ATCR Lab, 4/28/16

Repeated Measures 1

Lab Summary

The purpose of this lab is to cover basic manipulation of longitudinal data and to indicate how some longitudinal analyses fit with more standard analyses. Before you start this tutorial, download the GAbabies dataset and the OAI datasets from the course web site and start a record of your STATA session with a log file. This Georgia babies 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 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).`

Getting started: descriptive statistics

First use the `xtset` command to tell Stata about the hierarchical and longitudinal nature of the dataset:

```
xtset momid birthord
```

Next describe the longitudinal pattern in the data.

```
xtdes
```

Next, use the `xtsum` command to understand the within and between mom variation in some of the variables:

```
xtsum birthord momage initage bweight
```

Why are some of the entries zero for the standard deviations? Notice any data errors? Finally, look at some of the individual birthweight trajectories to get a sense of the data. The "`if`" statement restricts the plots to a reasonable number of moms

```
twoway (connected bweight birthord if momid<2500), by(momid)
```

Relationship of birthweight to birth order

We'll start by looking at descriptive summaries. Does birthweight appear to be related to birth order?

```
twoway scatter bweight birthord
```

This is a bit hard to judge. A slightly better method is to fit a smooth curve through the data. Recall our way of drawing a smooth curve:

```
lowess bweight birthord, bw(0.4)
```

What does this tell you about the relationship? How comfortable would you feel treating birth order as a continuous variable and using a linear relationship?

Some people prefer tabular summaries over graphical. Use the `table` command to generate descriptive statistics by birth order. The format command isn't necessary but makes the table easier to read

```
. table birthord, c(mean bweight sd bweight n bweight) format(%6.0f)
```

Does this seem consistent with the graph?

Analyses

Let's proceed to a more formal analysis by regressing bweight on birthord:

```
regress bweight birthord
```

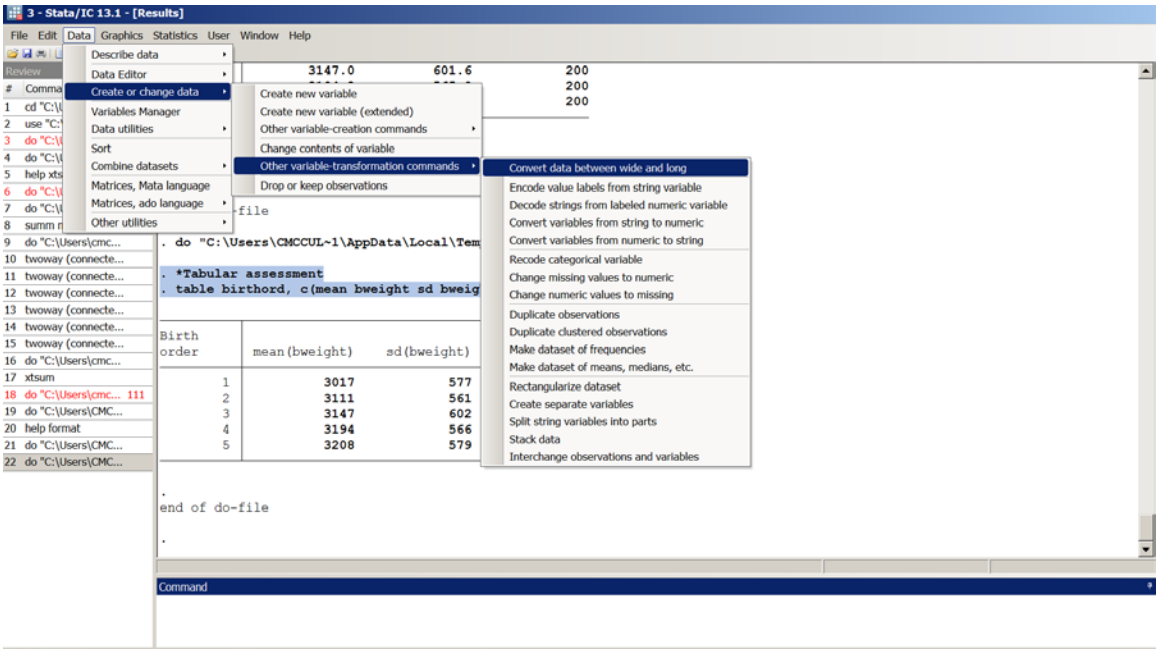
Is there a statistically significant relationship? Is there an alternate explanation for why birth order might not be causally associated with birth weight that could explored using this data? How would you check?

An alternate way to compare changes in birth weight with order is to look at the difference between the last and the first birth and conduct a paired t-test. Unfortunately, the data are not in the right format to easily do so. The data are currently listed with one observation per child (i.e., each child is a row in the data matrix). To subtract the birth weights for the first and last child we need one row per mom (why?). Fortunately, STATA has a simple command to rearrange the data. It is `reshape` and it can be used to take data in the current format (called the "long" format) and put it into a format with one data row per mom (the "wide" format). Or the reverse.

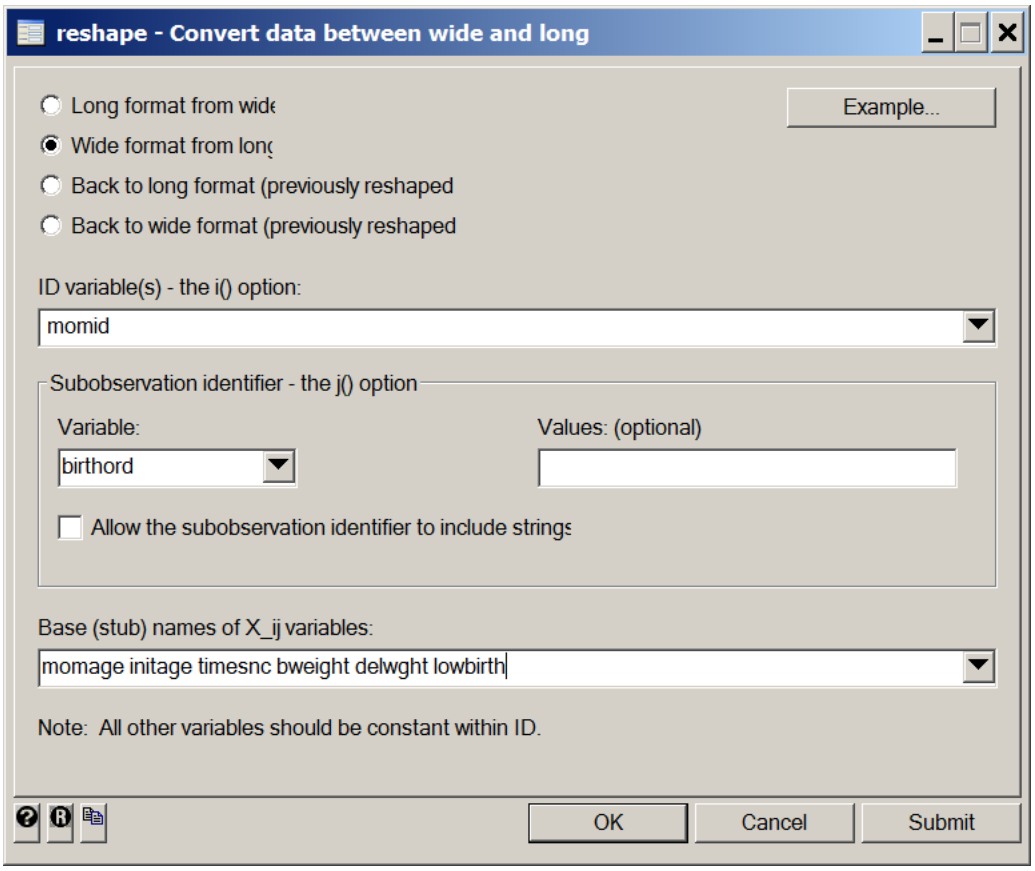
```
reshape wide momage timesnc bweight delwght lowbirth, i(momid) j(birthord)
```

In the text form of this command we first say what type of reformatting we want (`reshape wide` or `reshape long`), then we list the variables that are *not* constant within a cluster (in this case a cluster is a mom). After the comma goes the cluster variable, designated inside the i(·), and the variable giving the order within a cluster, designated inside the j(·). It may be easier to do this through the menus. To navigate through the menus use:

Data > Create/change variables > Other variable transformation commands > Convert data between wide/long



Then fill in the reshape window as follows and click on Submit:



Now we can easily calculate the differences between the last and first birthweights and conduct a t-test.

```
generate bwdiff=bweight5-bweight1
ttest bwdiff=0
```

How would you expect this to compare to the regression, given that the t-test ignores the three intermediate births? How do the t-statistics (and hence p-values) for testing birth order compare?

This "wide" format also makes it easier to understand the relationships between the repeated measures. Here is the graph showing the association of the five birth weights that we saw in class:

```
graph matrix bweight1 bweight2 bweight3 bweight4 bweight5
```

And a numerical summary:

```
corr bweight1 bweight2 bweight3 bweight4 bweight5
```

If you want to go back to the "long" format, all you have to do is to give the command

```
reshape long
```

or go back to the reshape menu and click on "back to long format".

Now let's analyze the data. For now the important thing to know is that the command below performs a regression of birth weight on birth order, taking account of the clustering on mom:

```
mixed bweight birthord || momid:, reml
```

(Don't forget the colon after momid).

How does the p-value compare to the t-test and the regression? Does it make sense? What is the relationship between the birth order coefficient in `mixed` and average value of BWDIFF? An alternative analysis method is with the xtgee command. Compare the results from that analysis:

```
xtgee bweight birthord
```

OAI dataset

Let's move on to a different data set. The purpose of this analysis is to show how more standard analyses fit with some simple repeated measures analyses. Open the OAI (www.oai.ucsf.edu) dataset. The variables in the dataset are:

1. ID (participant ID)
2. Visit (baseline = 0 months, visit 1 = 12 months)
3. Age at the visit
4. Sex of the participant
5. xr_koa = evidence of knee osteoarthritis on Xray at baseline
6. sx_koa = xr_koa with reported symptoms
7. WOMAC pain score (measures pain on a scale from 0-50, higher being worse)
8. Body mass index (BMI).

We are going to compare the change in pain scores in men and women. First get a sense of the data by generating some descriptive statistics.

```
table visit sex, c(mean womac n womac)
```

What is the change in pain score in women? In men? And what is the difference in the changes? Test your statistical intuition: do you expect the changes to be statistically significantly different between men and women?

Analysis of difference scores

A simple and effective analysis is to calculate the difference scores within a person and compare those using a t-test. Reshape the data to wide, calculate the difference scores and compare the difference scores between men and women using a t-test. How does this compare to the descriptive statistics?

(after reshaping wide)

```
gen ch_womac=womac_pain12-womac_pain0
ttest ch_womac, by(sex)
```

Analysis using hierarchical methods

Reshape back to long and tell Stata about the data using `xtset` and use `mixed` to perform a hierarchical analysis:

```
mixed womac_pain visit##sex || id:, reml
```

Why did we need to include the interaction? How does this compare to the t-test? Compare the `xtgee` command to `mixed` and the t-test.

Analysis adjusting for baseline

Some people argue that, instead of analyzing change scores, one should adjust for the baseline value. Let's go back to the wide format and try this.

```
reshape wide
```

As a basis for comparison consider another way to do the t-test above, namely with a regression command:

```
regress ch_womac i.sex
```

Now check to see what we get with two ways to adjust for baseline pain score:

```
regress womac_pain12 i.sex womac_pain0
```

Some people like a minor variation in that they use the change score as the outcome and adjust for baseline values:

```
regress ch_womac i.sex womac_pain0
```

Focusing on the sex effect, how do these analyses compare to the above analyses?

Morals:

1. In this simple scenario, the longitudinal analysis gives exactly the same results as the t-test on the difference scores. Reassuring. In this simple situation, why do anything else? In more complicated situations, however, like with missing data and multiple visits, the longitudinal analysis uses exactly the same command for two or more time points and is much easier. It is also a way to deal with observations that are unequally spaced in time.
2. Adjusting for the baseline value of the outcome is rarely a good idea in an observational study. It gets you away from analyzing changes over time and can generate spurious results. Use with care!