Stata: An Introduction

BUILD Online Research Module

CSULB

Summer 2020

Acknowledgements

- Dr. Alan Acock, Author of “A Gentle Introduction to Stata”
- CSULB BUILD Program (NIH Award #RL5GM118978)

BUILD
CSULB

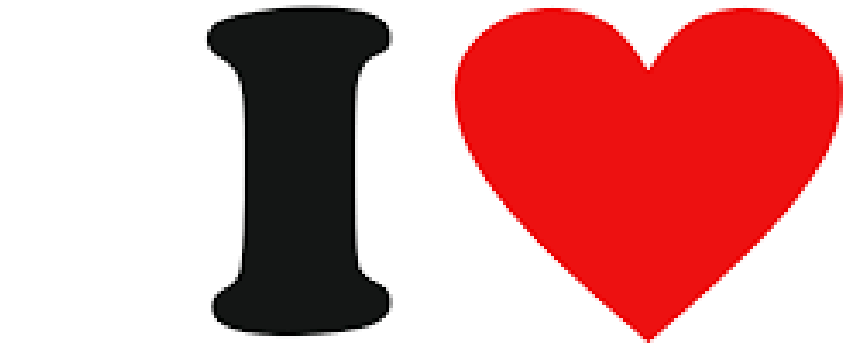
Agenda

- I. Background
- II. Getting Started
- III. Entering Data
- IV. Preparing Data for Analysis
- V. Working with commands, do-files and results
- VI. Descriptive Statistics and Graphs – Single Variables
- VII. Statistics and Graphs – Two Categorical Variables
- VIII. Tests for one or two means
- IX. Bivariate Correlation and Regression
- X. Multiple Regression
- XI. Logistic Regression
- XII. A Public Health Example
- XIII. What's Next

I. Background

- Public Health Background
  - Statistics training via coursework and real-world research projects
  - Quantitative Software experience:
    - Excel
    - MPlus
    - SPSS
    - **Stata**

I. Background



- Stata fan
  - Why I like Stata:
    - Easy to use
    - Fast
    - Saves Time (Do-Files)
    - Helps you when you make a mistake
    - Efficient
    - Capable of complex analyses

I. Background

- Today's Workshop
  - Basic understanding of statistics
  - Created for the Stata novice
  - Covering the basics of Stata

II. Getting Started

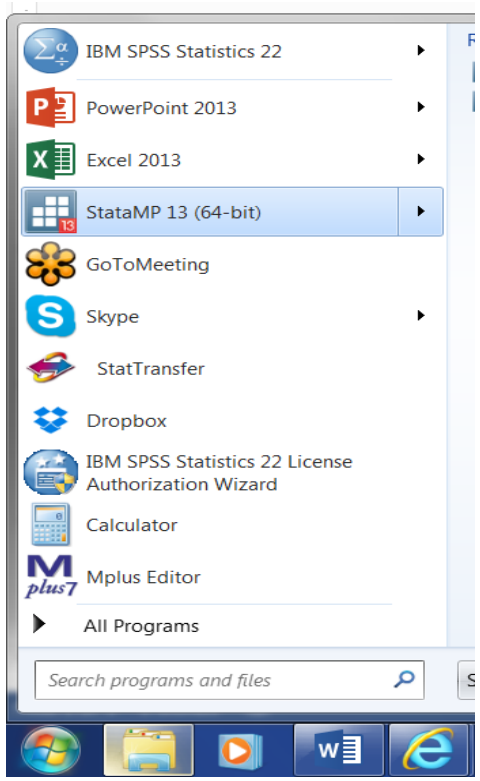
- Before we begin
  - Short Term Trial for Students
    - https://www.stata.com/customer-service/short-term-license/
    - Student Plans
      - https://www.stata.com/order/new/edu/gradplans/student-pricing/
  - Module material
    - Do File
    - Sample data set and codebook (upon formal request)
  - Personal Recommendation
    - “Pause” often to follow along as we go through different examples

II. Getting Started

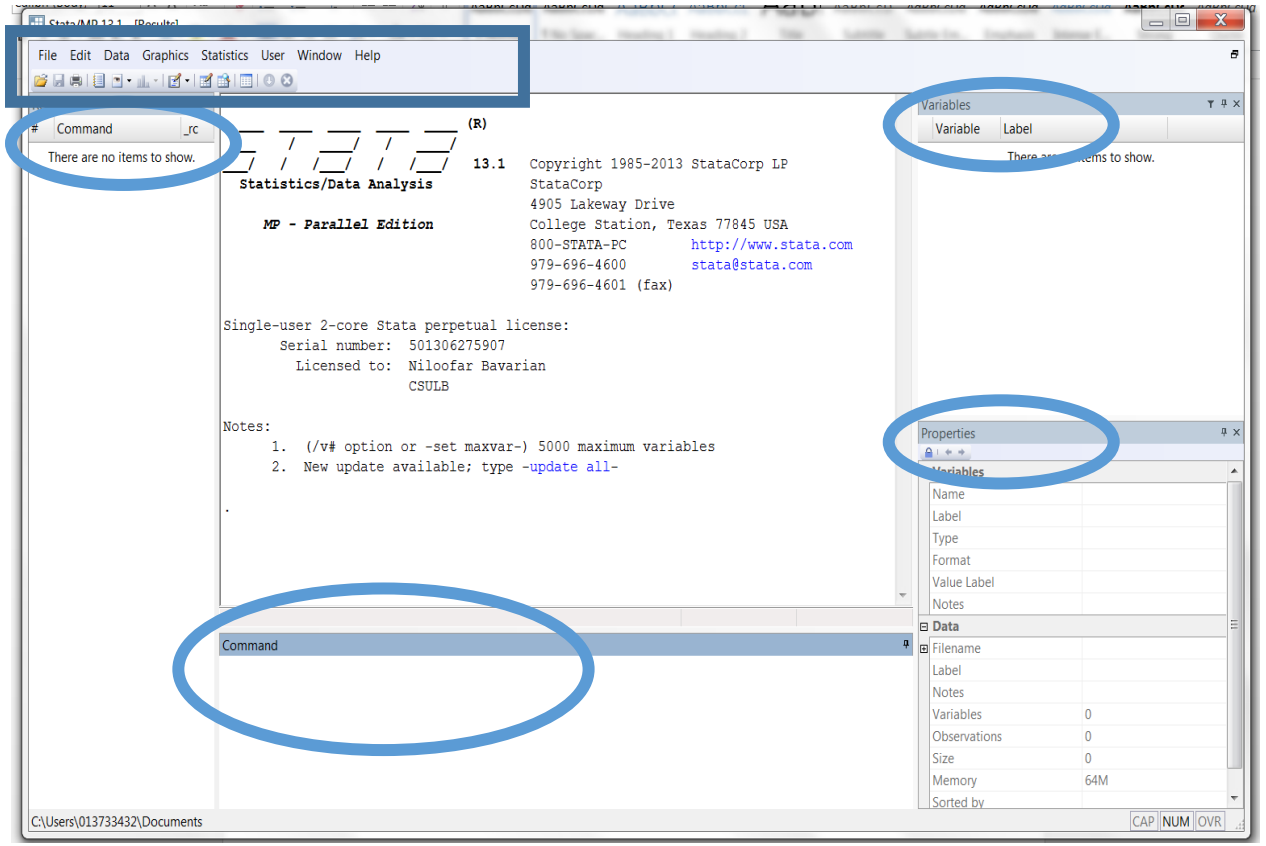
- OBJECTIVES
  - Understand the Stata interface
  - Open an existing dataset
  - Complete a short Stata session

II. Getting Started

- Open Stata

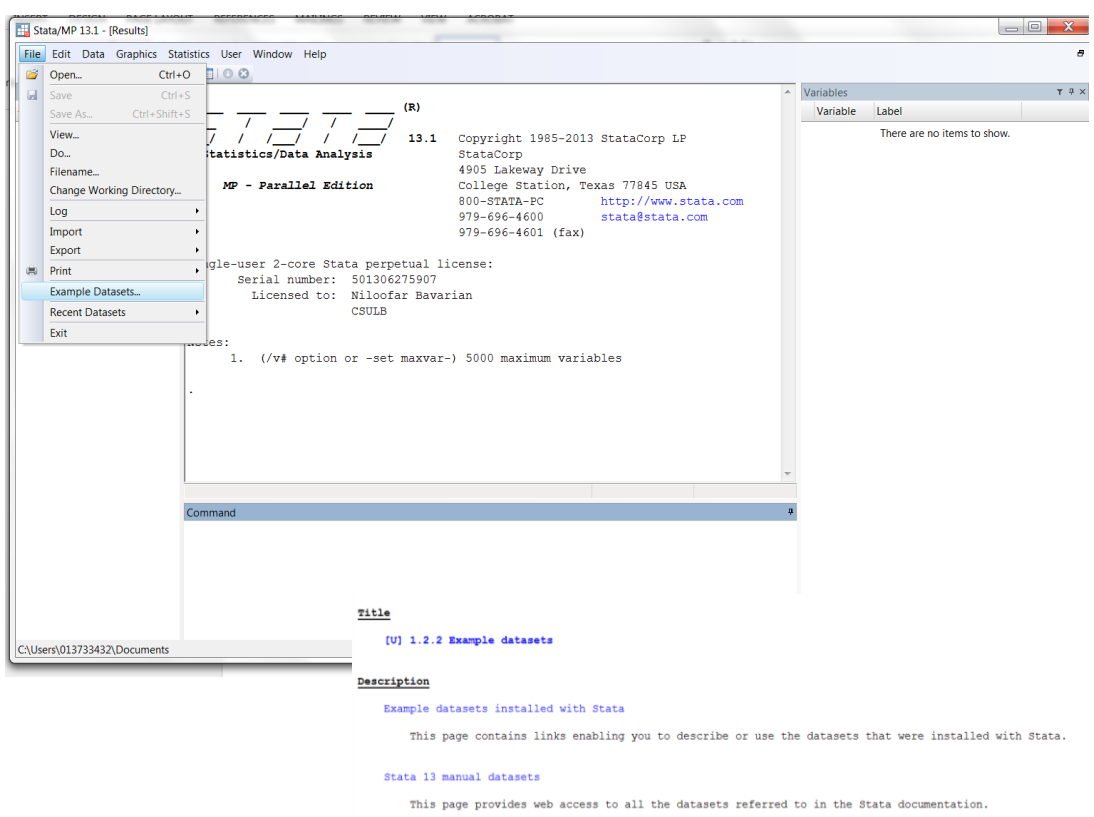


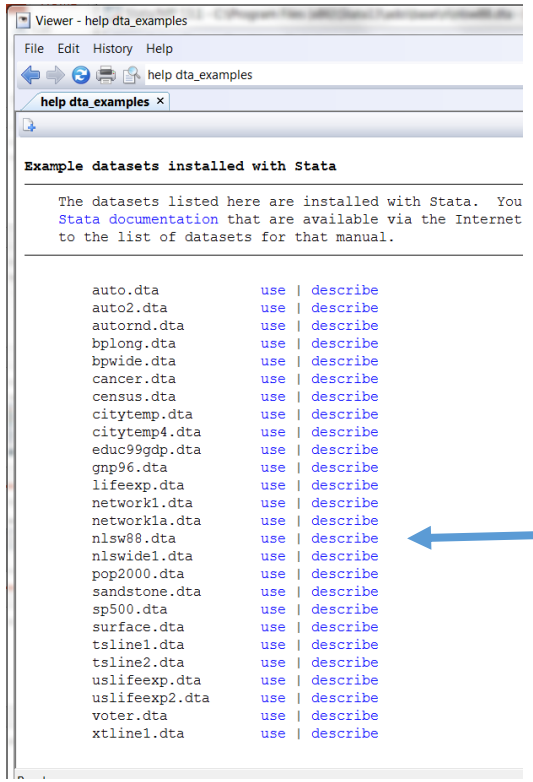
- The Stata Interface



II. Getting Started

Open an Existing Dataset

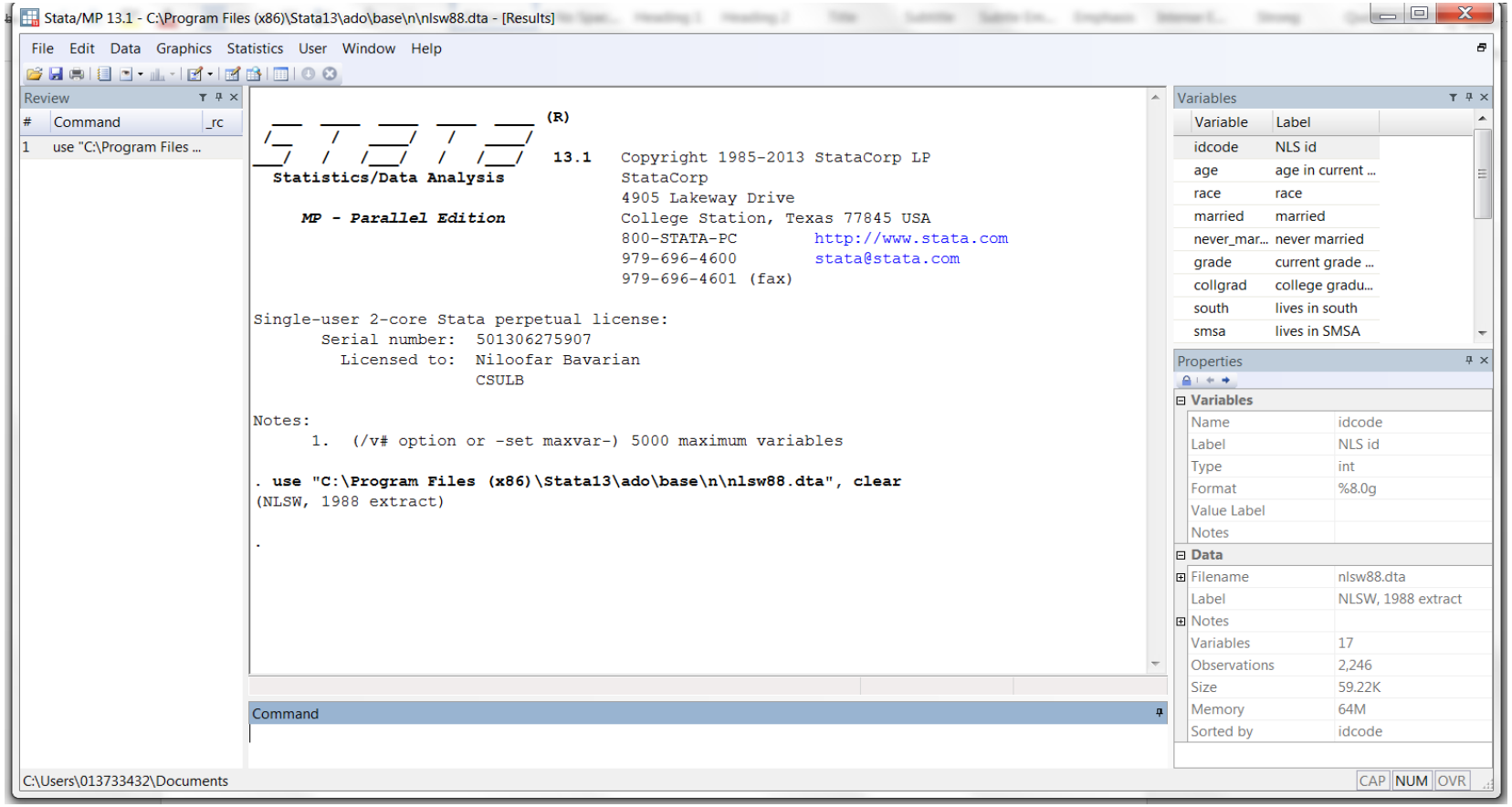




II. Getting Started

Open an existing dataset

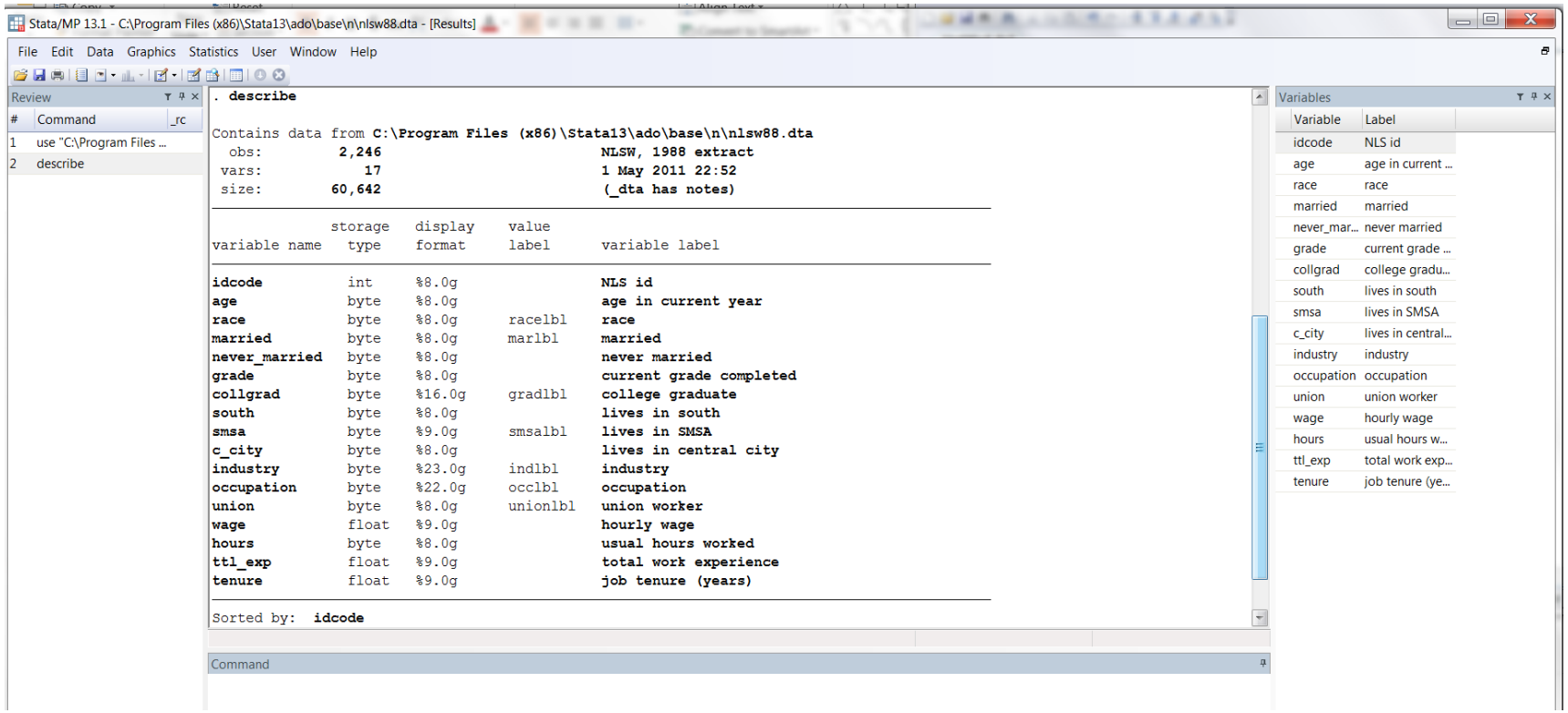
National Longitudinal Study of Women, NLSW, 1988



II. Getting Started

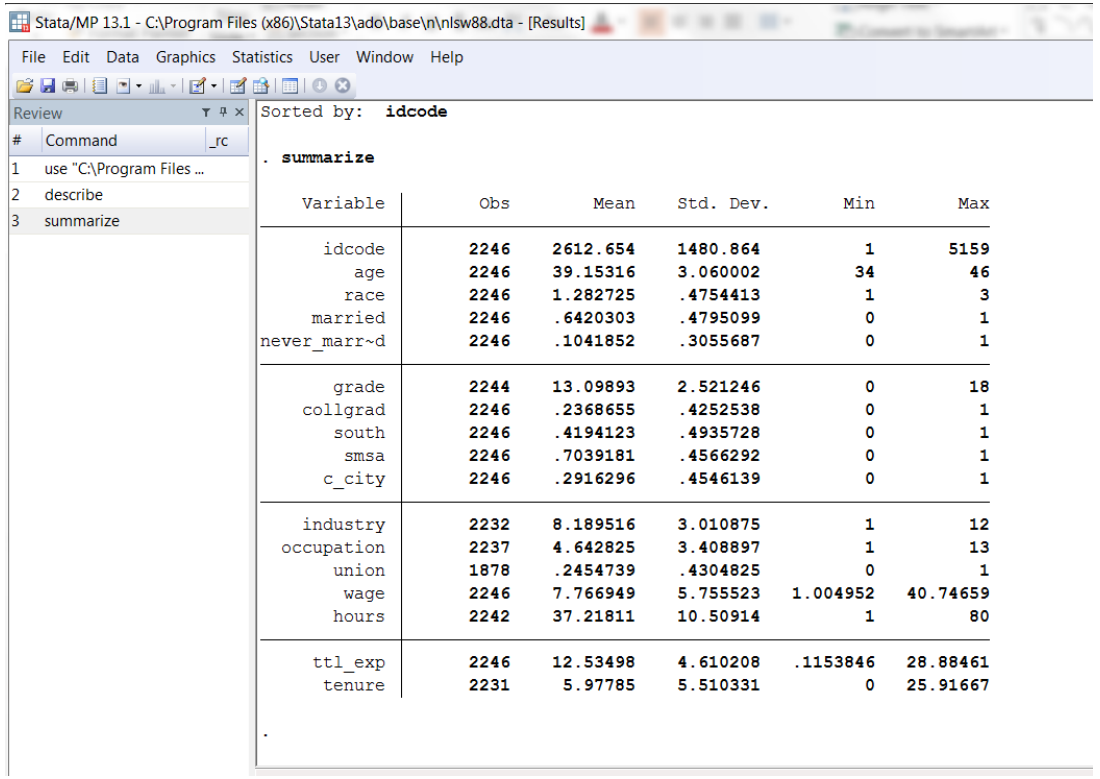
Open an existing dataset

describe



II. Getting Started

- Complete a short Stata session
  - `summarize`



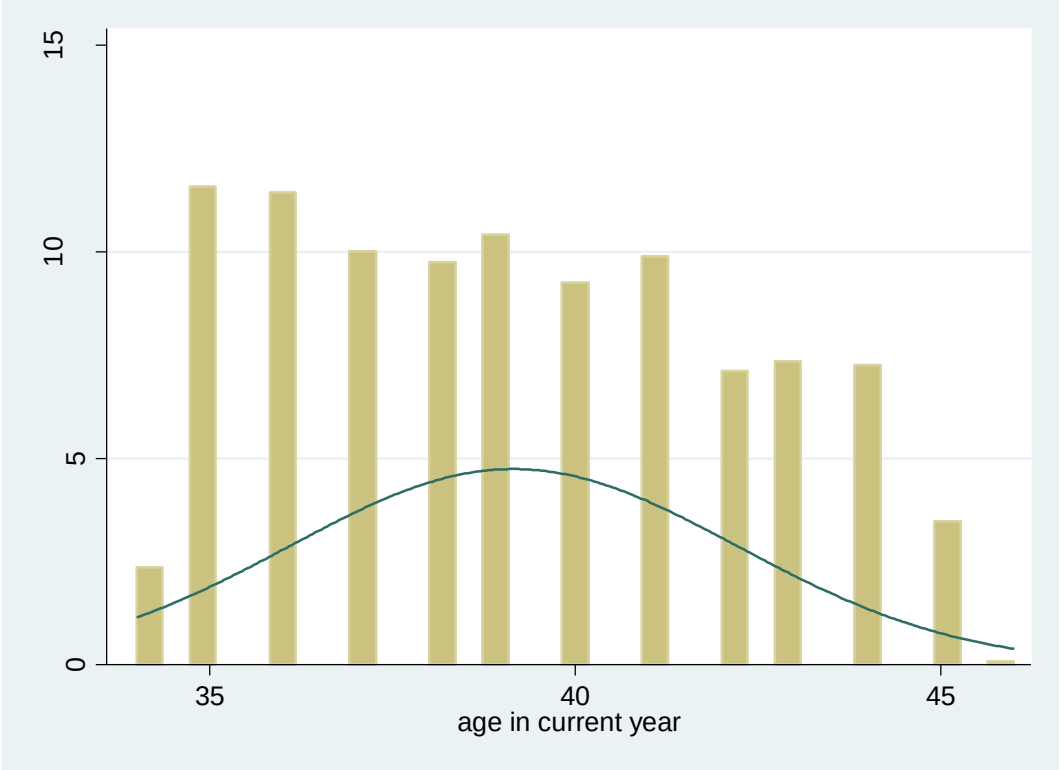
II. Getting Started

- Complete a short Stata session
  - `summarize age`
  - `summarize age if south == 1`

```
. summarize age

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |      2246    39.15316    3.060002        34         46

. summarize age if south == 1

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
         age |       942    39.17834    3.118291        34         45
```

II. Getting Started

- Complete a short Stata session
  - `histogram age, percent normal`



III. Entering Data

- Objectives
  - Understand the importance of a codebook
  - Create variables in Stata
  - List resources for Data Management in Stata

III. Entering Data

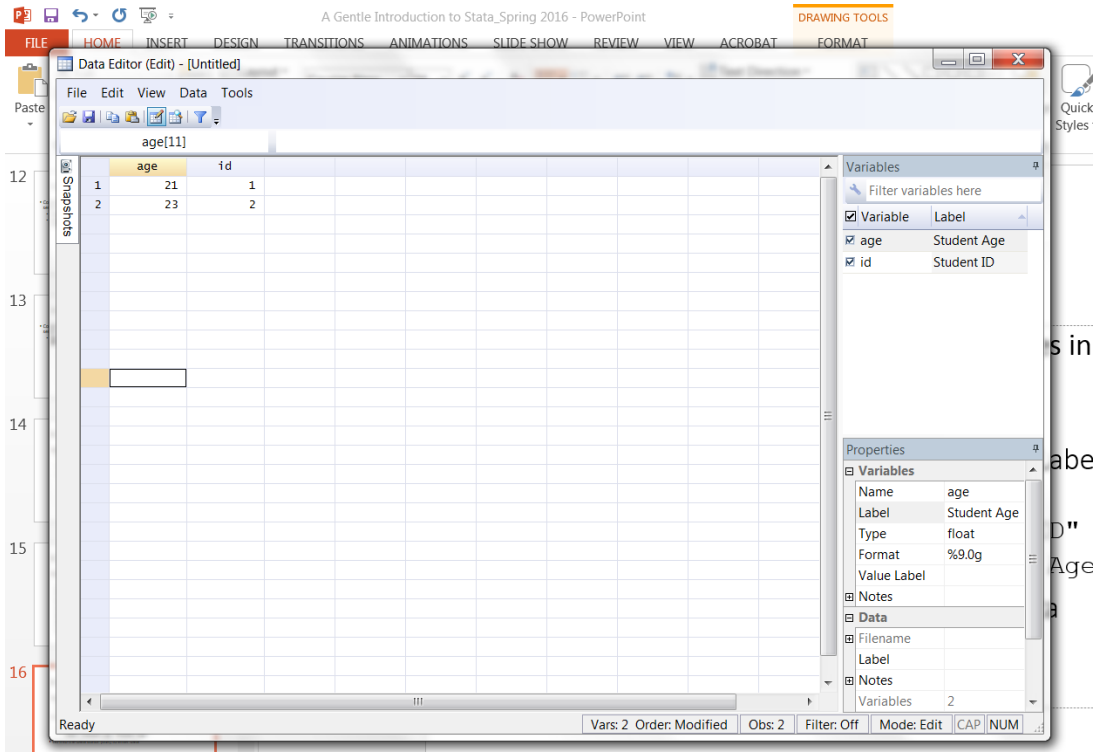
- First start with a codebook

BEACH Q-V2 Codebook

#	Variable Name	Question	Response Options/How Scored				
	id	Student ID Number	Continuous				
	classid	Classroom ID Number	Continuous				
	deptid	Course Department	Descriptive				
			Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree
1	_inatt1_t1	It is difficult for me to pay attention during classes	1	2	3	4	5
2	_hyper1_t1	I often feel restless	1	2	3	4	5
3	_inatt2_t1	It is difficult for me to concentrate on my academic work	1	2	3	4	5
4	_hyper2_t1	I am an impulsive person	1	2	3	4	5
5	_inatt3_t1	I have difficulty keeping track of my different school assignments	1	2	3	4	5
6	_hyper3_t1	I rarely plan ahead	1	2	3	4	5
7	_ss1_t1	I like "wild" parties	1	2	3	4	5
8	_ss2_t1	I enjoy getting into situations where I do not know how things will turn out	1	2	3	4	5
9	_ss3_t1	I prefer friends who are unpredictable	1	2	3	4	5

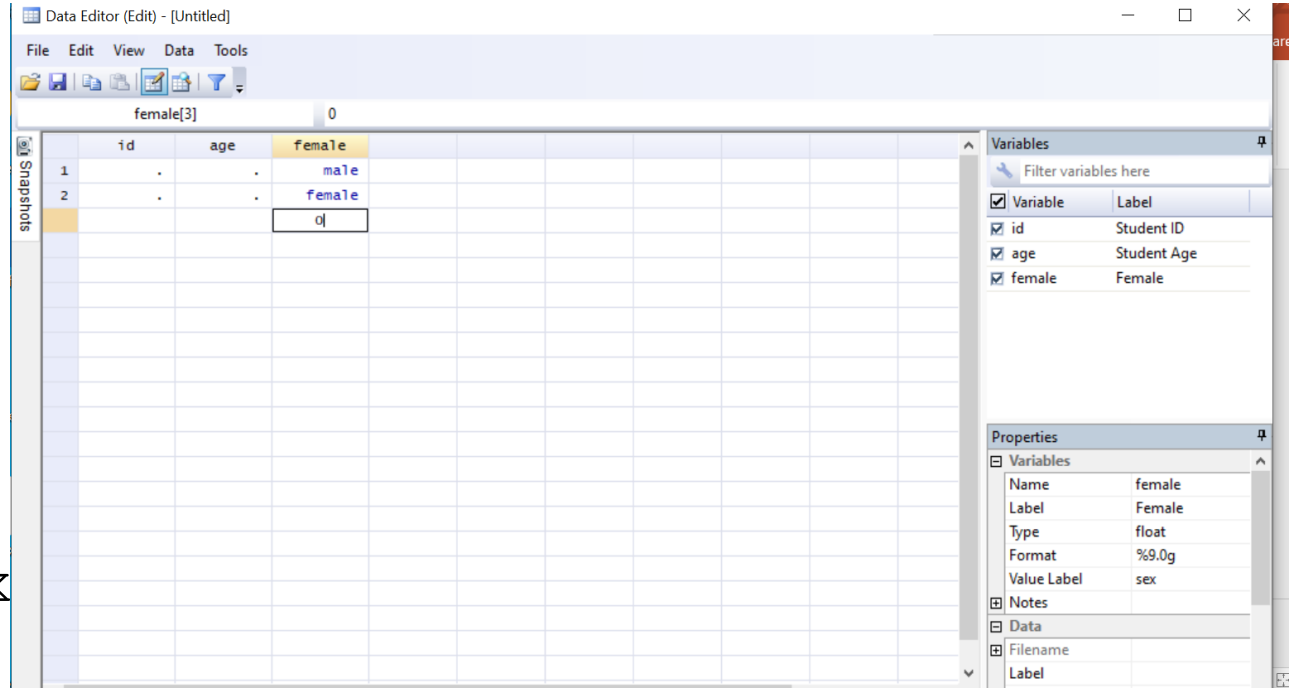
III. Entering Data

- Use the Command Box to Create Variables in Stata
  - `gen id=.`
  - `gen age = .`
- Use the Command Box to Enter Variable Labels
  - `. label variable id "Student ID"`
  - `. label variable age "Student Age"`
- Go into the Data Editor (Edit) to enter data



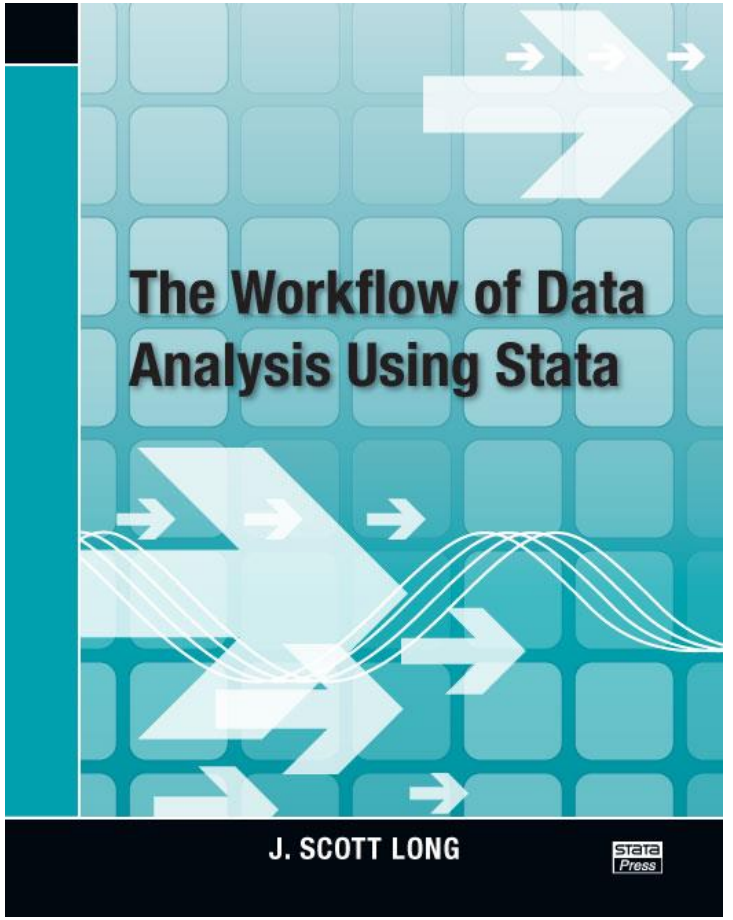
III. Entering Data

- Use the Command Box to Create Variables in Stata (Categorical)
  - `gen female=.`
  - `label variable female "Female"`
- Create labels for response categories
  - `label define sex 0 "male" 1 "female"`
  - `label values female sex`
  - Try entering 0 and 1 for in your Data Editor



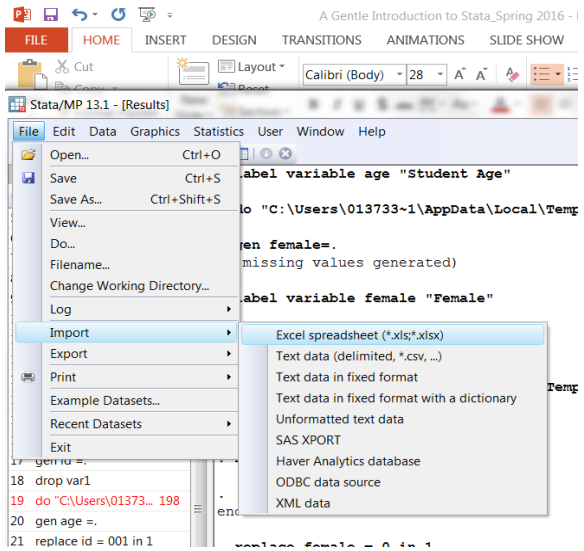
III. Entering Data

- Resources for Data Management

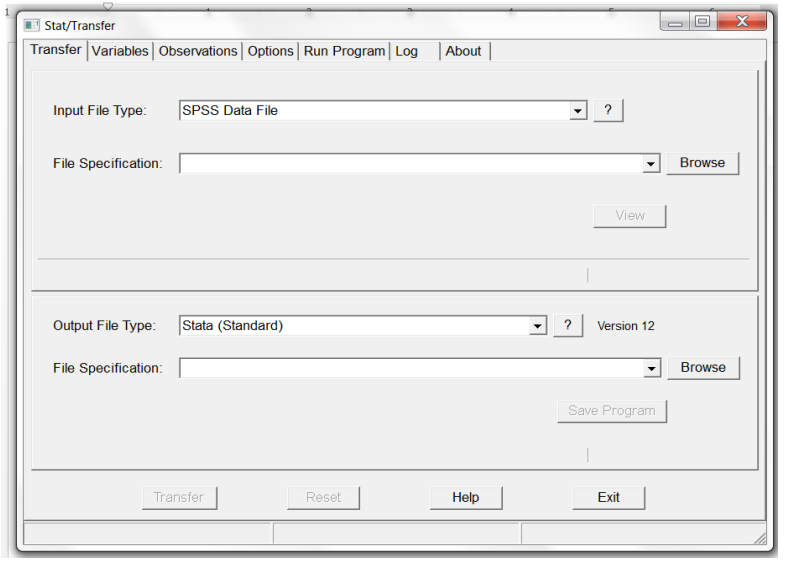


III. Entering Data

- Some of us may
  - 1) Prefer to enter data in Excel
  - 2) Work with secondary data that came in a different type of data file (e.g., SPSS)



- If this is the case
- 1)Stata has an IMPORT function
- 2) Stat Transfer is another option



IV Preparing Data for Analysis

- Objectives
  - Understand tools to prepare for data analysis such as
    - Checking for entry error
    - Recoding missing values
    - Reverse coding variables
    - Recoding variables
    - Checking alpha
    - Creating composite variables

IV Preparing Data for Analysis

Checking for entry Error

Practice with opening NLSW data set

- `Use "fre" command to scan all variables for entry error`
  - `fre age-tenure`

Recoding Missing Values

- Pretend the grade variable had a few responses with a -9 response
- Use `mvdecode`
- `mvdecode grade, mv(-9=.)`

IV Preparing Data for Analysis

Reverse Coding Variables

- Imagine you have a variable called "`safety`" with response options
  - 1=Very safe – 5 = Not safe at all
- You may choose to reverse code so that a higher score on safety reflects more feelings of safety
  - **`revrs safety`**
  - Will create a new variable called `revsafety`
  - The `revsafety` variable will now be reverse coded

Recoding variables

- Imagine you have a variable "`packs`" – number of packs smoked a day
  - 1 pack – 50
  - 2 packs – 20
  - 3 packs – 10
  - 4 packs – 2
- You may want to combine category 3 and 4; to do so
  - `recode packs (1=1 "1 pack")(2=2 "2 packs") (3/4=3 "3+ packs"), generate (packs_re)`

IV Preparing Data for Analysis

Checking alpha

- Imagine you have five items you think measure an underlying concept of happiness
  - `happy1 happy2 happy3 happy4 happy5`
- You want to check their internal consistency before you make a composite variable
  - `alpha happy1-happy5, item asis`

Creating composite variables

- Imagine those five items have a strong alpha, to create on composite variable
  - `egen happiness = rowmean (happy1-happy5)`

V. Working with commands, do-files and results

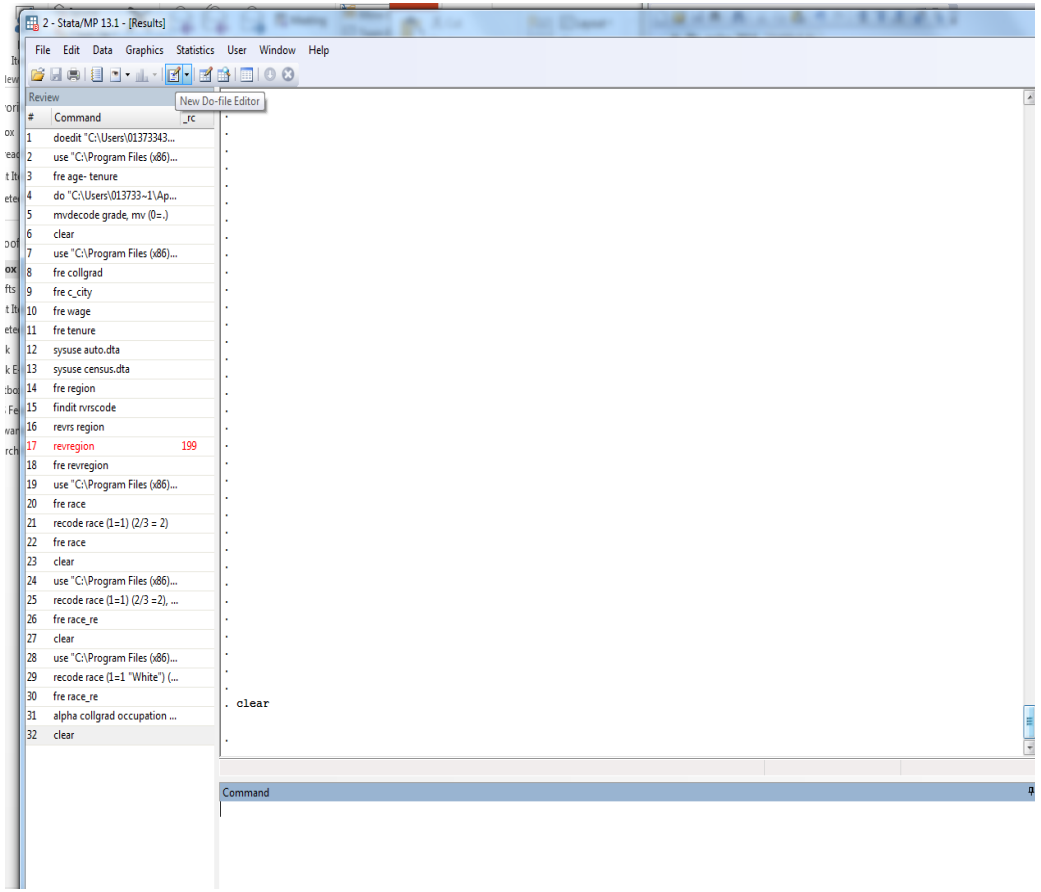
- Objectives
  - Understand why the Do-File is your new best friend

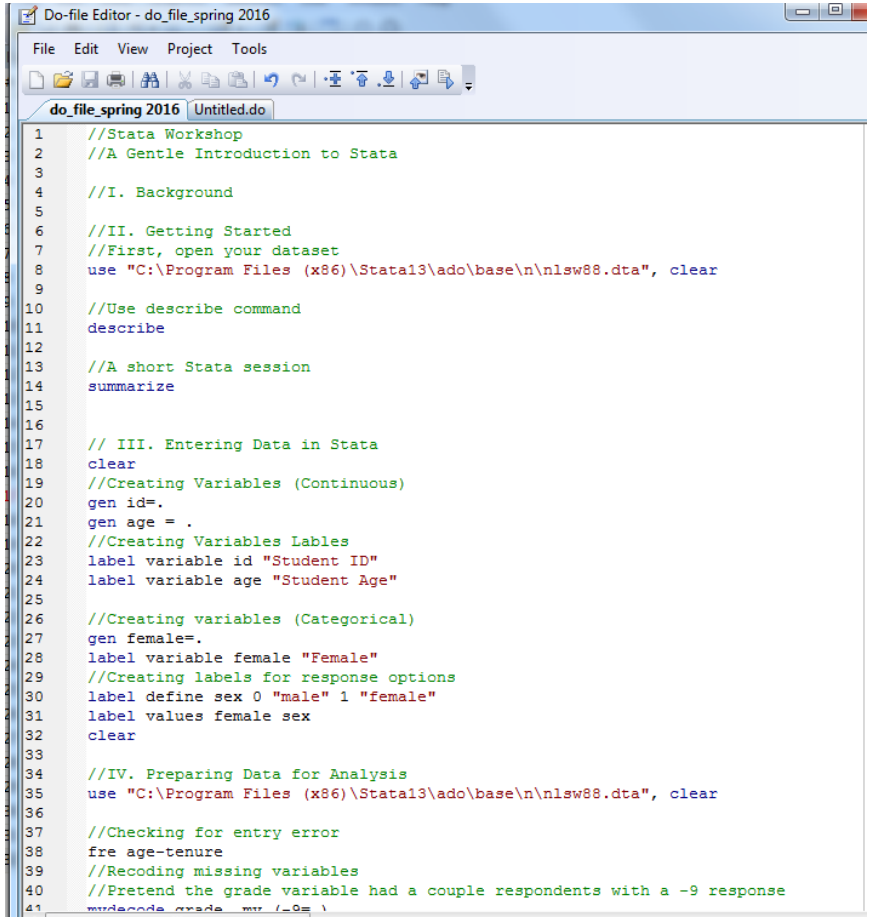
V. Working with commands, do-files and results

- What is a Do-File?
  - A text file in Stata where you can save (and run!) all your commands and notes
    - As opposed to typing everything in the command box
  - Allows you to go back to a dataset and quickly repeat your analysis
    - Why is this important?

V. Working with commands, do-files and results

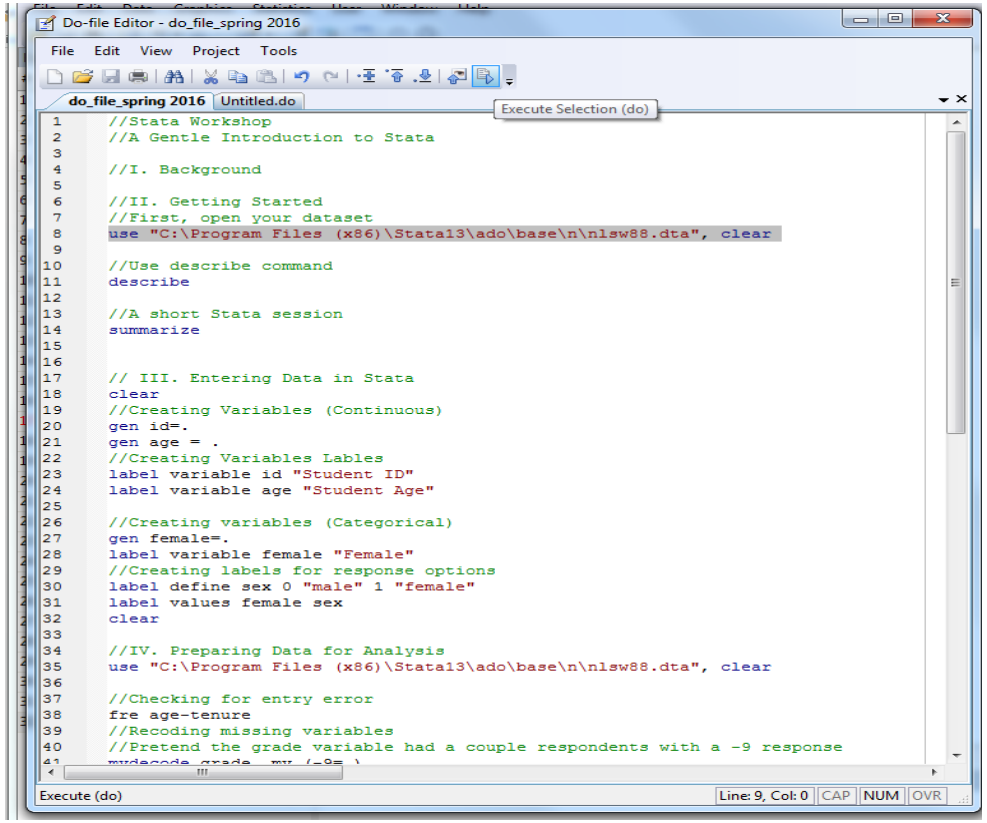
- Opening your do-file

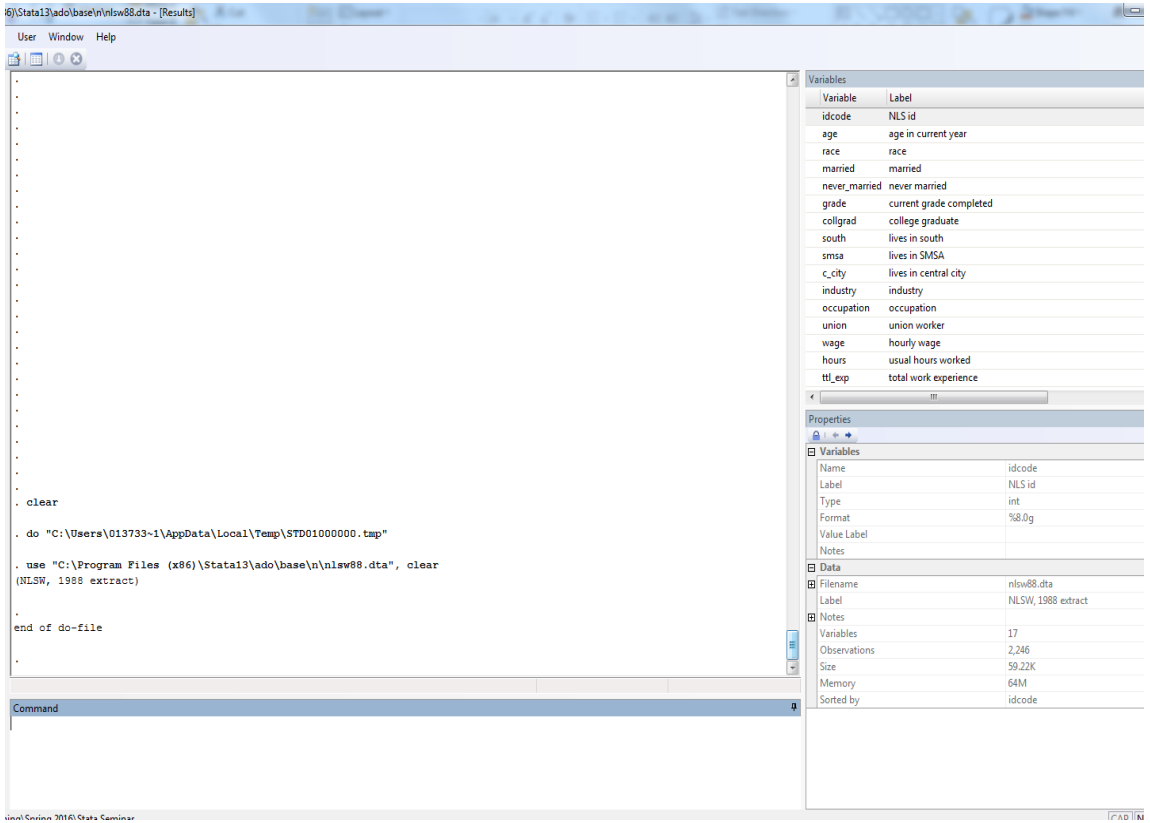




V. Working with commands, do-files and results

- Running the Do-File
  - Highlight and Execute





VI. Descriptive Statistics and Graphs – Single Variables

- Objectives
  - Calculate descriptive statistics
  - Construct graphical depictions of data

VI. Descriptive Statistics and Graphs – Single Variables

Descriptive Statistics

- What are examples of descriptive statistics?

Using Stata

- `use "C:\Program Files (x86)\Stata13\ado\base\n\nlsw88.dta", clear`
- `summarize wage, detail`

VI. Descriptive Statistics and Graphs – Single Variables

Descriptive Statistics - Output

- `. summarize wage, detail`

```
                        hourly wage
-------------------------------------------------------------
      Percentiles      Smallest
 1%     1.930993       1.004952
 5%     2.801002       1.032247
10%     3.220612       1.151368       Obs                2246
25%     4.259257       1.344605       Sum of Wgt.        2246

50%      6.27227                      Mean           7.766949
                        Largest       Std. Dev.      5.755523
75%     9.597424       40.19808
90%     12.77777       40.19808       Variance       33.12604
95%     16.52979       40.19808       Skewness       3.096199
99%     38.70926       40.74659       Kurtosis       15.85446
```

VI. Descriptive Statistics and Graphs – Single Variables

Descriptive Statistics - Output

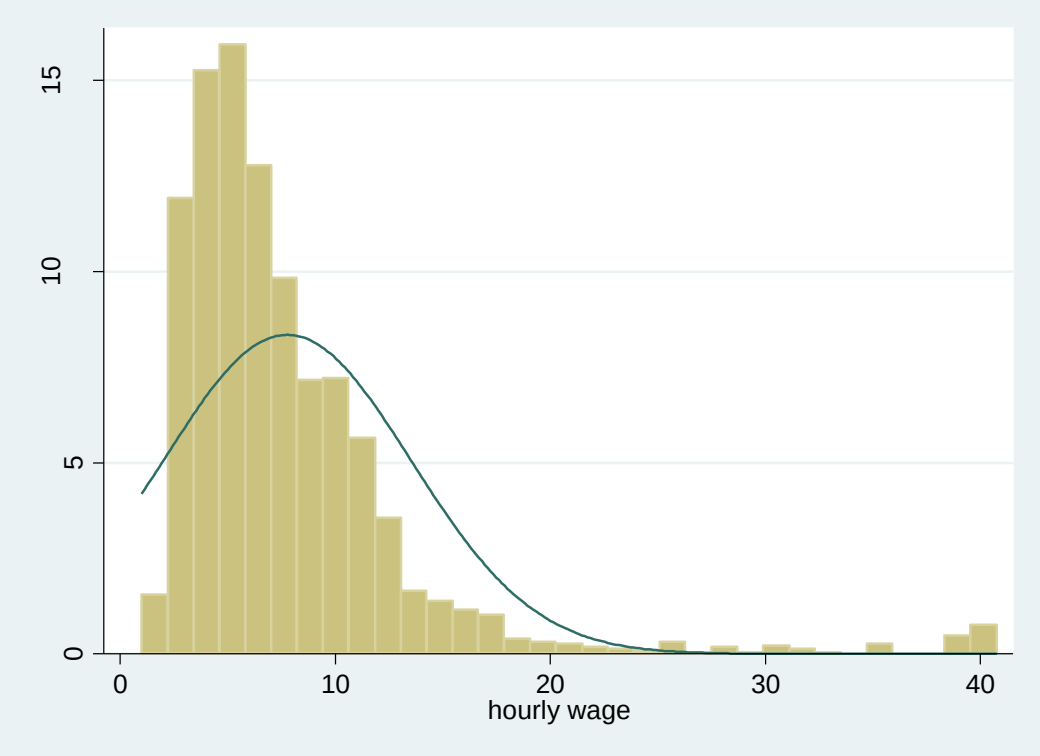
```
. sktest wage

                    Skewness/Kurtosis tests for Normality
                                                       ------- joint ------
    Variable |    Obs    Pr(Skewness)    Pr(Kurtosis)  adj chi2(2)    Prob>chi2
-------------+-------------------------------------------------------------------
        wage |   2.2e+03   0.0000          0.0000             .              .
```

VI. Descriptive Statistics and Graphs – Single Variables

Descriptive Statistics - Graphs

- What kind of variable is "`wage`"?
- What graphic illustration is most appropriate

`histogram wage, percent normal`



VII. Statistics and Graphs – Two Categorical Variables

- Objectives
  - Conduct a cross-tabulation
  - Conduct a two-variable Chi-Square test

VII. Statistics and Graphs – Two Categorical Variables

Categorical Variables

- What are examples of categorical variables/response options?

NLSW Dataset

- Two examples of categorical variables
  - `Married` (Yes/No)
  - `Race` (White/Black/Other)

- Research Question
  - Is there an association between race and marriage?

VII. Statistics and Graphs – Two Categorical Variables

Stata Command

- `tabulate race married, chi2 expected row`

Stata Output

```
. tabulate race married, chi2 expected row

Key

    frequency
expected frequency
  row percentage
```

race	married: single	married: married	Total
white	487	1,150	1,637
	586.0	1,051.0	1,637.0
	29.75	70.25	100.00
black	309	274	583
	208.7	374.3	583.0
	53.00	47.00	100.00
other	8	18	26
	9.3	16.7	26.0
	30.77	69.23	100.00
Total	804	1,442	2,246
	804.0	1,442.0	2,246.0
	35.80	64.20	100.00

```
Pearson chi2(2) = 101.4215   Pr = 0.000
```

VIII. Tests for one or two means

- Objectives
  - Conduct a one-sample test of means
  - Conduct a two-sample test of group means

VIII. Tests for one or two means

One-sample test of means

- Let's pretend the average hourly compensation rate in the U.S. in 1988 was $10
  - You want to know if the average hourly rate `wage` for your sample is equal to national average

- `ttest wage == 10`

Output

```
. ttest wage == 10

One-sample t test
------------------------------------------------------------------------------
Variable |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
    wage |    2246    7.766949    .1214451    5.755523    7.528793    8.005105
------------------------------------------------------------------------------
    mean = mean(wage)                                         t = -18.3873
Ho: mean = 10                                 degrees of freedom =     2245

    Ha: mean < 10               Ha: mean != 10                 Ha: mean > 10
 Pr(T < t) = 0.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 1.0000
```

VIII. Tests for one or two means

Two-sample test of group means

- Let's say we want to know if the mean `wage` is the same for single versus `married` participants

  - `ttest wage, by (married)`

Output

```
. ttest wage, by (married)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
  single |     804    8.080765     .223456    6.336071    7.642138    8.519392
 married |    1442    7.591978    .1421835    5.399229    7.313069    7.870887
---------+--------------------------------------------------------------------
combined |    2246    7.766949    .1214451    5.755523    7.528793    8.005105
---------+--------------------------------------------------------------------
    diff |            .4887873    .2531718               -.0076882    .9852627
------------------------------------------------------------------------------
    diff = mean(single) - mean(married)                           t =   1.9307
Ho: diff = 0                                     degrees of freedom =     2244

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.9732         Pr(|T| > |t|) = 0.0537          Pr(T > t) = 0.0268
```

IX. Bivariate Correlation and Regression

- Objectives
  - Construct a scattergram
  - Calculate correlation
  - Estimate a regression model

IX. Bivariate Correlation and Regression

Example

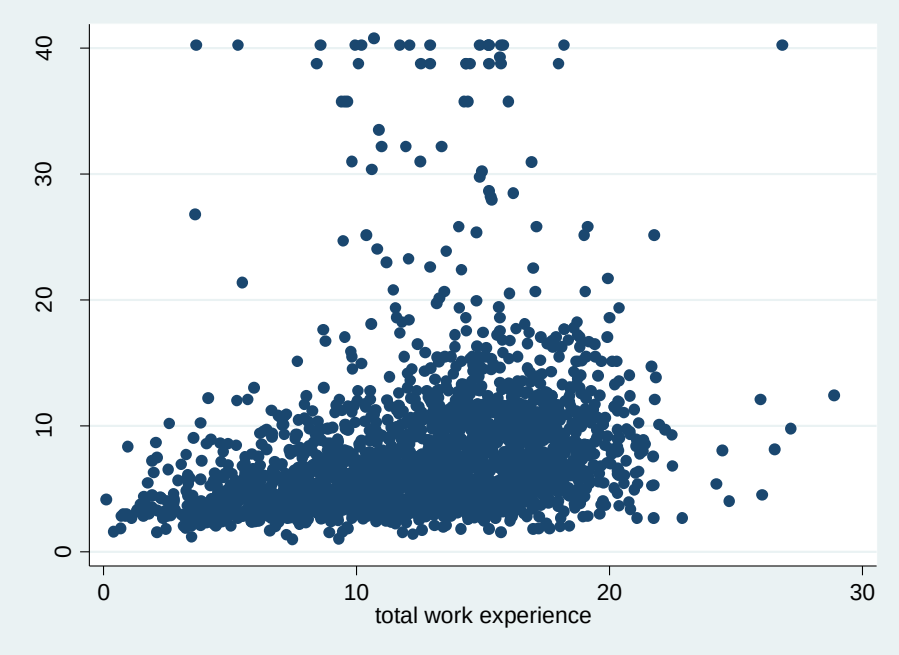
- You are interested in examining the relationship between total work experience (`ttl_exp`) and hourly wage (`wage`)

- You hypothesize that wage is dependent on work experience

IX. Bivariate Correlation and Regression

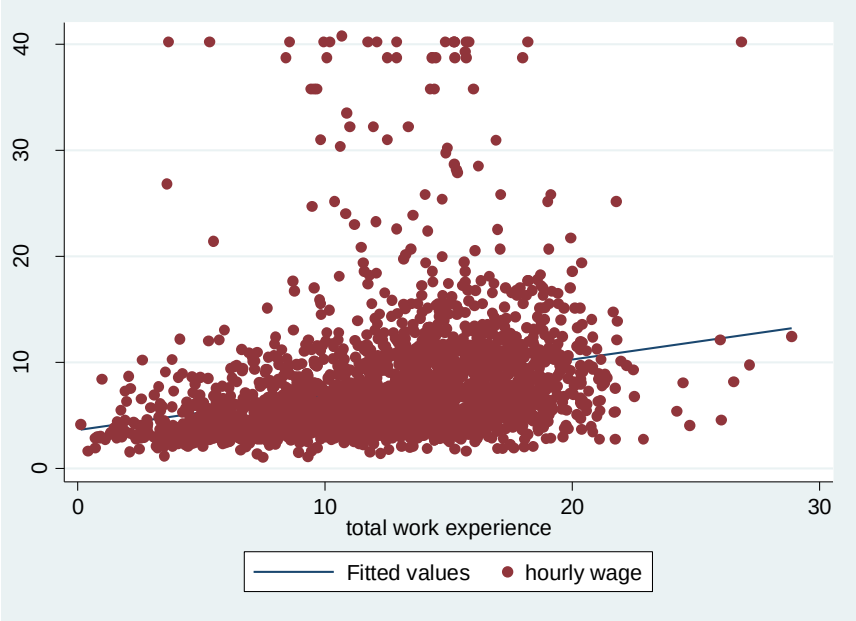
Visual Learners

- May want to see a scattergram
- `scatter wage ttl_exp`



Visual Learners

- Some may want to see the fitted line through the scattergram
- `twoway (lfit wage ttl_exp) (scatter wage ttl_exp)`



IX. Bivariate Correlation and Regression

By the Numbers

- May prefer the correlation matrix
- `pwcorr wage ttl_exp, obs sig star(0.05)`

By The Numbers

-

```
             |      wage  ttl_exp
-------------+------------------
        wage |    1.0000
             |
             |      2246
             |
     ttl_exp |    0.2655*  1.0000
             |    0.0000
             |      2246     2246
```

IX. Bivariate Correlation and Regression

Is the IV associated with the DV?

- Stata can run various regressions based on the distribution
- If we assume the DV has a normal distribution, simply use the following format:

```
regress dv iv
regress wage ttl_exp
```

Output

```
. regress wage ttl_exp

      Source |       SS       df       MS              Number of obs =    2246
-------------+------------------------------           F(  1,  2244) =  170.14
       Model |  5241.29609     1  5241.29609           Prob > F      =  0.0000
    Residual |  69126.6713  2244   30.805112           R-squared     =  0.0705
-------------+------------------------------           Adj R-squared =  0.0701
       Total |  74367.9674  2245  33.1260434           Root MSE      =  5.5502

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ttl_exp |   .3314291   .0254087    13.04   0.000     .2816021    .3812562
       _cons |   3.612492   .3393469    10.65   0.000     2.947026    4.277959
------------------------------------------------------------------------------
```

X. Multiple Regression

- Objectives
  - Estimate a multiple regression model

IX. Multiple Regression

Example

- You are interested in examining correlates of `wage`
- In addition to work experience, you think job `tenure` (in years), and whether someone is a college `grad` may also be associated with wage

X. Multiple Regression

Are the IVs associated with the DV?

- If we assume the DV has a normal distribution, simply use the following format:

`regress dv iv1 iv2 …ivn`

`regress wage ttl_exp tenure collgrad`

Output

```
. regress wage ttl_exp tenure collgrad

      Source |       SS       df       MS              Number of obs =    2231
-------------+------------------------------           F(  3,  2227) =  108.04
       Model |  9414.34005     3  3138.11335            Prob > F      =  0.0000
    Residual |  64687.4876  2227  29.0469185            R-squared     =  0.1270
-------------+------------------------------           Adj R-squared =  0.1259
       Total |  74101.8276  2230  33.2295191            Root MSE      =  5.3895

------------------------------------------------------------------------------
        wage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     ttl_exp |    .273864   .0304097     9.01   0.000     .2142298    .3334983
      tenure |   .0321168   .0253609     1.27   0.206    -.0176167    .0818502
    collgrad |   3.251781   .2698484    12.05   0.000       2.7226    3.780962
       _cons |   3.389638   .3397438     9.98   0.000      2.72339    4.055886
------------------------------------------------------------------------------
```

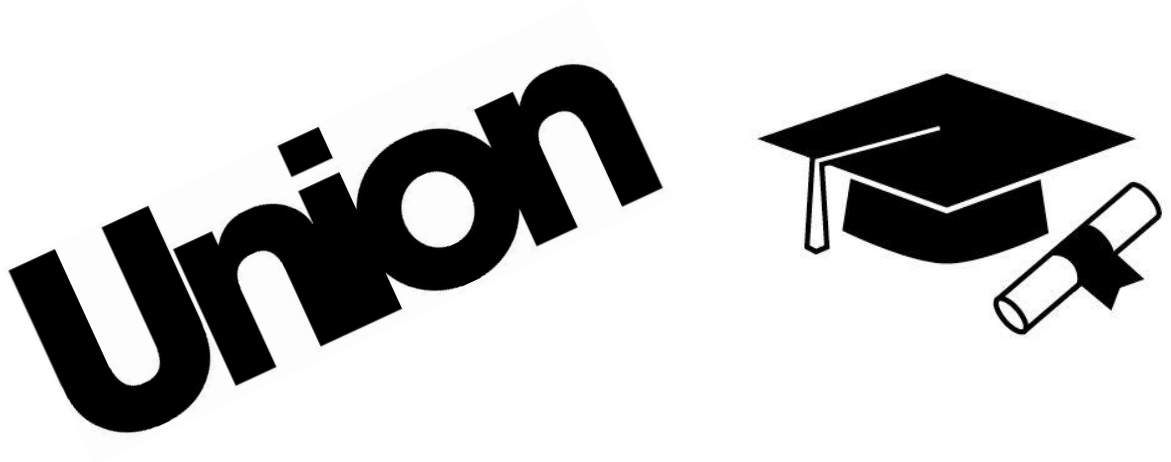
XI. Logistic Regression

- Objectives
  - Estimate a simple logistic regression model
  - Estimate a multiple logistic regression model

XI. Logistic Regression

Simple Logistic Regression

- You are interested in examining whether being a union worker (`union`) is dependent on being a college graduate (`collgrad`)
- What is the dependent variable here?
- How is it measured?



XI. Logistic Regression

Simple Logistic Regression

- `logistic union collgrad`

```
. logistic union collgrad

Logistic regression                                     Number of obs   =       1878
                                                        LR chi2(1)      =      17.30
                                                        Prob > chi2     =     0.0000
Log likelihood = -1037.9748                             Pseudo R2       =     0.0083

------------------------------------------------------------------------------
       union | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    collgrad |    1.64747   .1951016     4.22   0.000     1.306213    2.077883
       _cons |    .284287   .0182102   -19.64   0.000     .2507452    .3223157
------------------------------------------------------------------------------
```

XI. Logistic Regression

Multiple Logistic Regression

- You are interested in examining whether being a union worker (`union`) is dependent on being a college graduate (`collgrad`), age (`age`), and race/ethnicity (`race`)
- What is the dependent variable here?
- How is it measured?
- How are the IVs measured?

XI. Logistic Regression

Multiple Logistic Regression

- `xi: logistic union collgrad age i.race`

```
. xi: logistic union collgrad age i.race
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

Logistic regression                               Number of obs   =       1878
                                                  LR chi2(4)      =      33.72
                                                  Prob > chi2     =     0.0000
Log likelihood = -1029.7628                       Pseudo R2       =     0.0161

------------------------------------------------------------------------------
       union | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    collgrad |   1.734733   .2081249     4.59   0.000     1.371228    2.194602
         age |   1.013744   .0181203     0.76   0.445     .9788438    1.049889
    _Irace_2 |    1.60398    .191092     3.97   0.000      1.26996    2.025852
    _Irace_3 |    1.68724   .7450893     1.18   0.236     .7100424    4.009309
       _cons |   .1422289   .1008741    -2.75   0.006     .0354229    .5710734
------------------------------------------------------------------------------
```

XII. A Public Health Example

Prescription Stimulant Misuse

Background

- **Prescription Stimulant Misuse**: use of stimulants without a valid prescription, use in excess of a prescription, and/or use for purposes other than prescribed
- Health-related complications and Misuse
  - Increased blood pressure
  - Increased heart rate
  - Addiction
  - Psychosis
- Disproportionately affects college population
  - Research exploring prevalence and correlates important for prevention and intervention

XII. A Public Health Example

Misuse of Prescription Stimulants

Pilot Study

- Random Sampling
  - One Stage Cluster Sample
  - 19 classes
  - 94.71% response proportion
  - Analytic N = 499 students
  - Sample representative of undergraduate population
  - **Participants completed 100-item BEACH-Q questionnaire**

XII. A Public Health Example

Data Entry

- Survey
- Codebook
- Excel
- Import

Once in Stata

- Variable response options
- Variable labels
- Descriptive Data
- Inferential Data

XIII. What's Next

- Tobit regression
- Poisson regression
- Nested regression
- Factor analysis
- Longitudinal data analysis
- Hierarchical Models
- Structural Equation Modeling

XIII. What's Next

Resources

- Stata textbooks
- UCLA IDRE

idre

UCLA

Thank You!

Before & After Stata

Acknowledgements

- Dr. Alan Acock, Author of “A Gentle Introduction to Stata”
- CSULB BUILD Program (NIH Award #RL5GM118978)

BUILD
CSULB