Using STATA for meta analysis and regression

Jenny Chen

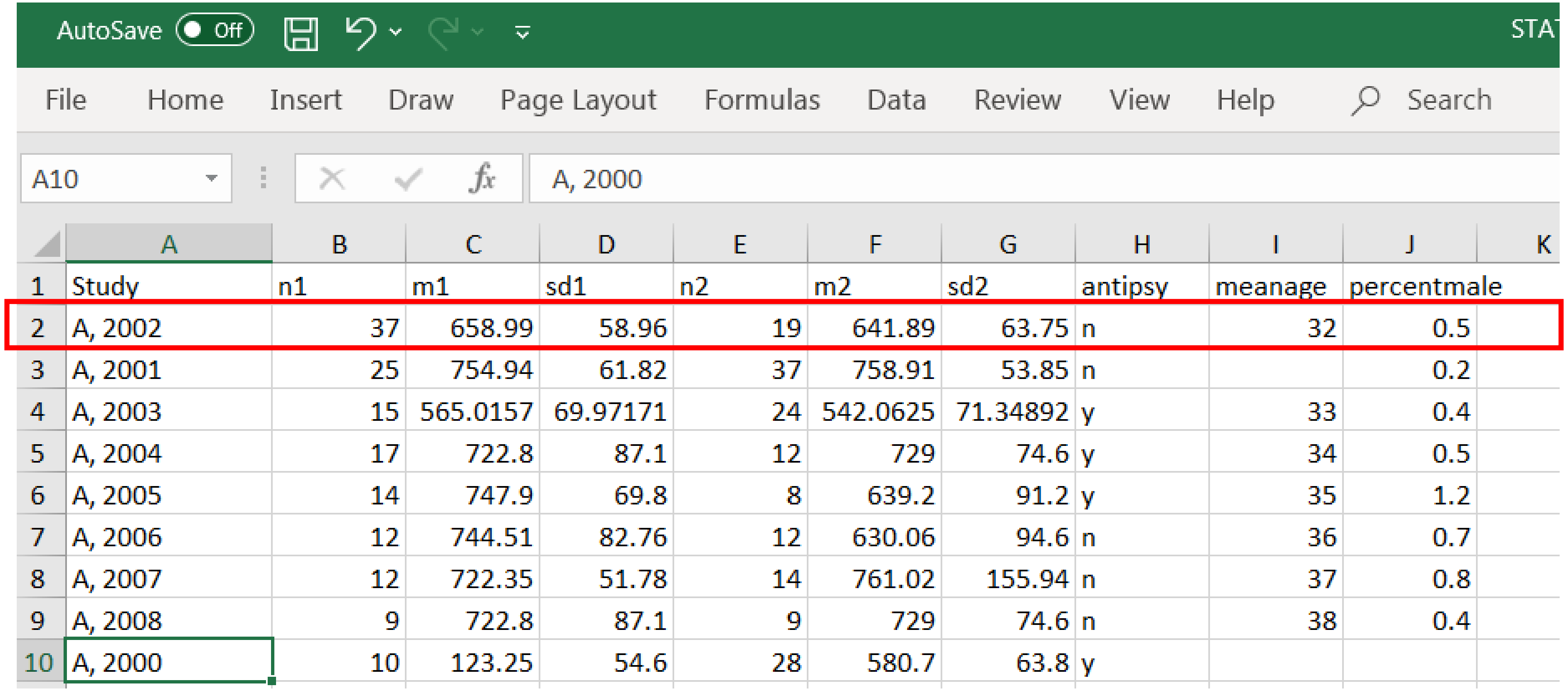
Neuropsychopharmacology Lab

What is STATA?

- **STATA** is a statistical software that we use to analyse data for the purpose of a meta-analysis. STATA give you the option of running a meta-regression, which can help us identify variables (covariates) contributing to heterogeneity
- Other popular programs (e.g. RevMan) are more user-friendly but have limitations
- STATA uses command-line interface and also allows user-written commands

To Start…

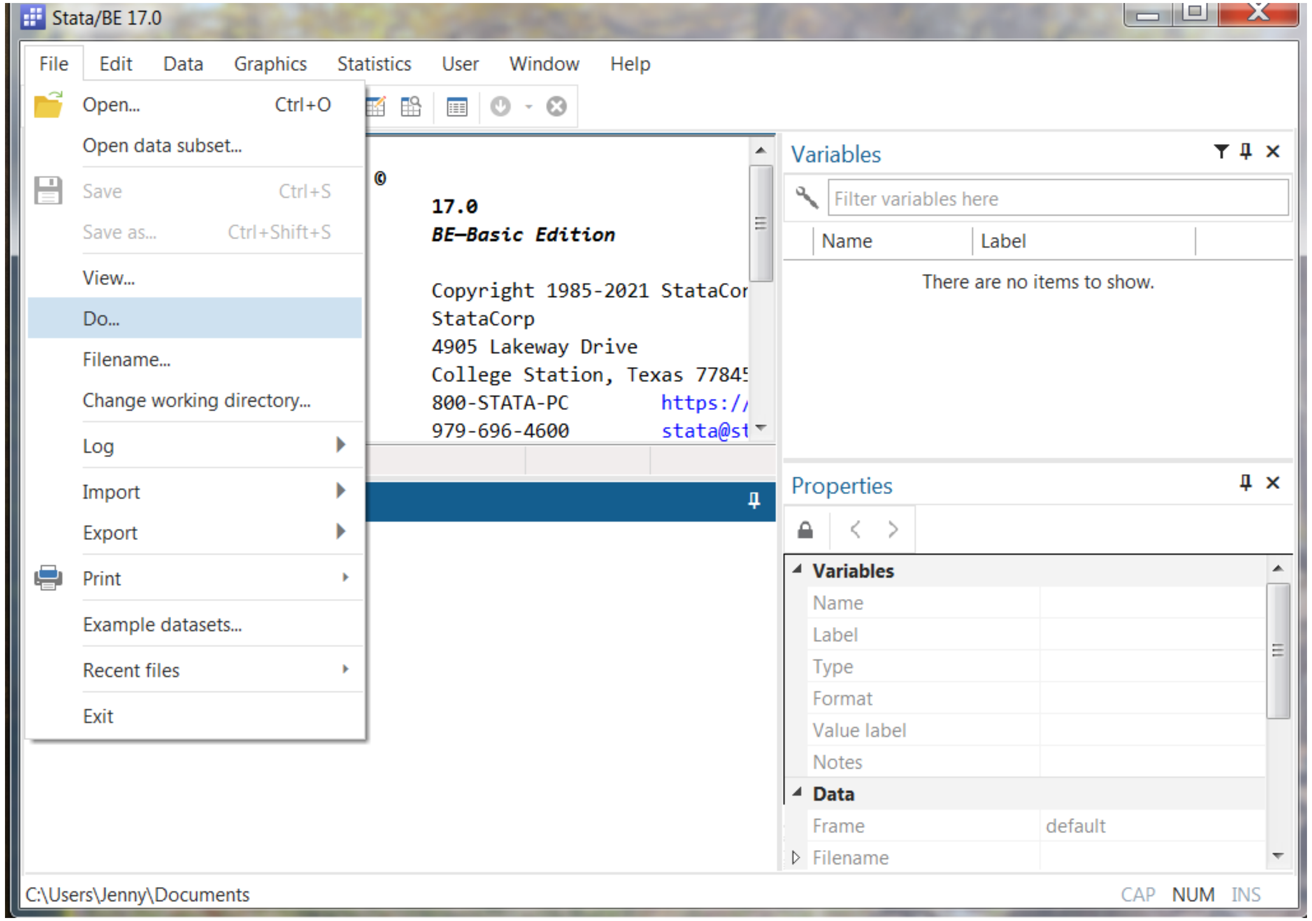
- You can find the files you need for STATA in neuroshared:\STATA files
  - **Please make a copy of it in your own folders and do not modify the files in this folder!**
- We will be using a toy dataset with a continuous outcome (STATA demo.xlsx)



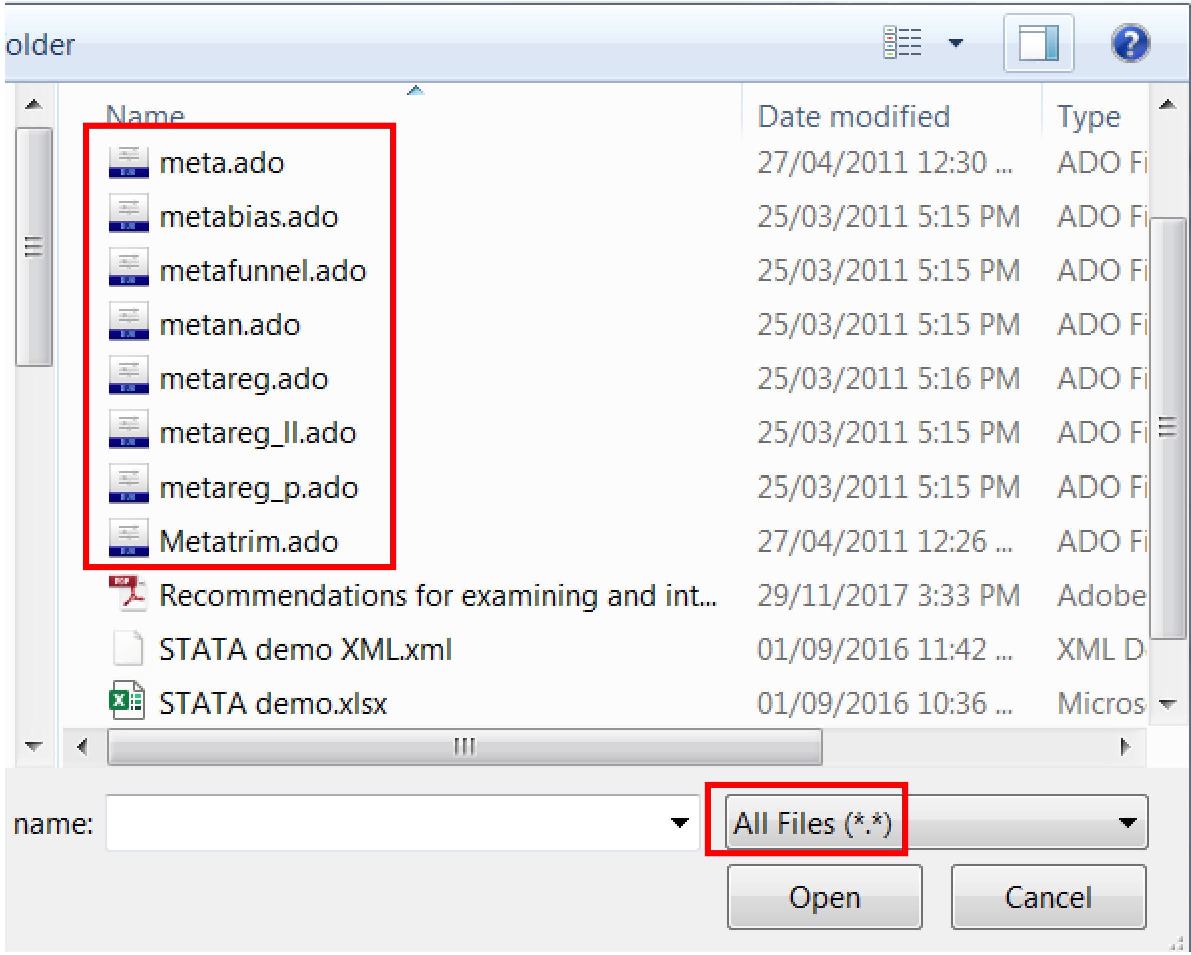
A10 | A, 2000

	A	B	C	D	E	F	G	H	I	J	K
1	Study	n1	m1	sd1	n2	m2	sd2	antipsy	meanage	percentmale	
2	A, 2002	37	658.99	58.96	19	641.89	63.75	n	32	0.5	
3	A, 2001	25	754.94	61.82	37	758.91	53.85	n		0.2	
4	A, 2003	15	565.0157	69.97171	24	542.0625	71.34892	y	33	0.4	
5	A, 2004	17	722.8	87.1	12	729	74.6	y	34	0.5	
6	A, 2005	14	747.9	69.8	8	639.2	91.2	y	35	1.2	
7	A, 2006	12	744.51	82.76	12	630.06	94.6	n	36	0.7	
8	A, 2007	12	722.35	51.78	14	761.02	155.94	n	37	0.8	
9	A, 2008	9	722.8	87.1	9	729	74.6	n	38	0.4	
10	A, 2000	10	123.25	54.6	28	580.7	63.8	y			

Open STATA and load .ado files

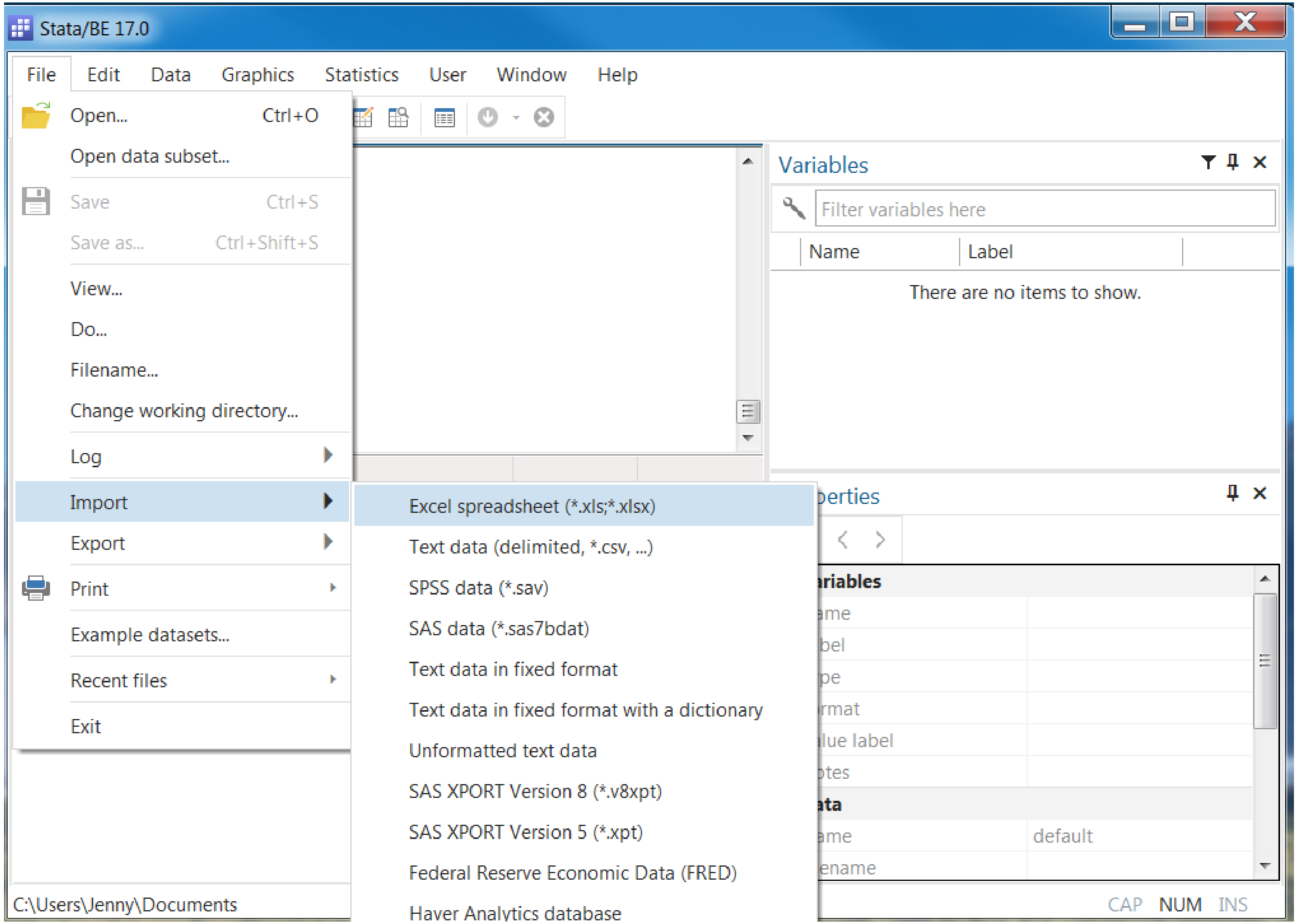


```
®
   17.0
   BE—Basic Edition

   Copyright 1985-2021 StataCor
   StataCorp
   4905 Lakeway Drive
   College Station, Texas 77845
   800-STATA-PC        https://
   979-696-4600        stata@st
```

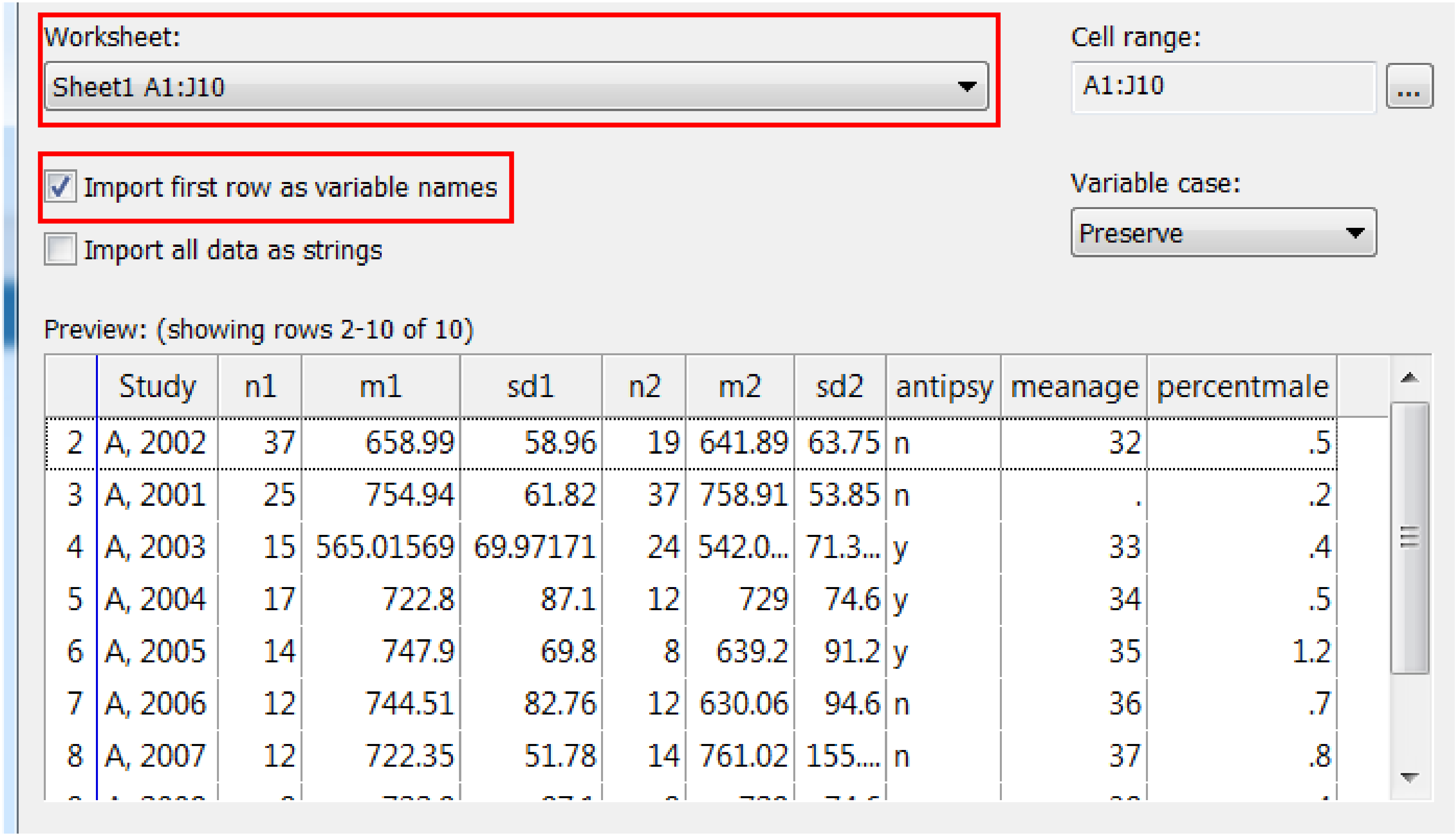


Name	Date modified	Type
meta.ado	27/04/2011 12:30 ...	ADO Fi
metabias.ado	25/03/2011 5:15 PM	ADO Fi
metafunnel.ado	25/03/2011 5:15 PM	ADO Fi
metan.ado	25/03/2011 5:15 PM	ADO Fi
metareg.ado	25/03/2011 5:16 PM	ADO Fi
metareg_ll.ado	25/03/2011 5:15 PM	ADO Fi
metareg_p.ado	25/03/2011 5:15 PM	ADO Fi
Metatrim.ado	27/04/2011 12:26 ...	ADO Fi
Recommendations for examining and int...	29/11/2017 3:33 PM	Adobe
STATA demo XML.xml	01/09/2016 11:42 ...	XML D
STATA demo.xlsx	01/09/2016 10:36 ...	Micros

Load your data



Load your data



Worksheet: Sheet1 A1:J10

Cell range: A1:J10

☑ Import first row as variable names

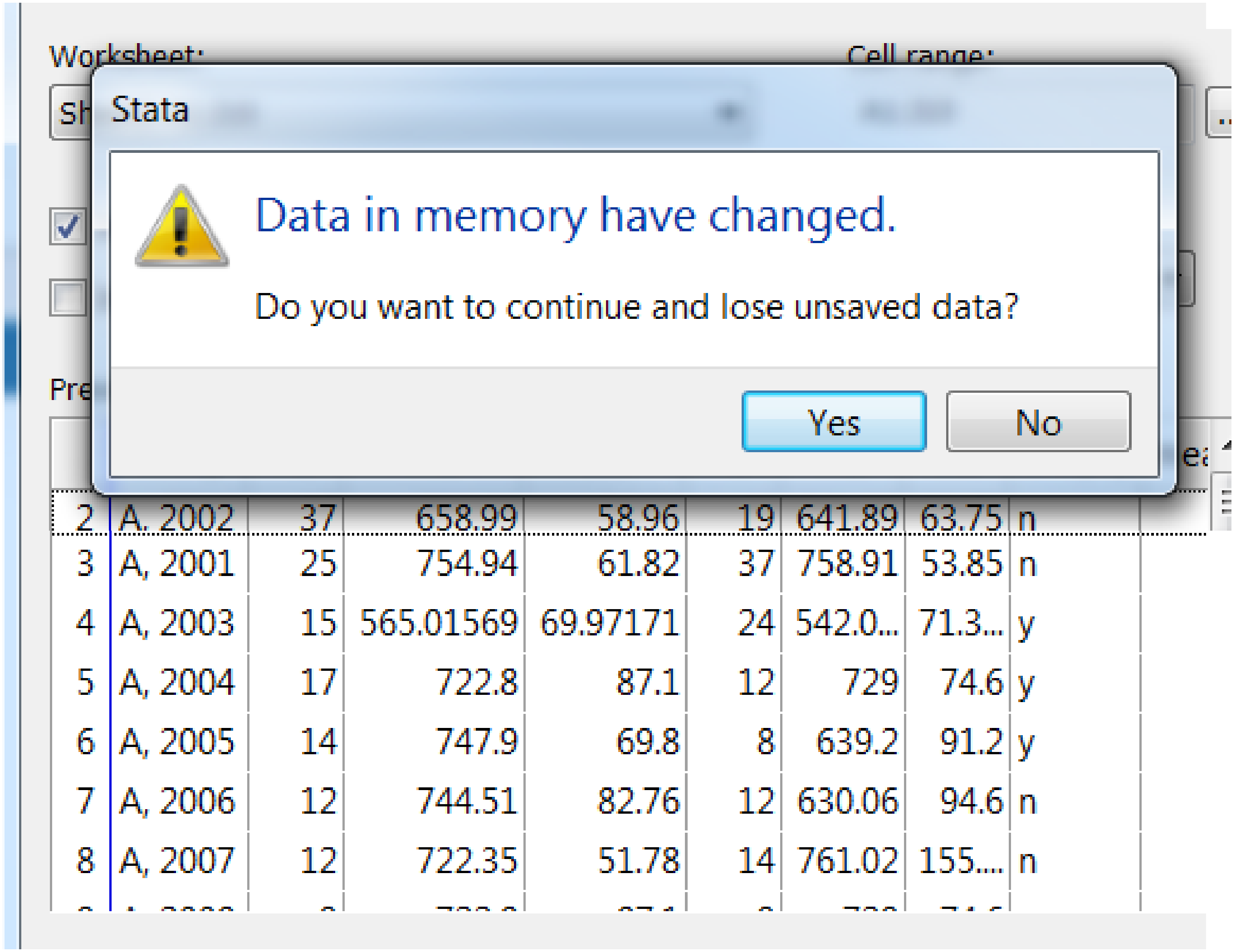
□ Import all data as strings

Variable case: Preserve

Preview: (showing rows 2-10 of 10)

	Study	n1	m1	sd1	n2	m2	sd2	antipsy	meanage	percentmale
2	A, 2002	37	658.99	58.96	19	641.89	63.75	n	32	.5
3	A, 2001	25	754.94	61.82	37	758.91	53.85	n	.	.2
4	A, 2003	15	565.01569	69.97171	24	542.0...	71.3...	y	33	.4
5	A, 2004	17	722.8	87.1	12	729	74.6	y	34	.5
6	A, 2005	14	747.9	69.8	8	639.2	91.2	y	35	1.2
7	A, 2006	12	744.51	82.76	12	630.06	94.6	n	36	.7
8	A, 2007	12	722.35	51.78	14	761.02	155....	n	37	.8

Load your data

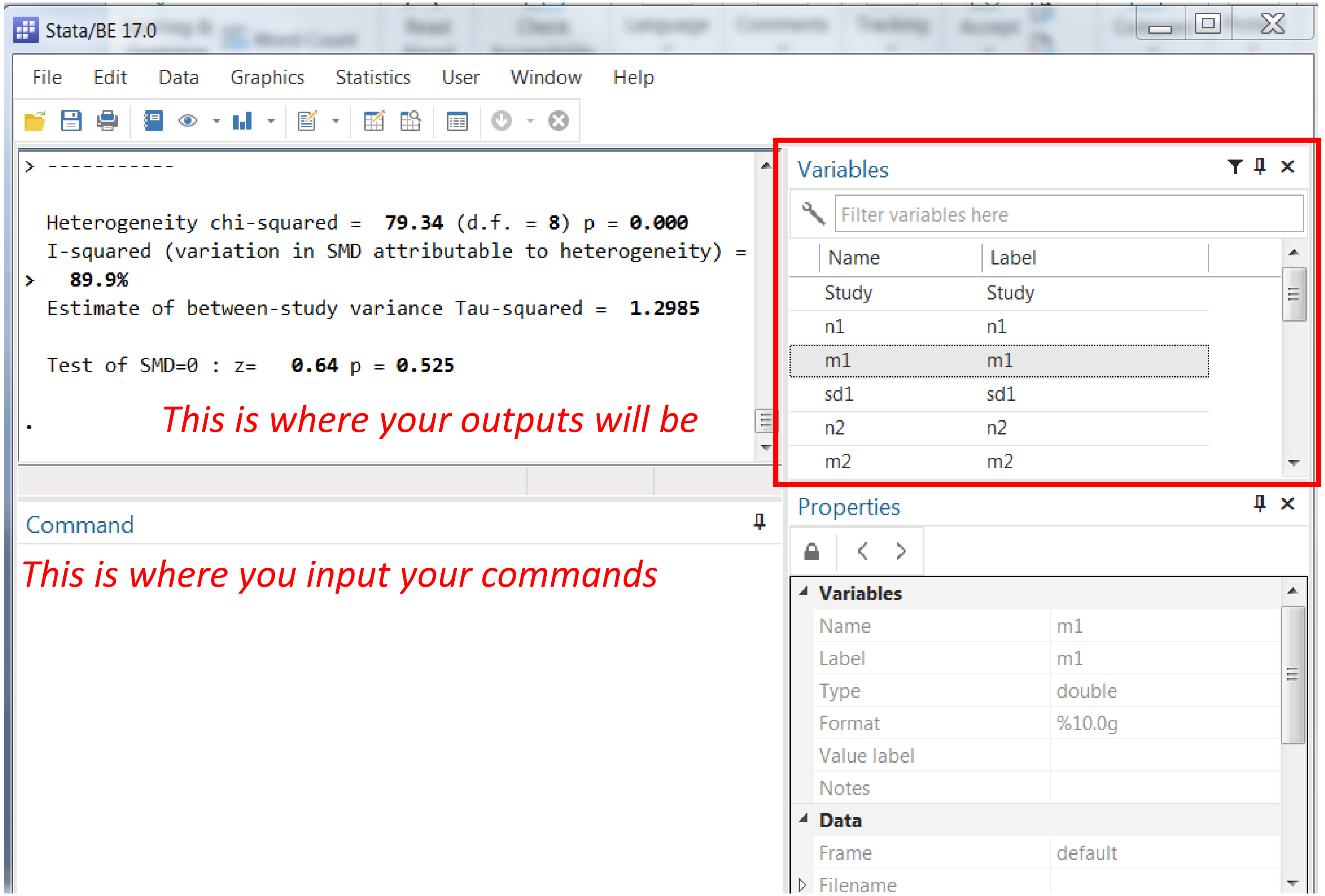


Data in memory have changed.

Do you want to continue and lose unsaved data?

2	A. 2002	37	658.99	58.96	19	641.89	63.75	n
3	A, 2001	25	754.94	61.82	37	758.91	53.85	n
4	A, 2003	15	565.01569	69.97171	24	542.0...	71.3...	y
5	A, 2004	17	722.8	87.1	12	729	74.6	y
6	A, 2005	14	747.9	69.8	8	639.2	91.2	y
7	A, 2006	12	744.51	82.76	12	630.06	94.6	n
8	A, 2007	12	722.35	51.78	14	761.02	155....	n

Load your data



```
> -----------

  Heterogeneity chi-squared =  79.34 (d.f. = 8) p = 0.000
  I-squared (variation in SMD attributable to heterogeneity) =
>   89.9%
  Estimate of between-study variance Tau-squared =  1.2985

  Test of SMD=0 : z=   0.64 p = 0.525

.
```

This is where your outputs will be

Command

This is where you input your commands

Variables

Filter variables here

Name	Label
Study	Study
n1	n1
m1	m1
sd1	sd1
n2	n2
m2	m2

Properties

Variables	
Name	m1
Label	m1
Type	double
Format	%10.0g
Value label	
Notes	
Data	
Frame	default
Filename	

Some decisions to make before your run your meta analysis

Random or Fixed?

- **Random effects** model most often used for heterogeneous studies (use different methods, different time values available, different age ranges across studies)
  - Random effects most often used when you have large data over a long period of time or if you think there is a possibility that the authors may have used different methods to obtain their outcomes
- **Fixed effects**, we impose time-independent effects for each entity that are possibly correlated with the regressors, aka you're saying that heterogeneity is constant.

Some decisions to make before your run your meta analysis

SMD or WMD?

- **Standardized mean differences** - if there were different scales used but the same outcome is measured
  - Eg. All studies measure depression but some use CES-D, some use HAM-D, some use GDS.
  - Expresses the size of the intervention effect in each study relative to the variability observed in that study.
- **Mean difference** (more correctly, 'difference in means') is a standard statistic that measures the absolute difference between the mean value in two groups in a clinical trial
  - Estimates the amount by which the experimental intervention changes the outcome on average compared with the control
  - Can be used when outcome measurements in all studies are made on the same scale, eg. If they all use the HAM-D
  - *Aside: Analyses based on this effect measure have historically been termed weighted mean difference (WMD) analyses in* Cochrane. *This name is potentially confusing: although the meta-analysis computes a weighted average of these differences in means, no weighting is involved in calculation of a statistical summary of a single study. Furthermore, all meta-analyses involve a weighted combination of estimates, yet we do not use the word 'weighted' when referring to other methods.*

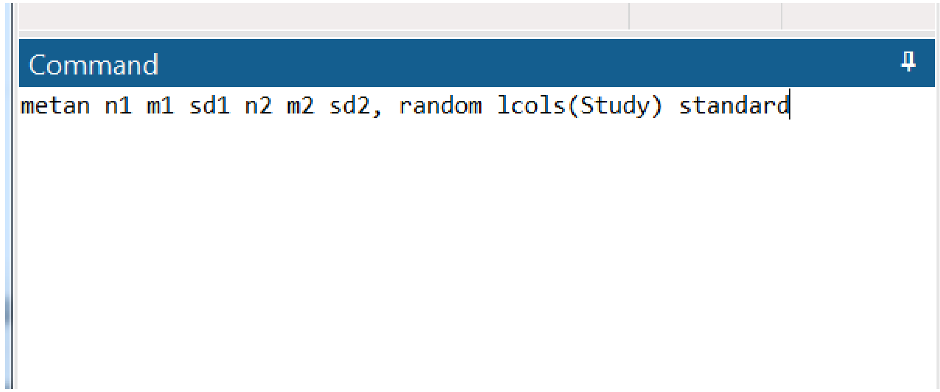
Syntax for comparison of continuous outcomes in different groups (the “base” meta analysis)

- Program needed: meta.ado, metan.ado
- Syntax: metan n1 m1 sd1 n2 m2 sd2, random lcols(Study) standard

  *** You must run this before running a regression or any other analysis***

- Random or fixed – interchange these for random effects or fixed effects model
- Lcols(Study) – tells STATA to present the study names in the left column of the effect size table
- Standard or nostandard – interchange these for standardized mean differences or weighted mean differences in the model

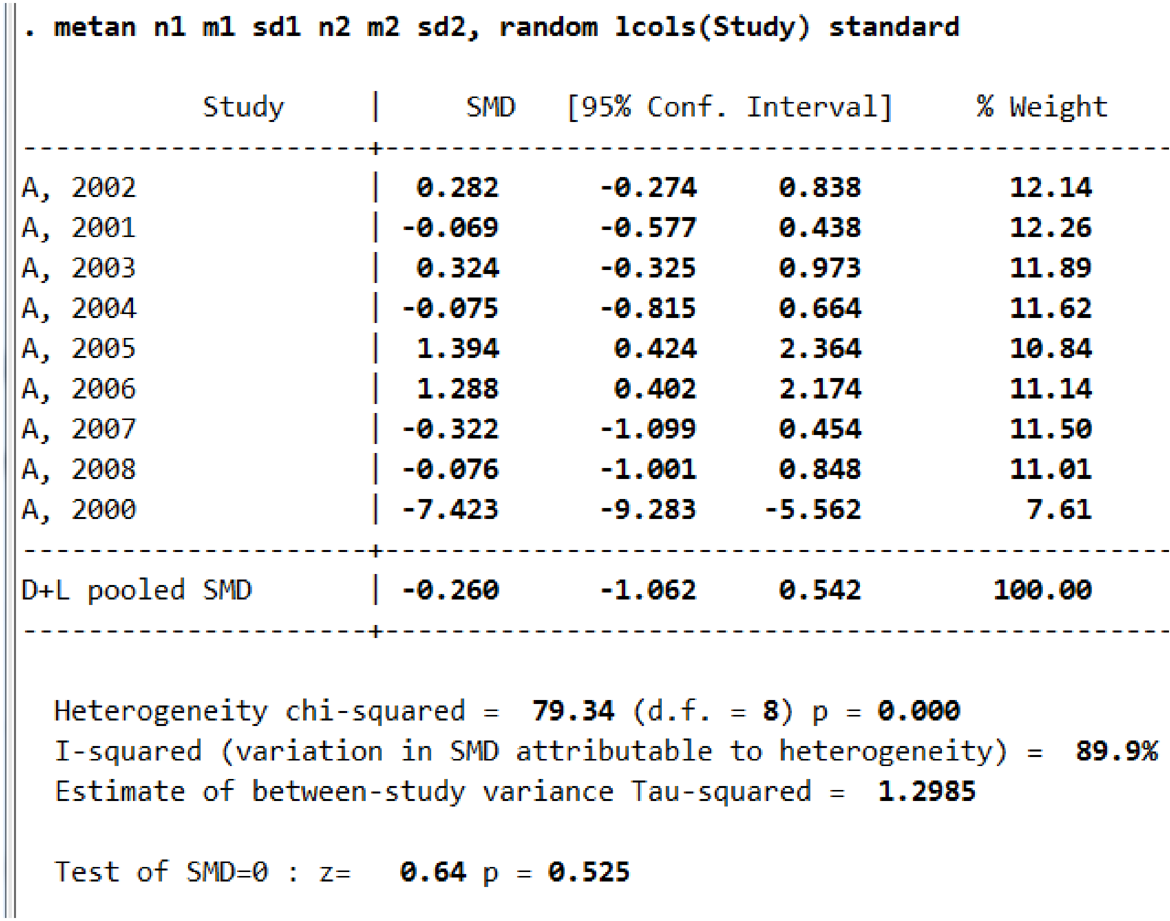
** Note that STATA is a **case sensitive**, so **Study** and **study** are different variables.



```
metan n1 m1 sd1 n2 m2 sd2, random lcols(Study) standard
```

Output for comparison of continuous outcomes in different groups (the “base” meta analysis)

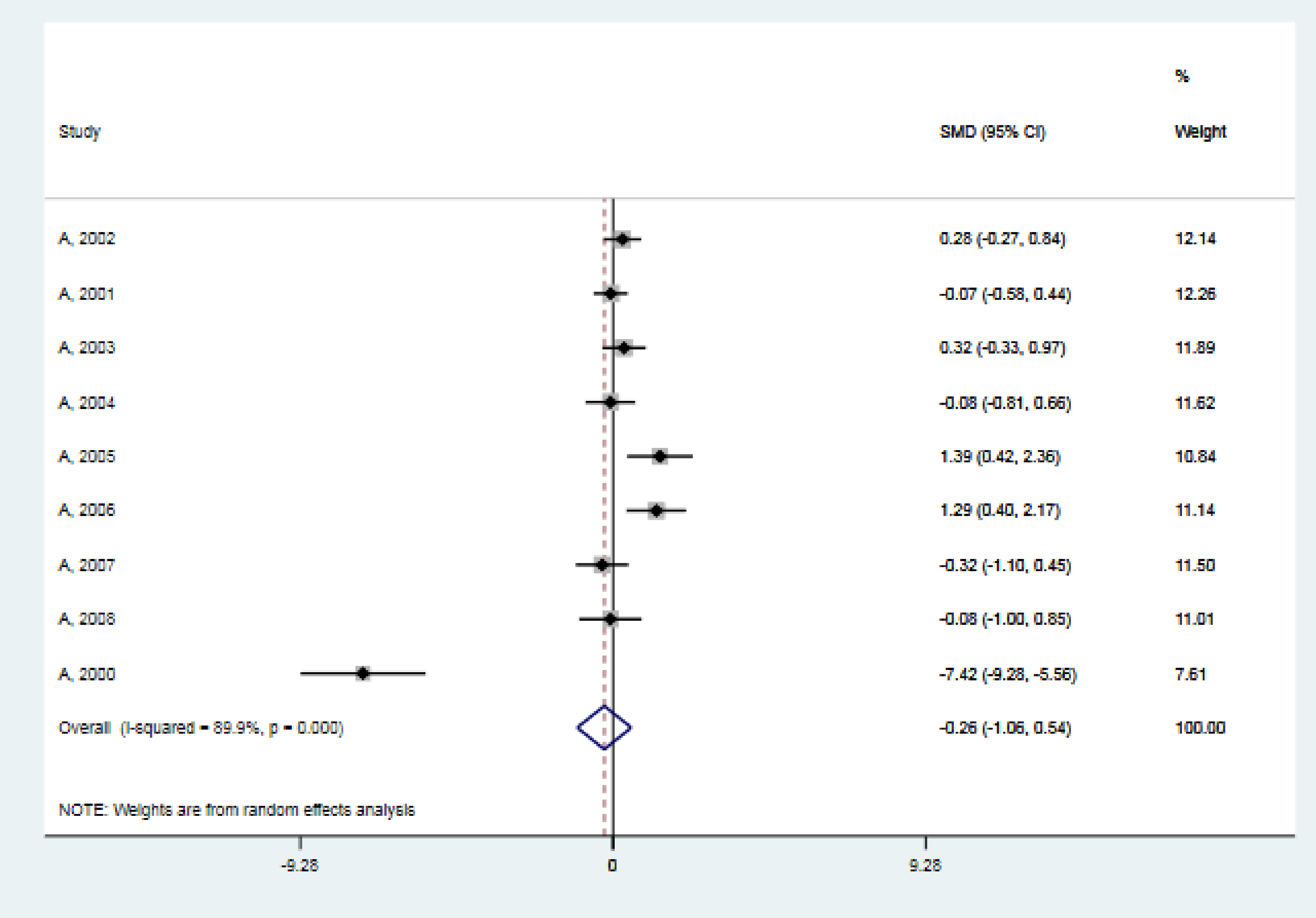
- For each study
  - SMD: 0.2=small effect, 0.5= moderate, 0.8= large
  - 95% CI: (how reliable that estimate is; if you are measuring this outcome with the other samples of the same size and population, 95% of those CIs produced would capture the true value. Sample size and spread of data influence CI)
  - % Weight
- Pooled SMD
- Heterogeneity chi-squared
- I-squared (**this is important! 0-40%= minimal heterogeneity; 30-60%= moderate; 50-90% substantial; 90-100% considerable)**
- Tau-squared
- z score (for SMD=0)
- p value **(P<0.05 = significance difference between groups)**
- forest plot (see next slide)



```
. metan n1 m1 sd1 n2 m2 sd2, random lcols(Study) standard

          Study        |     SMD    [95% Conf. Interval]     % Weight
-----------------------+-----------------------------------------------
A, 2002                |  0.282      -0.274     0.838          12.14
A, 2001                | -0.069      -0.577     0.438          12.26
A, 2003                |  0.324      -0.325     0.973          11.89
A, 2004                | -0.075      -0.815     0.664          11.62
A, 2005                |  1.394       0.424     2.364          10.84
A, 2006                |  1.288       0.402     2.174          11.14
A, 2007                | -0.322      -1.099     0.454          11.50
A, 2008                | -0.076      -1.001     0.848          11.01
A, 2000                | -7.423      -9.283    -5.562           7.61
-----------------------+-----------------------------------------------
D+L pooled SMD         | -0.260      -1.062     0.542         100.00
-----------------------+-----------------------------------------------

  Heterogeneity chi-squared =  79.34 (d.f. = 8) p = 0.000
  I-squared (variation in SMD attributable to heterogeneity) =  89.9%
  Estimate of between-study variance Tau-squared =  1.2985

  Test of SMD=0 : z=   0.64 p = 0.525
```

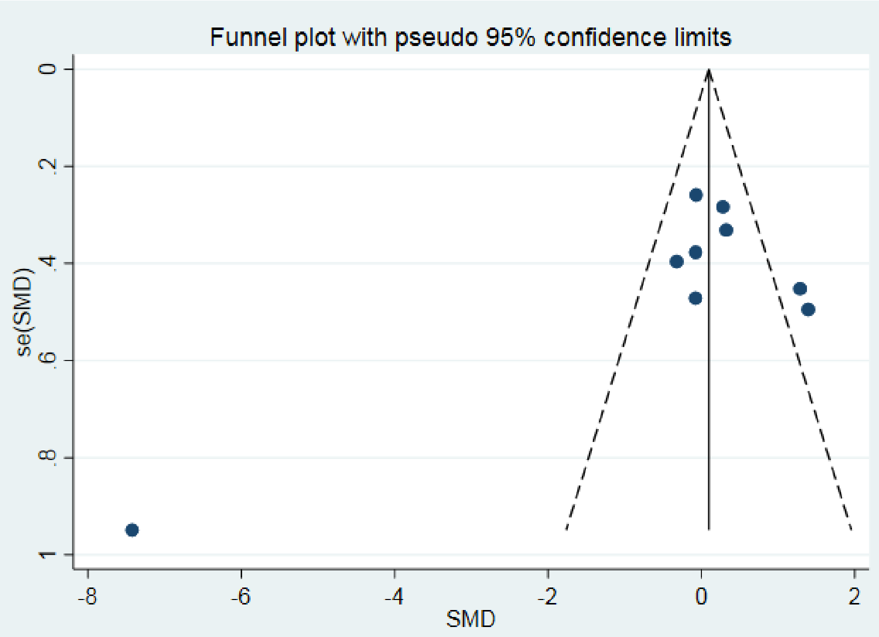
Forest Plot

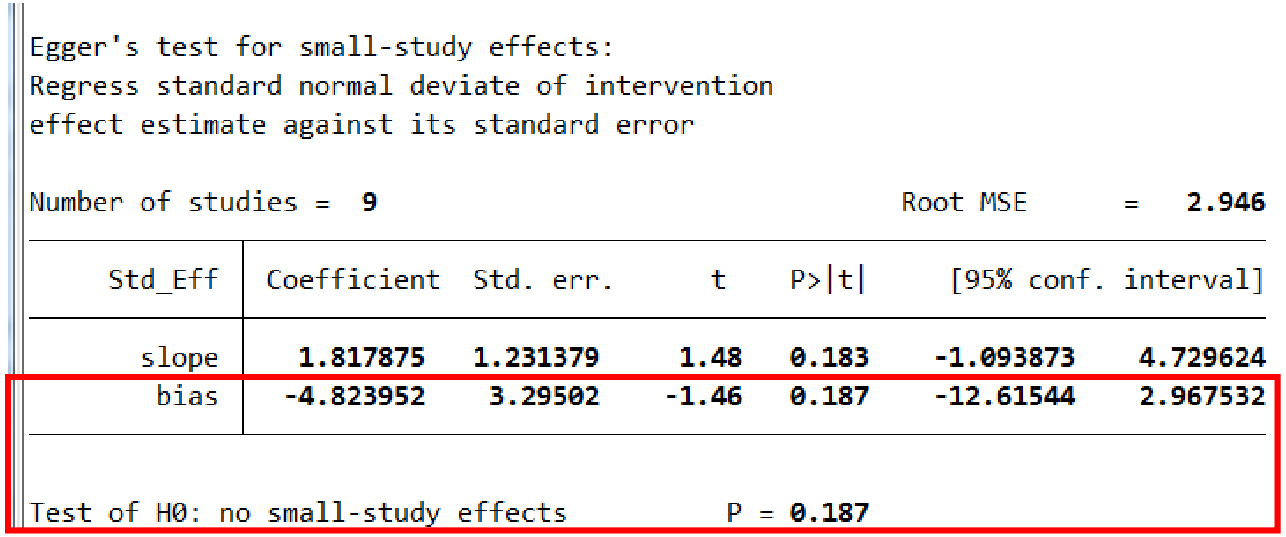
- The “basic” forest plot should be formatted (syntax in later slides)
- You can also use the Graph Editor within the graph to change the title and labels



Publication Bias

- A funnel plot is a visualization to investigate whether there was publication bias (qualitative)
  - look for asymmetry and outliers
- Egger’s Test – a statistical output (quantitative)
  - A significant Egger’s Test (p<0.05) means there is a small study effect (small studies (i.e., with smaller precision) show larger effect sizes).
  - i.e. there is likely publication bias





```
Egger's test for small-study effects:
Regress standard normal deviate of intervention
effect estimate against its standard error

Number of studies =  9                              Root MSE      =    2.946
------------------------------------------------------------------------------
     Std_Eff | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
       slope |   1.817875   1.231379     1.48   0.183    -1.093873    4.729624
        bias |  -4.823952    3.29502    -1.46   0.187    -12.61544    2.967532
------------------------------------------------------------------------------

Test of H0: no small-study effects          P = 0.187
```

Syntax for risk of publication bias (after metan is run)

For investigating the risk of publication bias

- Program needed: metabias.ado, metafunnel.ado
- Syntax for funnel plot: metafunnel _ES _seES
- Syntax for Egger’s test of bias: metabias _ES _seES, egger

Syntax for Trim and Fill

If bias is found, you can adjust for it using trim and fill

1. trim (i.e. remove) smaller studies causing funnel plot asymmetry,
2. use the trimmed funnel plot to estimate the true ‘centre’ of the funnel,
3. replace the omitted studies and their missing ‘counterparts’ around the centre (i.e. fill)

- Program needed: metatrim.ado
- Syntax: metatrim _ES _seES
- The output will show if anything was trimmed or filled. The new numbers are what you can report (pooled est. CI, z-value, **p-value**)
- Look in the table for random effects model
- Caveat: trim and fill does not take into account reasons for funnel plot asymmetry other than publication bias and also perform poorly in the presence of substantial between-study heterogeneity

Subgroup Analysis

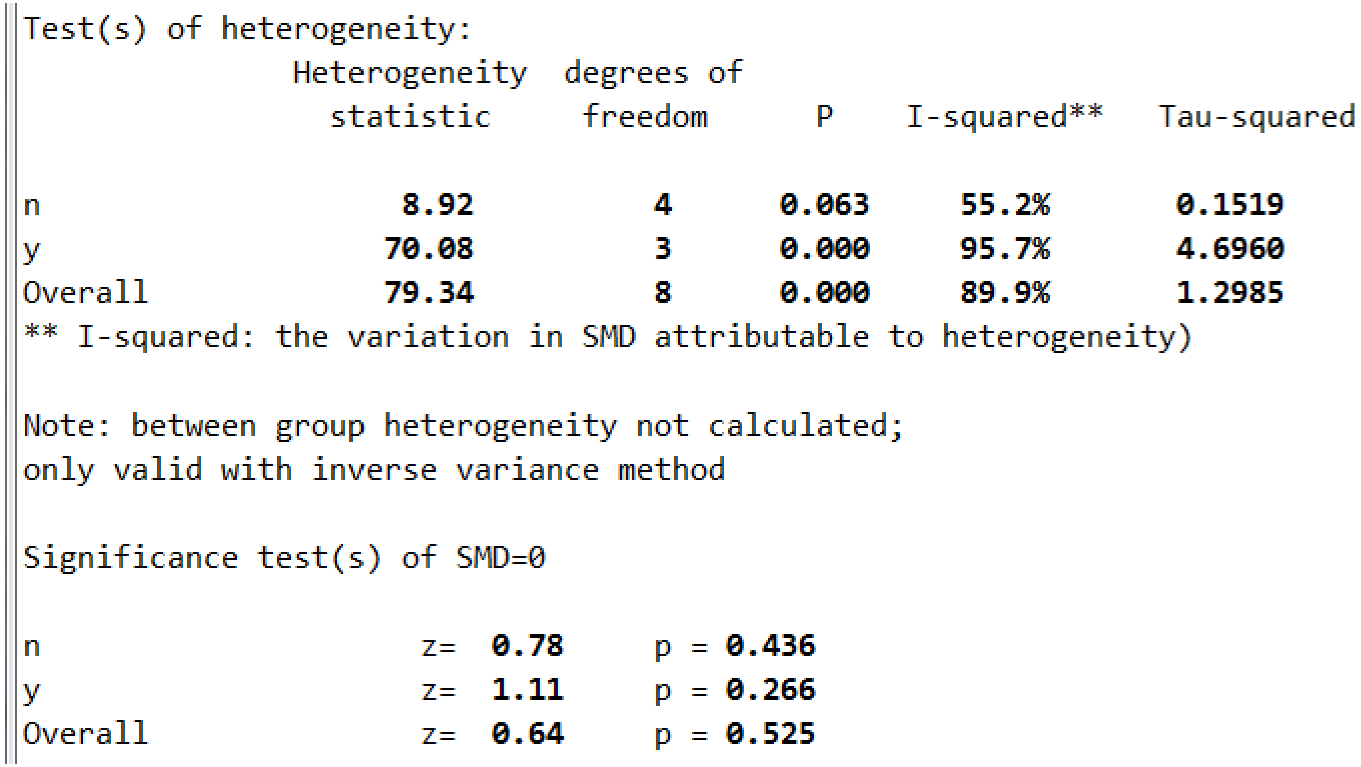
- This involves splitting all the participant data into subgroups
  - Eg. Subgroup analyses may be done for subsets of participants or for subsets of studies (eg. Studies that used antipsychotics).
  - May be done to investigate heterogeneity
  - Or answer specific questions about particular patient groups, types of intervention, or types of study
    - NOTE: Subgroup analyses are observational by nature and are not based on randomized comparisons
  - Subgroups are typically dichotomous (ie. one or the other)

Syntax for subgroup analysis

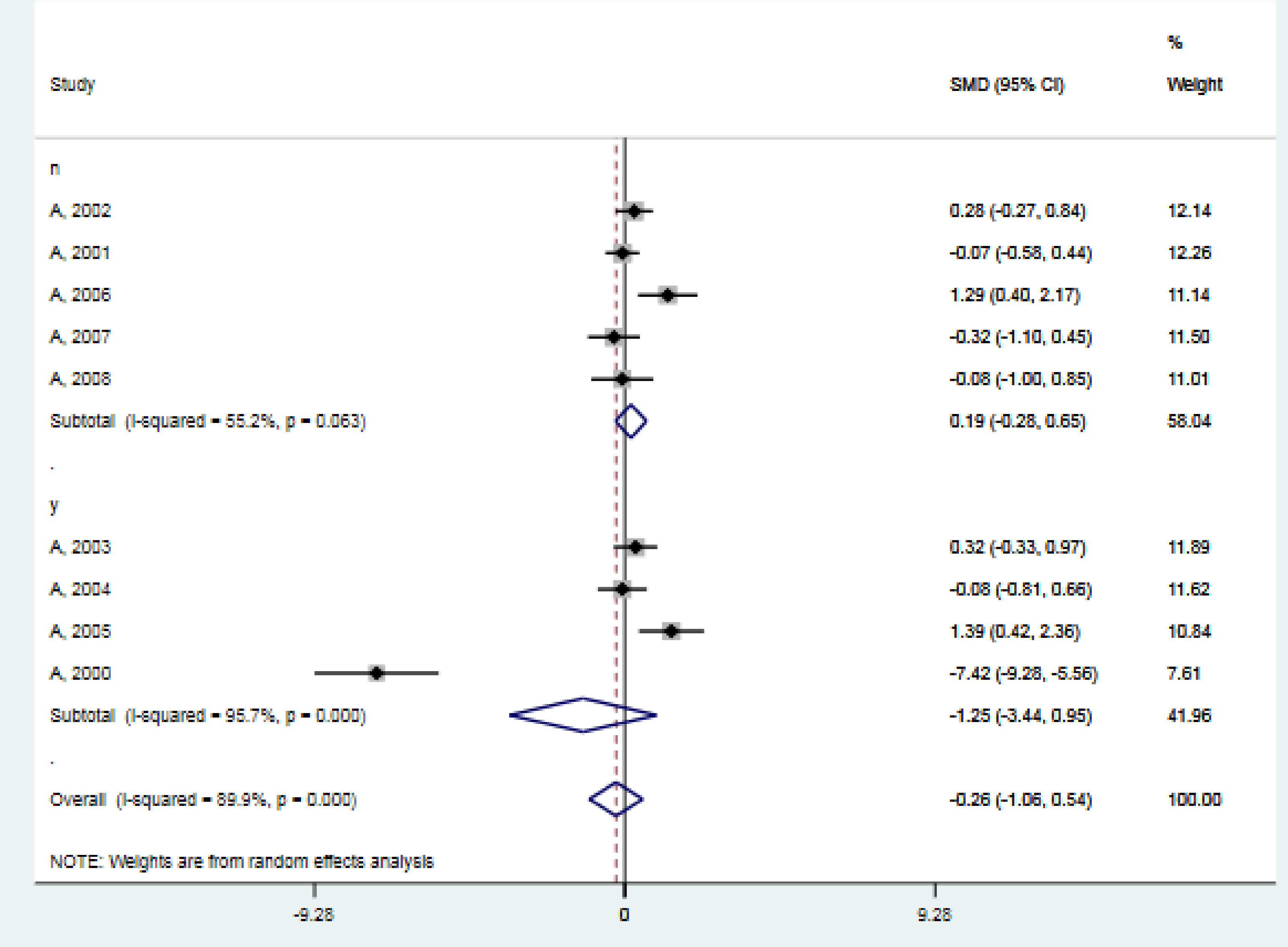
- Additional program needed: none
- Syntax: metan n1 m1 sd1 n2 m2 sd2, random lcols(Study) by(*characteristic name*)
    - Eg. metan n1 m1 sd1 n2 m2 sd2, random lcols(Study) by(antipsy)

What is outputted?

- Separates your pooled SMD by whether n (not on antipsy) and y (on antipsy).
- **New forest plot (next slide)**
- **New z scores and p values for n, y, and overall**



```
Test(s) of heterogeneity:
                   Heterogeneity  degrees of
                     statistic     freedom      P    I-squared**   Tau-squared

n                        8.92          4      0.063     55.2%        0.1519
y                       70.08          3      0.000     95.7%        4.6960
Overall                 79.34          8      0.000     89.9%        1.2985
** I-squared: the variation in SMD attributable to heterogeneity)

Note: between group heterogeneity not calculated;
only valid with inverse variance method

Significance test(s) of SMD=0

n                          z=   0.78     p = 0.436
y                          z=   1.11     p = 0.266
Overall                    z=   0.64     p = 0.525
```



Study	SMD (95% CI)	% Weight
n		
A, 2002	0.28 (-0.27, 0.84)	12.14
A, 2001	-0.07 (-0.58, 0.44)	12.26
A, 2006	1.29 (0.40, 2.17)	11.14
A, 2007	-0.32 (-1.10, 0.45)	11.50
A, 2008	-0.08 (-1.00, 0.85)	11.01
Subtotal (I-squared = 55.2%, p = 0.063)	0.19 (-0.28, 0.65)	58.04
.		
y		
A, 2003	0.32 (-0.33, 0.97)	11.89
A, 2004	-0.08 (-0.81, 0.66)	11.62
A, 2005	1.39 (0.42, 2.36)	10.84
A, 2000	-7.42 (-9.28, -5.56)	7.61
Subtotal (I-squared = 95.7%, p = 0.000)	-1.25 (-3.44, 0.95)	41.96
.		
Overall (I-squared = 89.9%, p = 0.000)	-0.26 (-1.06, 0.54)	100.00

NOTE: Weights are from random effects analysis

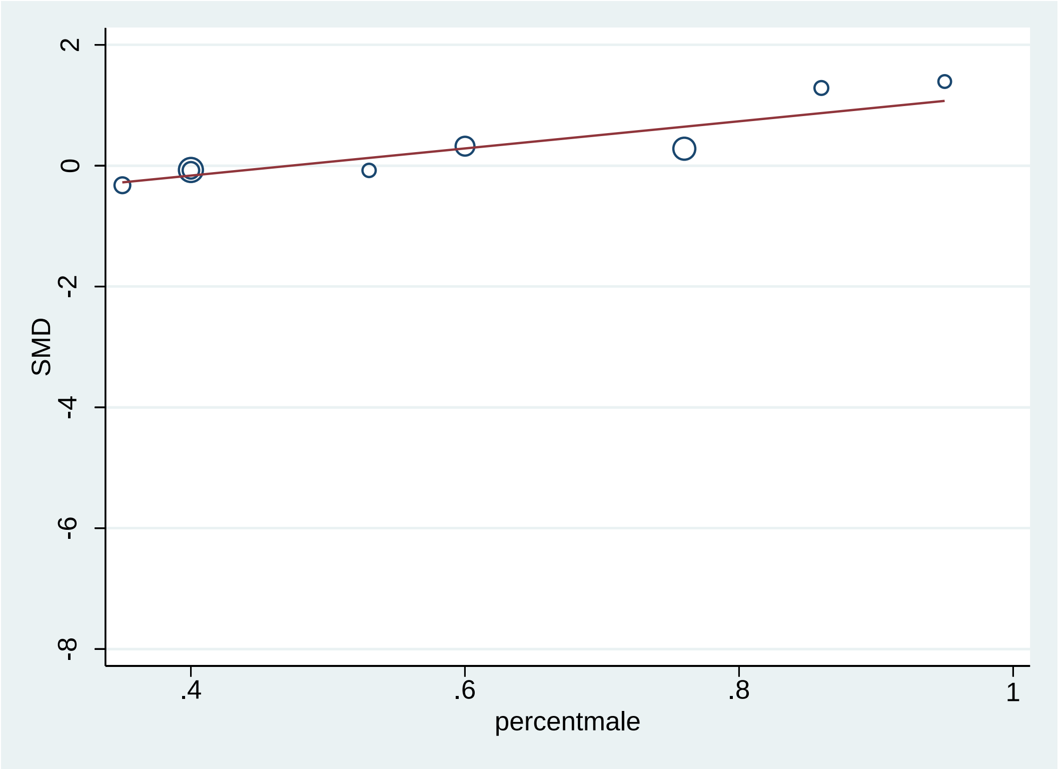
- SMDs for each study is groups by n and y (antipsy variable) with pooled SMD for the two subgroups

Meta-Regression

- Looks at relationship between effect size and a *continuous characteristic*
- Similar to simple regressions
  - **outcome variable** is predicted according to the values of one or more **explanatory variables**
- Additional program needed: metareg.ado, metareg_p.ado, metareg_II.ado
- Syntax: metareg _ES **characteristic name**, wsse(_seES) graph
- For multiple comparisons in the same meta-regression, the Syntax: metareg _ES **characteristic name* *SPACE* *second characteristic name**, wsse(_seES) graph
- For example: metareg _ES *percentmale*, wsse(_seES) graph
- Cochrane recommends at least ten observations (i.e. ten studies in a meta-analysis) should be available for each characteristic modelled, if the covariate is evenly distributed.

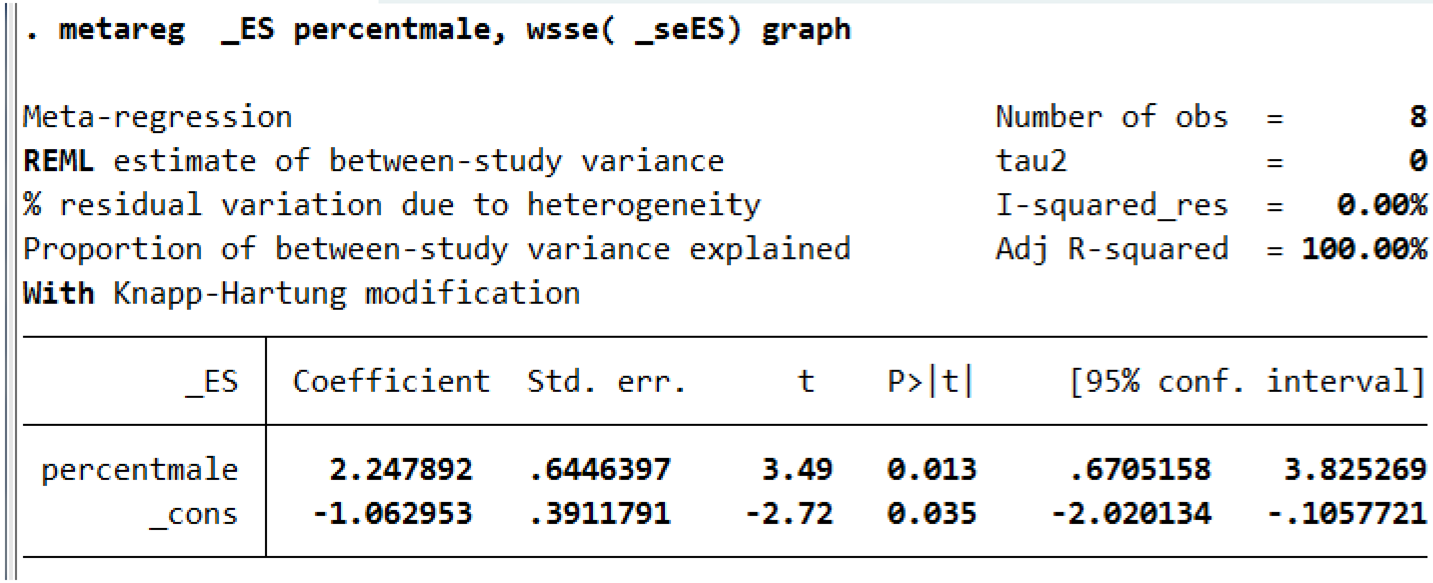
Meta-Regression Outputs

- Command gives you a bubble plot and summary table
- If the characteristic was a significant contributor to the outcome, then p<0.05.
- Interpretation of the intercept (_cons) depends on continuous variable



Of note:

- I-squared (did it lower the heterogeneity?)
- New p value (did it change the significance?)

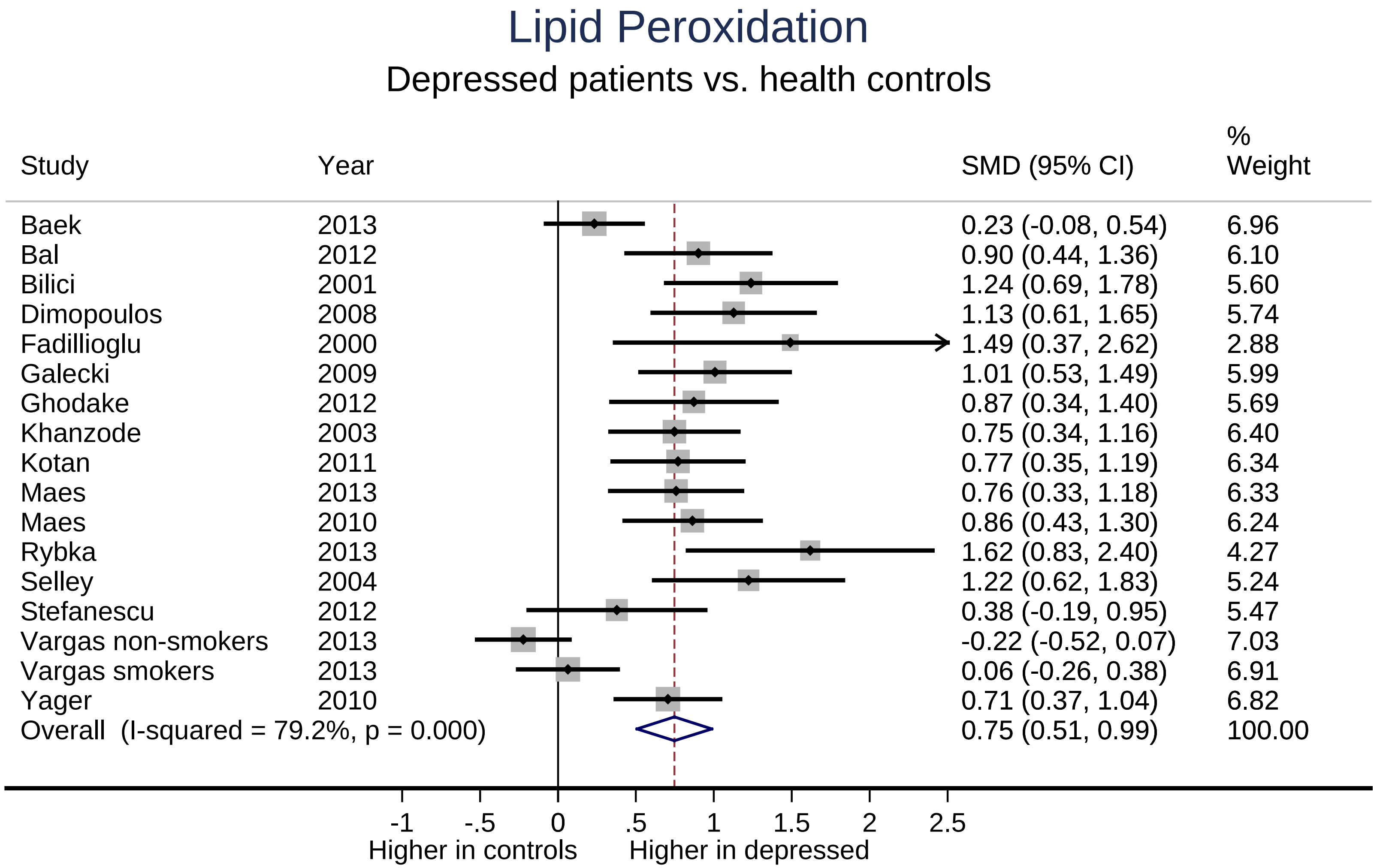


```
. metareg  _ES percentmale, wsse( _seES) graph

Meta-regression                                        Number of obs  =        8
REML estimate of between-study variance                tau2           =        0
% residual variation due to heterogeneity              I-squared_res  =    0.00%
Proportion of between-study variance explained         Adj R-squared  =  100.00%
With Knapp-Hartung modification
------------------------------------------------------------------------------
         _ES | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
 percentmale |   2.247892   .6446397     3.49   0.013     .6705158    3.825269
       _cons |  -1.062953   .3911791    -2.72   0.035    -2.020134   -.1057721
------------------------------------------------------------------------------
```

Customizing forest plots

Can use Graph Editor or code (below)

- metan n1 m1 sd1 n2 m2 sd2, random lcols(Study Year) sortby(Study) favours(Higher in controls # Higher in depressed) xlabel(-1,-0.5,0,0.5,1,1.5,2,2.5) astext(60) boxsca(150) xsize(20)ysize(13) force texts(120)
- sortby(Study) – tells STATA to sort the study names alphabetically in the figure
- favours(Higher in XXX # Higher in YYY) – tells STATA to label the left side of the forest plot for "effect is greater in XXX" and right side for "effect is greater in YYY"
    - This depends on your outcome and how you specified your group 1 and 2! (Example in next slide)
- xlabel(-1,-0.5,0,0.5,1,1.5,2,2.5) – assigns the values of the scale on the X axis
- astext(60) boxsca(150) xsize(20)ysize(13) force texts(120) – modify value until you get the presentation you want



Study	Year	SMD (95% CI)	% Weight
Baek	2013	0.23 (-0.08, 0.54)	6.96
Bal	2012	0.90 (0.44, 1.36)	6.10
Bilici	2001	1.24 (0.69, 1.78)	5.60
Dimopoulos	2008	1.13 (0.61, 1.65)	5.74
Fadillioglu	2000	1.49 (0.37, 2.62)	2.88
Galecki	2009	1.01 (0.53, 1.49)	5.99
Ghodake	2012	0.87 (0.34, 1.40)	5.69
Khanzode	2003	0.75 (0.34, 1.16)	6.40
Kotan	2011	0.77 (0.35, 1.19)	6.34
Maes	2013	0.76 (0.33, 1.18)	6.33
Maes	2010	0.86 (0.43, 1.30)	6.24
Rybka	2013	1.62 (0.83, 2.40)	4.27
Selley	2004	1.22 (0.62, 1.83)	5.24
Stefanescu	2012	0.38 (-0.19, 0.95)	5.47
Vargas non-smokers	2013	-0.22 (-0.52, 0.07)	7.03
Vargas smokers	2013	0.06 (-0.26, 0.38)	6.91
Yager	2010	0.71 (0.37, 1.04)	6.82
Overall (I-squared = 79.2%, p = 0.000)		0.75 (0.51, 0.99)	100.00

- Note: If group 1 (aka m1 s1 n1) was the depressed group, then you would see the values reported on the right of the middle line. Group 2 (aka m2 s2 n2) would therefore be the controls, and would be reported on the left of the middle line.

Any questions?