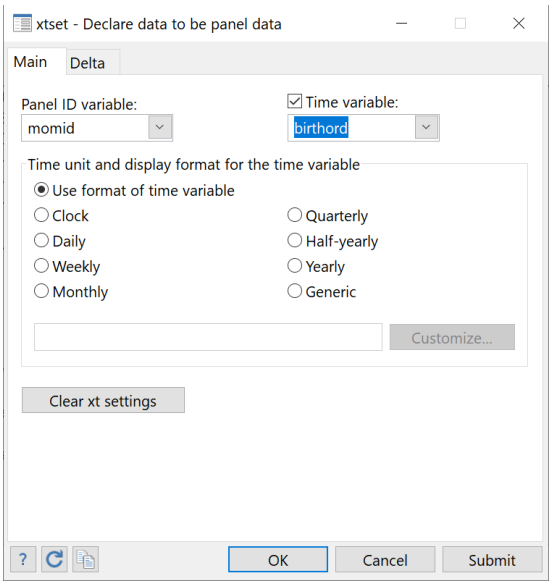
```
regress bweight birthord initage
```

As we noted previously, this is incorrect because we have not accounted for the clustering by mom. Next, let's see how the `xtgee` command can mimic the regression command.

First we need to tell Stata how the repeated measures are organized, namely as repeated measurements over birth order. Go through the menus:

Statistics > Longitudinal/Panel Data >Setup and Utilities > Declare dataset to be panel data

and fill in the Panel ID and Time variables as follows:



The `xtgee` command is given below, and the `corr(indep)` portion of the command tells STATA to pretend the data are independent even though they are clustered by mom.

```
xtgee bweight birthord initage, i(momid) corr(indep)
```

Or – navigate through the menus and choose an independence correlation structure.

Statistics > Longitudinal/Panel Data > Generalized estimating equations > GEE

choosing Gaussian and Identity and, on the Correlation tab, independent. How do the estimates and standard errors compare to the straight regression? Next give the command

```
xtcorr
```

or use the menu:

Statistics > Longitudinal/Panel Data > Generalized estimating equations > Display estimated within group correlations

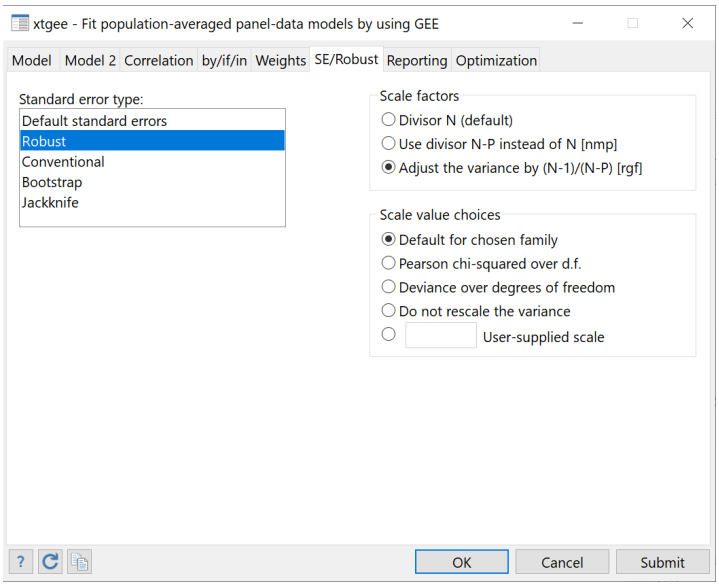
to view the correlation structure STATA was using. Now let's change the `corr(indep)` portion of the command to the (default) of exchangeable to see the impact.

```
xtgee bweight birthord initage, i(momid) corr(exch)
xtcorr
```

What is the form of the correlation structure invoked with "exchangeable"? Did it make a difference with regard to the estimates or standard errors or tests? Next, let's use an option called "robust". The correlation structure requested by `corr(·)` is then called the "working" correlation structure. When the `robust` option is used, the working structure is used for some intermediate calculations (the coefficients), but the correlation within the clusters is estimated from the data. A nice feature of this option is that inferences remain valid even if the assumed correlation structure is wrong. It is not a free lunch however. It works best with lots of small clusters and can give inaccurate estimates when there are only a few clusters or the clusters are large.

```
xtgee bweight birthord initage, i(momid) corr(exch) robust
xtcorr
```

Or by using the SE/Robust tab in the GEE menu:



In the above menu I have also checked the "rgf" option under scale factors, which provides a small sample adjustment like the t-test instead of a z-test, which makes a difference if the number of clusters is small.

The robust option is similar in spirit to assuming nothing about the correlation structure ("unstructured"). However, using "unstructured" first estimates all the pairwise correlations, which can be wasteful. Robust proceeds directly to calculating standard errors of coefficients without that intermediate calculation:

```
xtgee bweight birthord initage, i(momid) t(birthord) corr(uns)
```

Use the "Correlation" tab to specify "unstructured" as done above to specify independent or exchangeable

```
xtcorr
```

Finally, try the correlation structure "autoregressive" and specify order equal to 1 . Now, stop and take an assessment of all these fits, focusing on the coefficients and standard errors of birthorder. Did the estimates vary much? Which methods made the most difference in the standard errors and tests? In particular, the unstructured fit might imply that the correlations aren't all the same. How did the exchangeable structure do compared to the unstructured or robust fits?

Logistic regression with clustered data

Now let's try using `xtgee` to do a logistic regression.

Load the backpain data from the web site. It is a hypothetical version of the data from the Korff, et al article handed out in class. The variables we will be interested in are:

1. Doctor ID number (`doctor`).
2. Whether or not the patient understood the treatment (`undrstnd`).

3. Age in years (`age`).
4. Education divided into three categories (0 means <12 years, 1 means 13-16 years, 2 means ≥ 16 years education). (`educ`).

Using `undrstnd` as the dependent variable, fit a logistic regression model with predictors age (as a continuous variable) and education (as categorical):

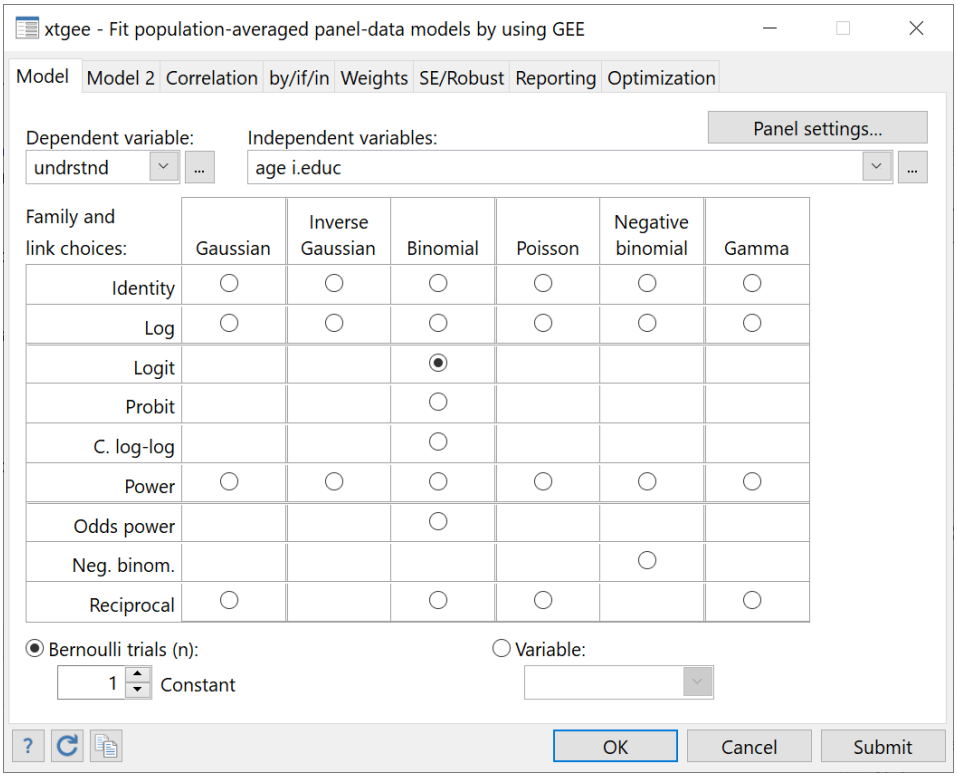
```
logit undrstnd age i.educ
```

Now let's take account of the clustering using the `xtgee` command. First `xtset` your data:

Then run the GEE command:

```
xtgee undrstnd age i.educ, i(doctor) family(binomial)
```

Use the "Model" tab to specify logit.:



We do see a couple of important differences from the previous use of `xtgee`. The `family(binomial)` option tells STATA that the data are binary and also invokes the logistic regression model. How do the results differ between the `xtgee` and `logit` commands?

Recall that the main point of the analysis is to determine the impact of practice style on the outcome variables (`cost`, `actlim`, and `undrstnd`). The other variables are variables we may wish to adjust for to equalize case mix among the doctors. Let's continue with the response `undrstnd`. Recall the basic form of the `xtgee` command for logistic regression (where `ef` causes the odds ratios to be printed):

```
xtgee undrstnd pred vars, i(doctor) corr(work corr) family(bi) ef
```

Which working correlation structure is the most reasonable one to use in this case? Use the `xtgee` command (and statistical significance) to decide which predictor variables to include in your final model (You don't have to consider interaction terms). Is practice style statistically significant? Provide an interpretation of the effect of practice style.

When you have reached a final model, compare the use of the logistic command. How do the coefficients compare? How does the test of practice style compare?

Again, working from your final model, change the link to a log link, so as to get a relative risk (as opposed to an odds ratio) interpretation. Specify the distribution as Poisson , which will default to a log link, and make sure to use a robust variance estimate. What is the interpretation of the practice style coefficients? How do the p-values compare to the xtgee-logistic? Why are the estimates so different?

Other regression models

What about days of activity limitation? What does the distribution of this look like? Generate a boxplot:

```
graph box actlim
```

There are two basic ways to treat this: 1) transform the data or 2) find a different distribution to describe this kind of data.

Here is a transformation approach:

```
generate sqrtal=sqrt(actlim)
xtgee sqrtal age i.educ thoraic i.pracstyl, i(doctor)
```

To find a different distribution, use the help command or the menus to find the options for `xtgee`. Which one do you think is appropriate? Try that one and compare to the above. Which do you believe?