Lab 5 & 6: Stata, graphics, and data

Robin Mejia

September 30, 2015

Reading and saving data in Stata

```
//remove any existing data and variables
clear

// Print working directory (check what the current working directory is)
pwd

// Set the working directory you want -- use your own here, not mine
cd "/Users/robinmejia/Dropbox/Long Data Fall 2015/Assignments/Assignment3"

// read in data using insheet comand
insheet using strength.csv

// Do things to your data if you want...

// Save a file called strength.dta (Stata format) in your working directory
save strength, replace
```

Schizophrenia Data Set

- There is a file called schizophrenia.dta in bCourses. Please download it and double click to open it in Stata
- 5 variables:
  - id: 437 people, measured 2 to 4 times each
  - week: discrete scale (0, …, 6)
  - gender: female/male (0, 1)
  - drug: untreated/treated (0, 1)
  - severity: discrete scale (1, …, 7)
- Taken from Robert Weiss's website, a good source for longitudinal datasets:
  - https://faculty.biostat.ucla.edu/robweiss/book-data-sets

Goals Today

- Perform basic exploratory data analysis on the relationship between drug status and severity of episodes.

- There is another binary variable in this dataset, gender. Please feel free to explore the code here by adapting it to the relationship between gender and severity of episodes.

One Dimensional Summaries

- How many unique individuals exist here?
- Of those, how many are being treated and how many are not?
- How many females and males?
- What is the distribution of observations over weeks?

One Dimensional Summaries

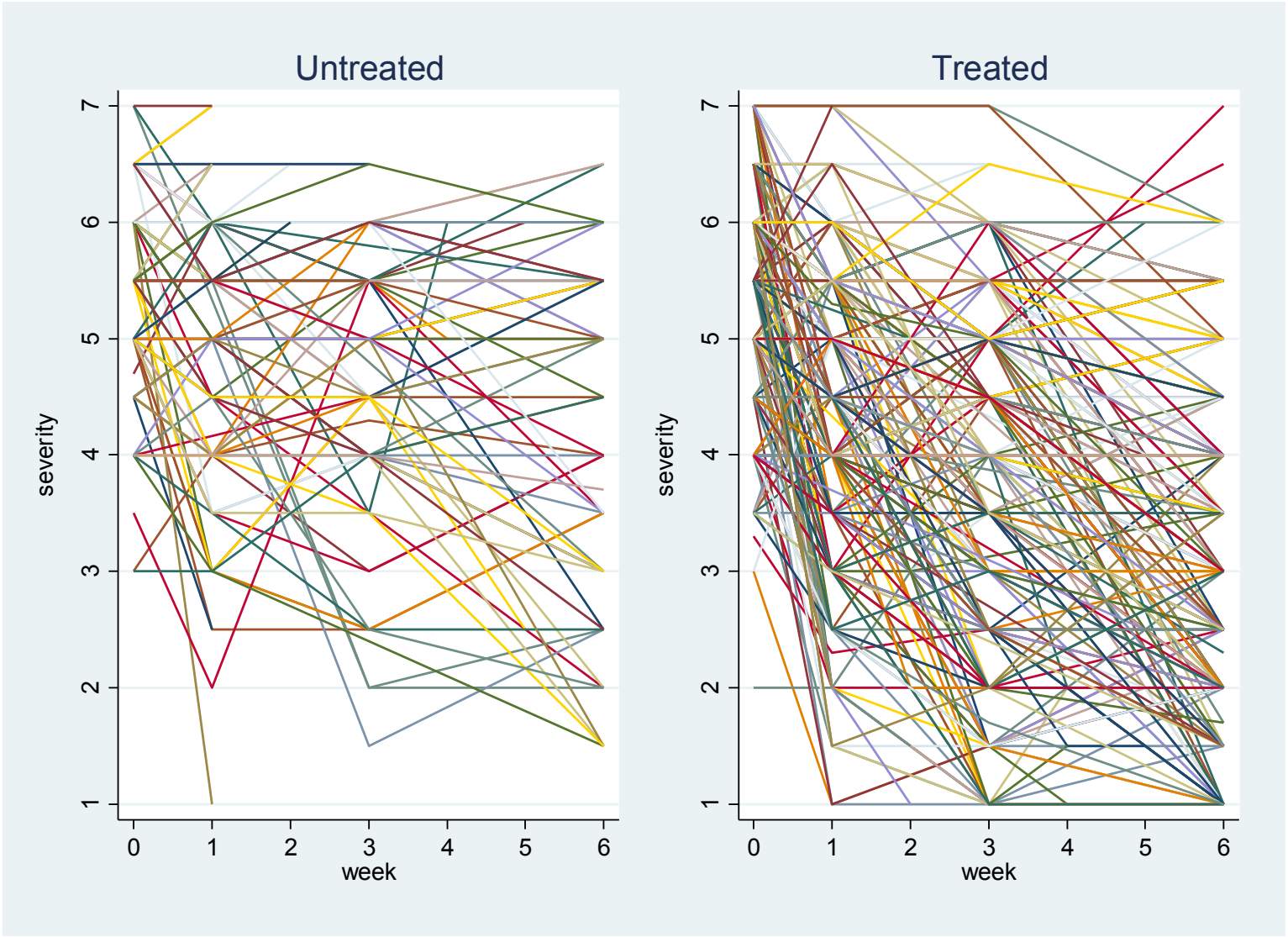
- xtset id week
- xtdescribe

Plot severity vs week for 2 groups

// This code wil plot severity by week for the treated and untreated.
// Try it and play with the settings to see what happens

```
xtline severity if drug==0, i(id) overlay t(week) xlab(0(1)6) ylab(1(1)7)
title(Untreated) legend(off)

graph save "untreated.gph", replace

xtline severity if drug==1, i(id) overlay t(week) xlab(0(1)6) ylab(1(1)7)
title(Treated) legend(off)

graph save "treated.gph", replace

gr combine "untreated.gph" "treated.gph"
graph save "untreated_treated.gph", replace
```

Scatterplot



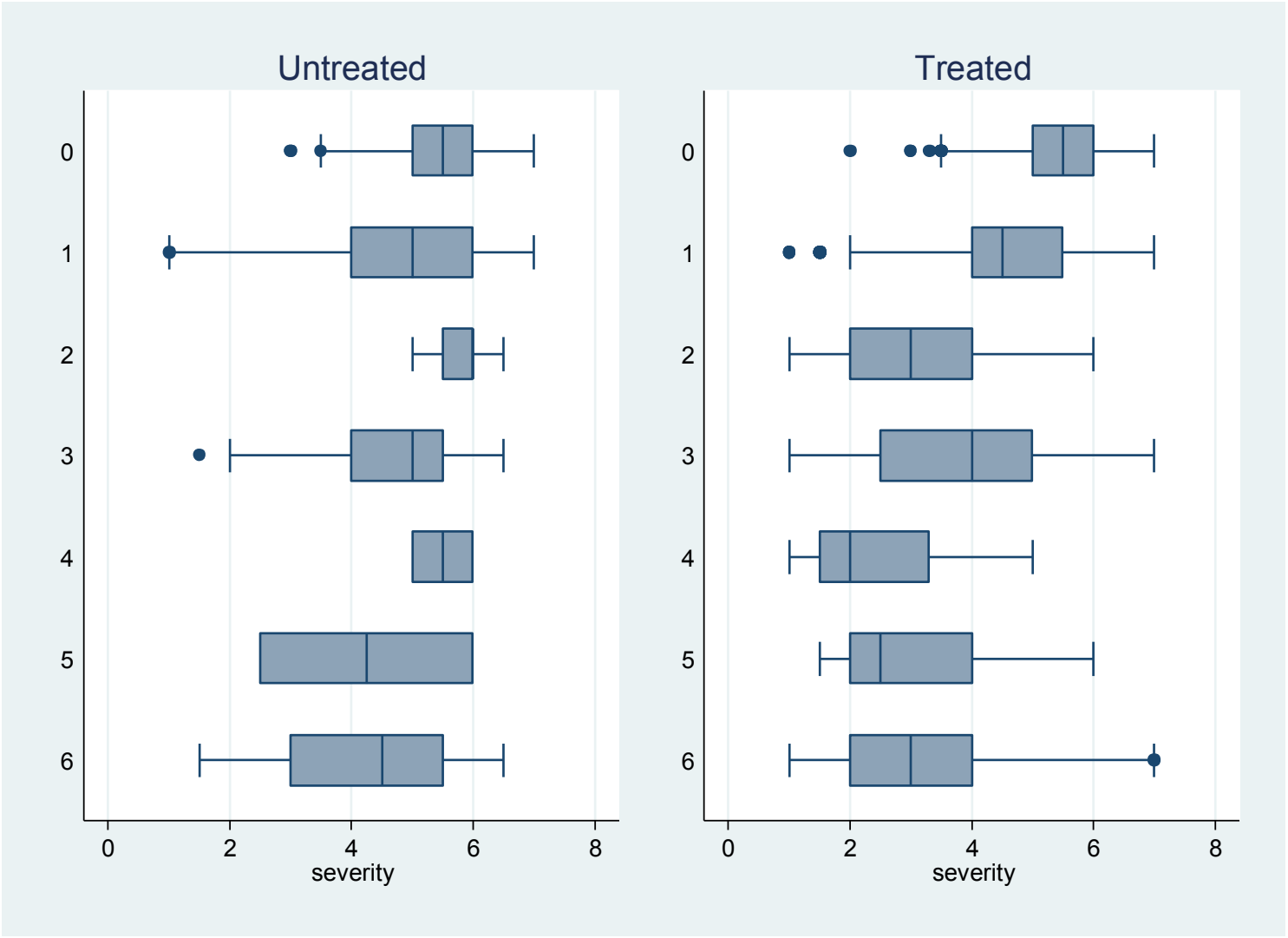
Observations from this plot

- There are fewer untreated than treated people (108 vs. 329)
- Untreated people seem to have a random trajectory – in fact, it seems almost horizontal.
- There might be a downward trend over time in the treated people – it's hard to tell
- This isn't so nice to look at…

Boxplots of severity by week, Treated vs. Untreated

```
graph hbox severity if drug==0, over(week) title(Untreated)
graph save "untreated_box.gph", replace

graph hbox severity if drug==1, over(week) title(Treated)
graph save "treated_box.gph", replace

gr combine "untreated_box.gph" "treated_box.gph"
graph save "treated_untreated_box.gph", replace
```

Boxplot



Plot regression lines by id

Another way to look for trends is to smooth the data

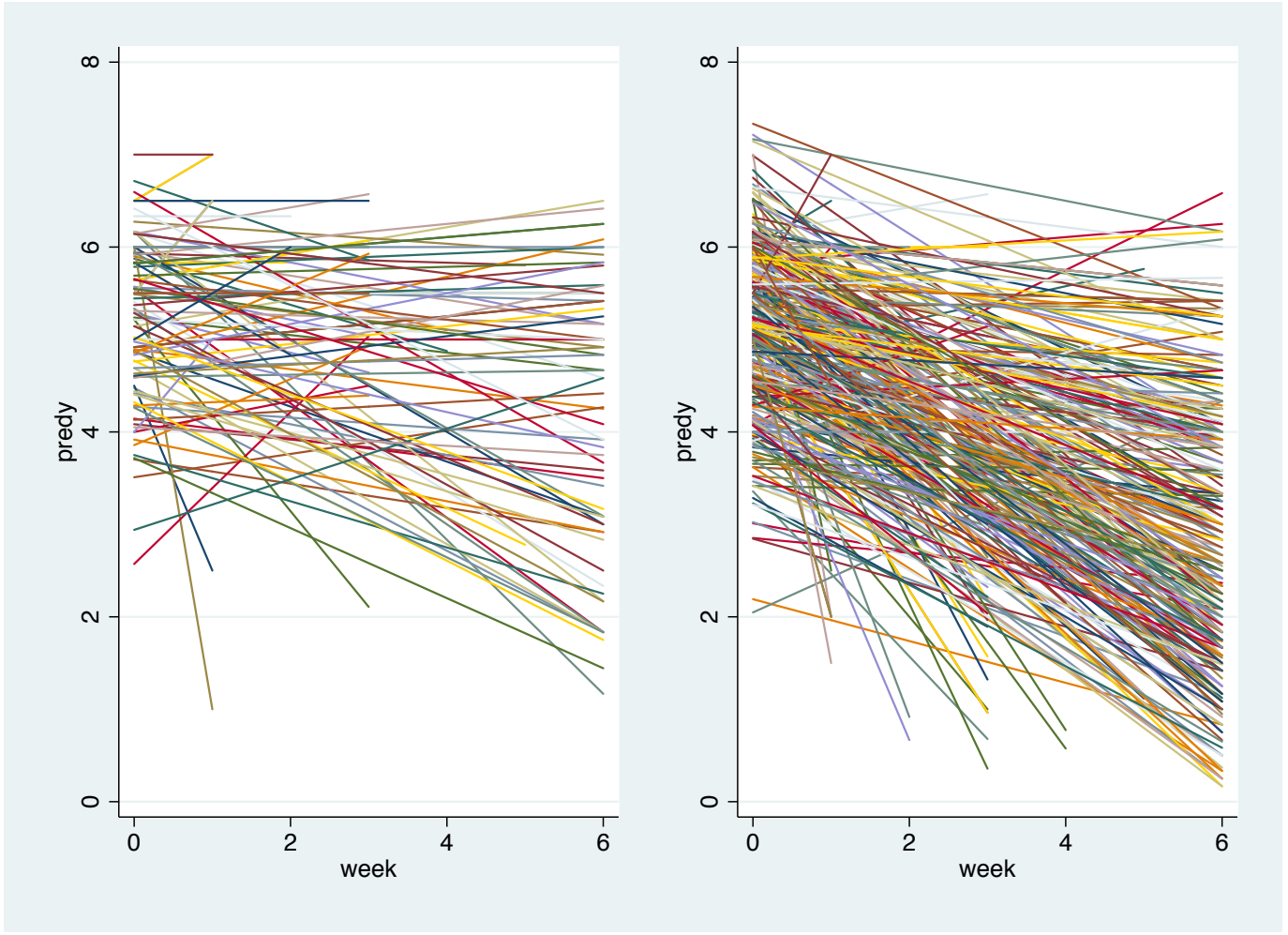
On the Bcourses site, there's a dofile to generate regression lines by id and capture the slope and intercept, with some tips for plotting

Plot regression lines by id

Another way to look for trends is to smooth the data

On the Bcourses site, there's a dofile to generate regression lines by id and capture the slope and intercept, with some tips for plotting

Plot regression lines by id



Test the effectiveness of the drug on the severity of the episode

- First, we need to define what “effectiveness” is.
- One way could be to look at the difference between the severity of the attack in the 0^{th} week vs. the 6^{th} week.
- How do we best achieve that? Look at the data in “wide” format

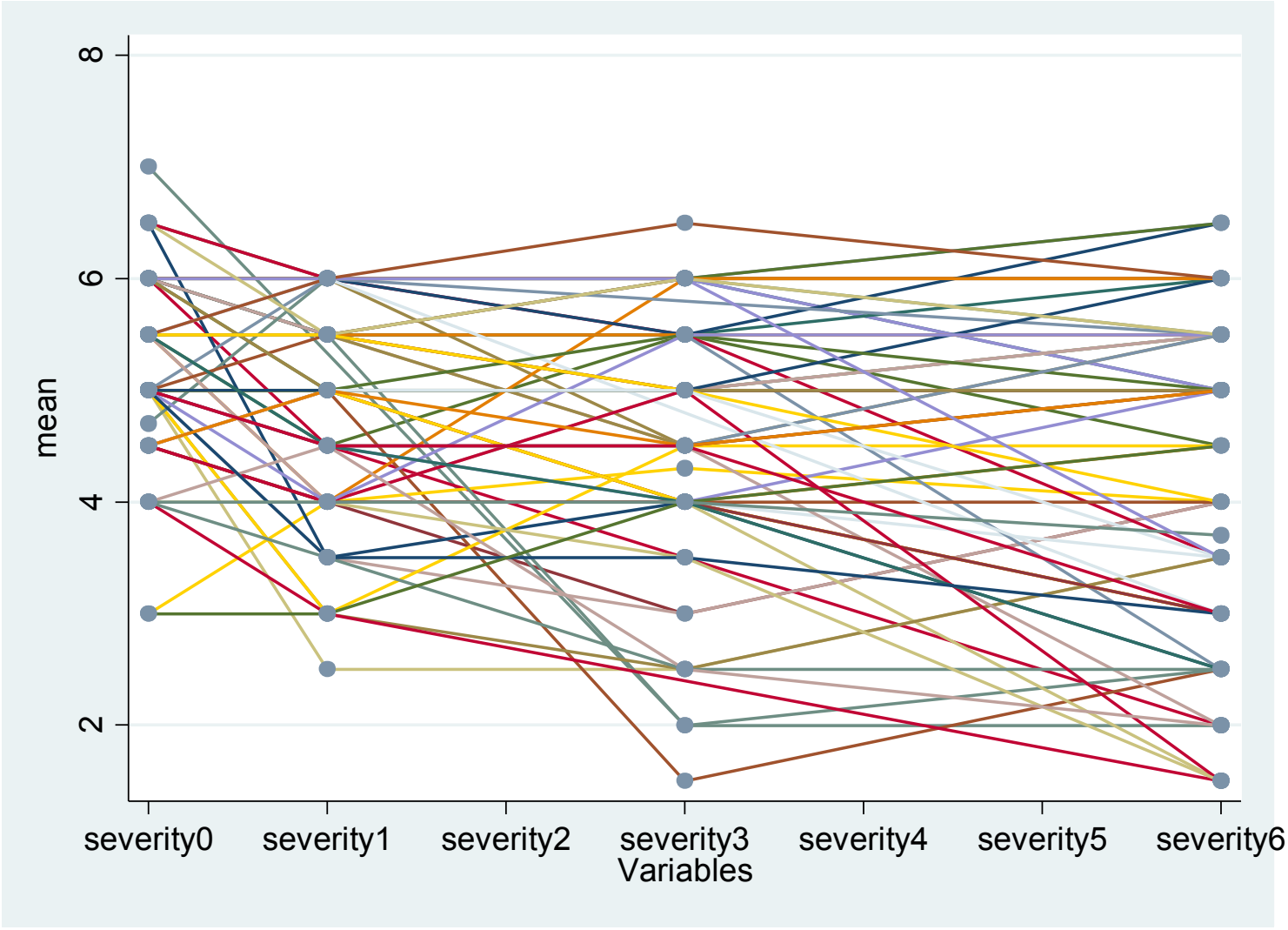
Code to go from long to wide format

```
drop nvals

//reshape to wide by id, creating multiple week variables
reshape wide severity drug gender, i(id) j(week)

// this keeps individuals for whole we have relevant observations
keep if severity0 != . & severity6 != .
```

Plot of the new data

- Let's look at our data again. Previously, we made a plot of data in the long format. We can use a Stata user-written function called "profileplot" to get a general idea of what the remaining data looks like in "wide" format.
- To install "profileplot," run findit profileplot in the command window and follow the instructions
- General because profile plot groups similar observations…pros and cons?

profileplot severity0 severity1 severity2 severity3 severity4 severity5 severity6 if drug0==0, by(id) legend(off)



Statistical Analysis

Now that we have visualized our data, we would like some statistical way to answer the following question:

Is there a difference in severity of episodes between the two treatment groups (being on the drug vs. off of it)? I.e. does the drug make a difference?

Two Sample T-test reminder

- Why do you do a 2 sample t-test?
  - Compare responses from two groups (Are women and men different heights? Did a cholesterol reduction drug actually work?)

- Necessary assumptions:
  - Each group is considered to be a sample from two separate populations (pop 1: treated with drug, pop 2: not treated with drug)
  - Responses are independent
  - Distribution of outcome variable is approximately normal (use a histogram to see this)

Two Sample T-test reminder

- H_0: $\mu_1 = \mu_2$ or equivalently $\mu_1 - \mu_2 = 0$
  (The two population means are equal)

- Alternate hypothesis can be one-sided or two-sided (≥, ≤, or ≠)

- Test statistic:

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2 / n_1 + s_2^2 / n_2}}$$

- Compare to a t-distribution with the smaller of n1-1 or n2-1 degrees of freedom (or if n1=n2, use n1+n2-2 degrees of freedom) to get a p-value

Run a t-test

```
keep severity0 drug0 gender0 severity6

gen difference = severity6 - severity0

ttest difference, by(drug0)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
       0 |      69   -.9927536    .1728765     1.43602   -1.337723    -.647784
       1 |     263   -2.326236    .0941662     1.52712   -2.511655   -2.140817
---------+--------------------------------------------------------------------
combined |     332   -2.049096    .0878668     1.60101   -2.221944   -1.876249
---------+--------------------------------------------------------------------
    diff |            1.333482    .2040787                .9320229    1.734941
------------------------------------------------------------------------------
    diff = mean(0) - mean(1)                                      t =   6.5342
Ho: diff = 0                     degrees of freedom =      330

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Other options

- You have slopes for each of your treated and untreated individuals. What about a t-test to test for a difference in the mean values of the slopes?

- You could also approach this question using the regression models we are covering in lecture. We will discuss this next week, but feel free to start now.

Your turn

We've looked at the data by drug treatment. Now, can you perform a similar analysis looking to see if there's a gender effect? Some ideas

- Create a scatterplot and a boxplot demonstrating the change over time of severity in each individual, conditioning on gender.
- Put the data in wide format and create a profileplot of each of the genders.
- Examine the difference in severity of attacks between weeks 0 and 6 between genders. Do a t-test . What do your results suggest?

Modeling

We've looked at graphics and summary measures, and thought about how to test for differences using summary measures.

- Why would you want to model?

Modeling

- We can include other variables
- Examine different assumptions and related variances
- Eventually, different mean models
- What models might you try with this data?

What models might you try on this data?

- Cross sectional model regressing strength on time and treatment
- Cross sectional model with robust standard errors
- Including gender
- Others?

Are these standard errors reasonable?

```
.
. regress severity week drug
```

Source	SS	df	MS
Model	951.170459	2	475.58523
Residual	2517.62532	1600	1.57351583
Total	3468.79578	1602	2.16529075

Number of obs	1603
F(2, 1600)	302.24
Prob > F	0.0000
R-squared	0.2742
Adj R-squared	0.2733
Root MSE	1.2544

severity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
week	-.3207993	.0141362	-22.69	0.000	-.3485266	-.2930719
drug	-.6397973	.0738503	-8.66	0.000	-.7846508	-.4949438
_cons	5.597435	.0713402	78.46	0.000	5.457505	5.737365

Is this reasonable?

```
.
. regress severity week drug
```

Source	SS	df	MS
Model	951.170459	2	475.58523
Residual	2517.62532	1600	1.57351583
Total	3468.79578	1602	2.16529075

Number of obs	1603
F(2, 1600)	302.24
Prob > F	0.0000
R-squared	0.2742
Adj R-squared	0.2733
Root MSE	1.2544

severity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
week	-.3207993	.0141362	-22.69	0.000	-.3485266	-.2930719
drug	-.6397973	.0738503	-8.66	0.000	-.7846508	-.4949438
_cons	5.597435	.0713402	78.46	0.000	5.457505	5.737365

Is this reasonable?

```
. regress severity week drug, robust

Linear regression                               Number of obs =    1603
                                                F(  2,  1600) =  301.24
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2742
                                                Root MSE      =  1.2544
```

severity	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
week	-.3207993	.0145419	-22.06	0.000	-.3493223	-.2922762
drug	-.6397973	.070897	-9.02	0.000	-.7788581	-.5007366
_cons	5.597435	.0635448	88.09	0.000	5.472796	5.722075

Adding gender

```
. regress severity week drug gender
```

Source	SS	df	MS
Model	952.266632	3	317.422211
Residual	2516.52915	1599	1.57381435
Total	3468.79578	1602	2.16529075

Number of obs	1603
F(3, 1599)	201.69
Prob > F	0.0000
R-squared	0.2745
Adj R-squared	0.2732
Root MSE	1.2545

severity	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
week	-.3209089	.0141381	-22.70	0.000	-.3486401	-.2931777
drug	-.6412884	.0738789	-8.68	0.000	-.7861981	-.4963787
gender	.0523612	.0627403	0.83	0.404	-.0707007	.1754231
_cons	5.573707	.0768032	72.57	0.000	5.423062	5.724353

Adding gender

```
. regress severity week drug gender, robust

Linear regression                               Number of obs =    1603
                                                F(  3,  1599) =  201.25
                                                Prob > F      =  0.0000
                                                R-squared     =  0.2745
                                                Root MSE      =  1.2545
```

severity	Coef.	Robust Std. Err.	t	P>|t|	[95% Conf.	Interval]
week	-.3209089	.0145409	-22.07	0.000	-.3494301	-.2923877
drug	-.6412884	.0710023	-9.03	0.000	-.7805559	-.502021
gender	.0523612	.0626472	0.84	0.403	-.070518	.1752404
_cons	5.573707	.0702801	79.31	0.000	5.435857	5.711558

Any other ideas?