Biostatistics 140.655
Longitudinal Data Analysis
Laboratory 1-2 – Notes

Topics:

- **Reading/Loading Data**
- **Plotting Data and Exploring Correlation**
- **Ordinary and Weighted Least Squares**

Reading/Loading Data

- Stata: infile (from text file), insheet (from spreadsheet), cut and paste

- SAS: infile (from text file)

```
data dataname;
        infile 'filename';
        input varnames;
```

Data sets are available at network neighborhood in both Stata and Text formats (see handout). To infile for SAS one should delete the variable names from the first row of the text files, then use the input option.

Examining Data

- Stata: use the editor; table or tabulate give frequencies

- Sas: PROC Print, PROC Freq

See also: Sas and Stata intro handouts.

Plotting Data / Exploring Correlation

- **Strength**
- **Time Dependence**
- **Inference Concerning Interaction (Effect Modification)**

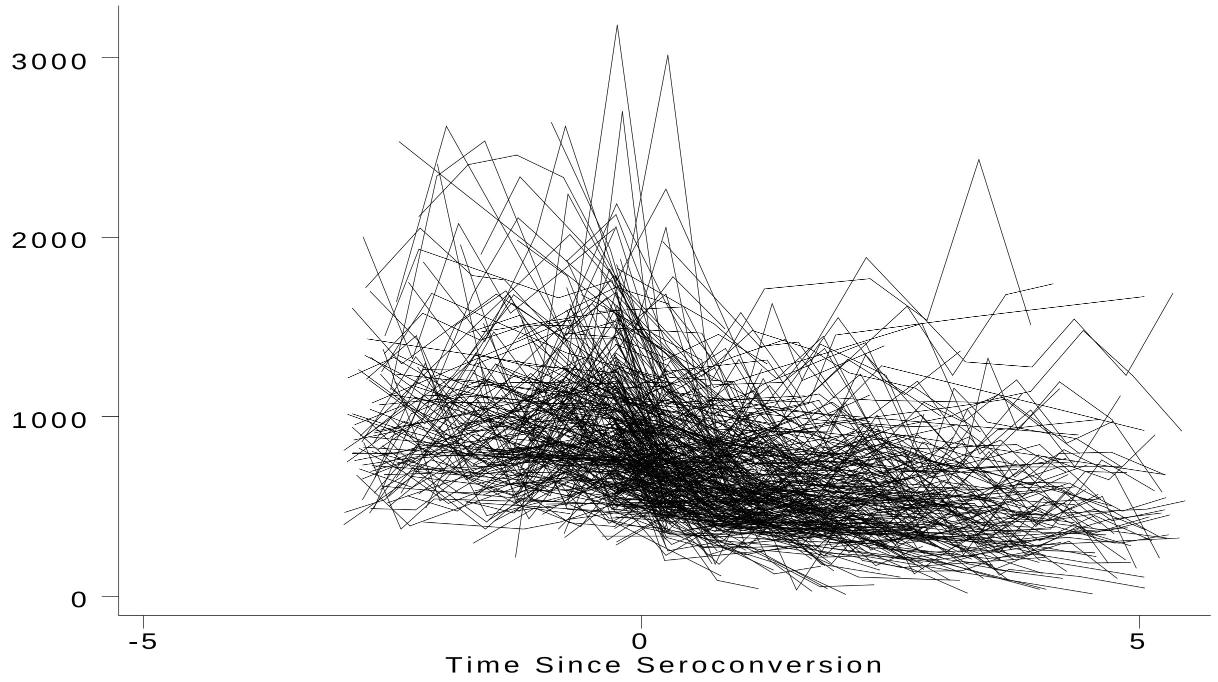
Creating a Longitudinal Plot (Stata)

- **Graph command (see Stata Handout)**
- **Sort Data Points within Sampling Unit, then by time**
- **Connect Points within sampling Unit**

Syntax:

form: command varlist, options

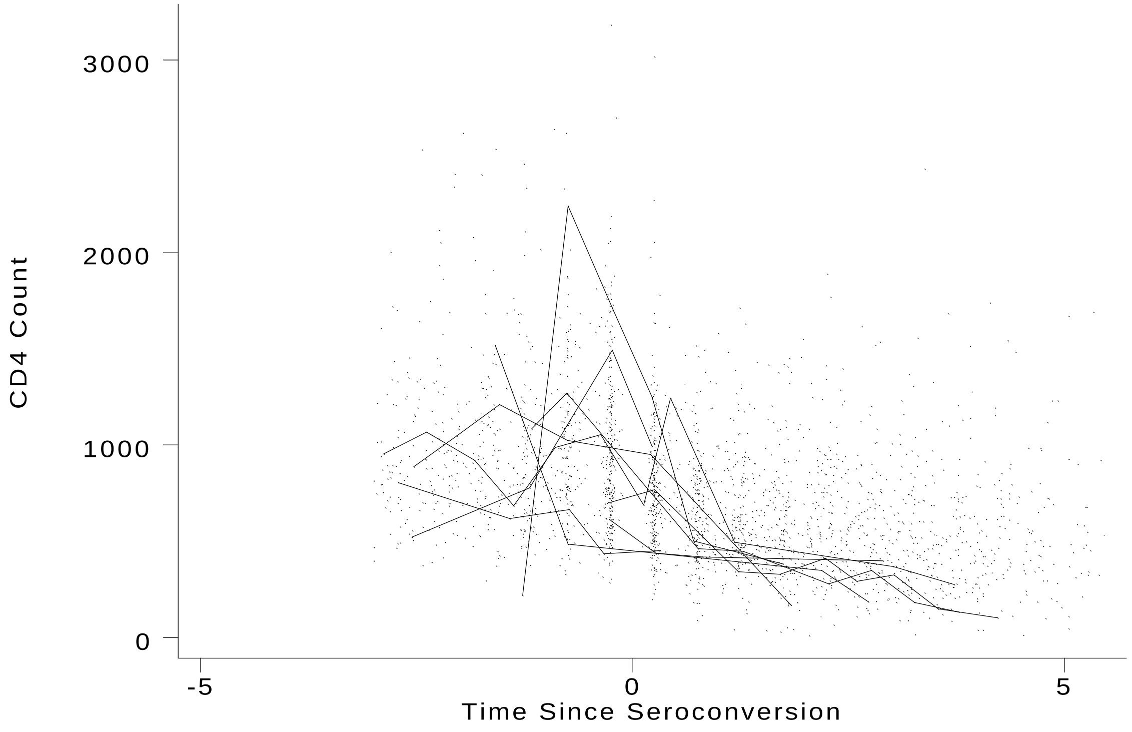
Label var time "Time Since Seroconversion" (labels time variable)



```
sort id time
graph count time, connect(L) symbol(i) xlab ylab
```

The graph above may not be very useful - it's too noisy.

Make a graph connecting only the individuals at the deciles of the distribution of starting CD4 count:



Syntax: (this is a complicated instance - the time points are not the same for all individuals)

sort id time   * sorts by time inside individuals

gen first = id!=id[_n-1]   *is it a new individual? Pick out each person's first observation

sort first count   *group all 'first' observations together; within this group, order people according to CD4 count

```
gen decile = first & (int(_n/37)==_n/37)
```

* this is the toughest part - an observation is a "decile" obs. if it's a person's first observation AND the observation number (_n) is an even multiple of 37 (there are 369 "first" observations)   The "int" function rounds the result of what's in the parenthesis (_n/37) to the nearest integer. Hence int(_n/37) will only be the same as _n/37 if _n/37 is a whole number.

This statement is strongly dependent on the way we have ordered the data and our knowledge of the number of individuals.

```
sort id time  * re-order the data by individual
```

```
replace decile=decile[_n-1] if id==id[_n-1]  *now that we know who are our
decile people, mark all of their observations that way.
```

```
gen count2 = count if decile  *make a variable of only the decile observations
```

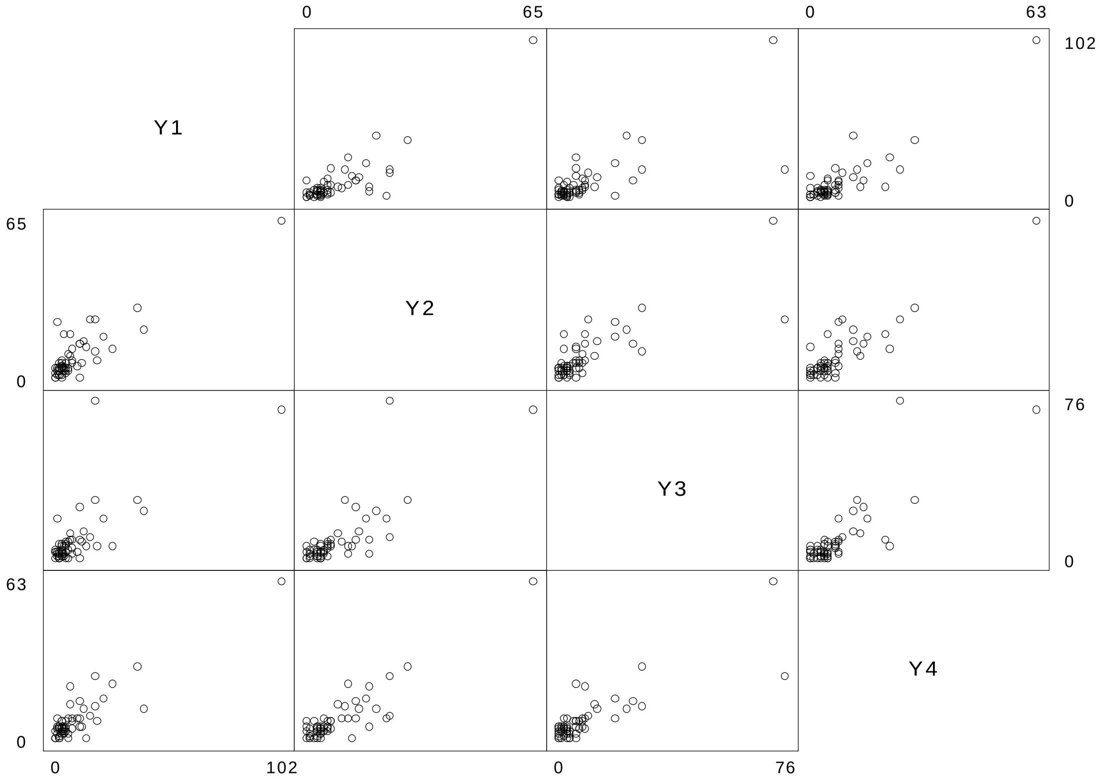
```
graph count count2 time, s(.i) c(.L) xlab ylab  *make the plot
```

Notice that we had to create an extra variable (count 2) so that we could connect the decile points and not the others.

WHEW!!

SAS: similar difficulties. Explore PROC GPLOT.

Scatterplot Matrix: (seizure data)



To construct this plot, you need the data in WIDE format (see data handout).

syntax:

graph varlist, matrix

in this case:

graph Y1 Y2 Y3 Y4, matrix

Further exploration: Autocorrelation function.

Stata function autocor.ado calculates and plots autocorrelation function (see handout).

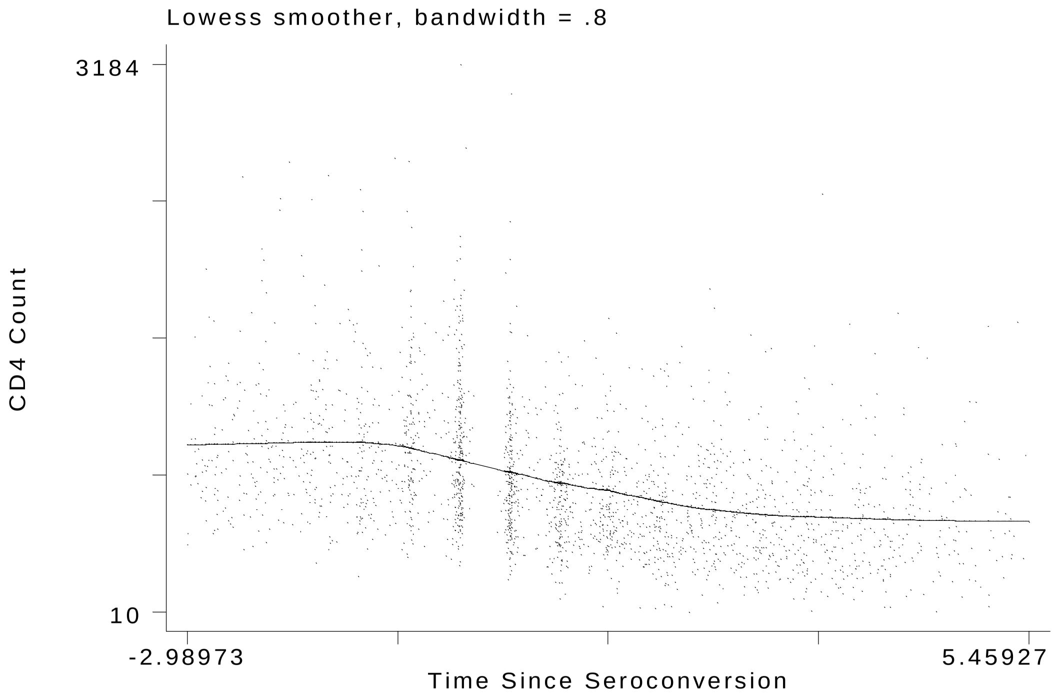
Residuals Plotting:

For some estimation commands, use PREDICT to construct residuals.

Then plot residuals vs. fitted values, predictors.

Graph Smoothing

- Various methods
- Stata: ksm statement: moving average, lowess smoothing
- SAS: PROC GPLOT, Symbol and smoothing options.



Least Squares Estimation

Simple linear regression commands utilize OLS, including regress, fit (Stata) and PROC REG (SAS). Weighted least squares is an option in most estimation commands in SAS; specifically variance weighted least squares can be accomplished using **vwls** in Stata.