**THIS IS THE ACTIVE FILE**

Read and edit this in Microsoft Word

with "show hidden text= on"

(that setting is usually in "tools/options/view")

# \*STATA\_APST470\_APST670

Mariah Evans and Jonathan Kelley’s introductory notes for the applied statistics regression class APST 470/670 and for Stats Club Monday group.

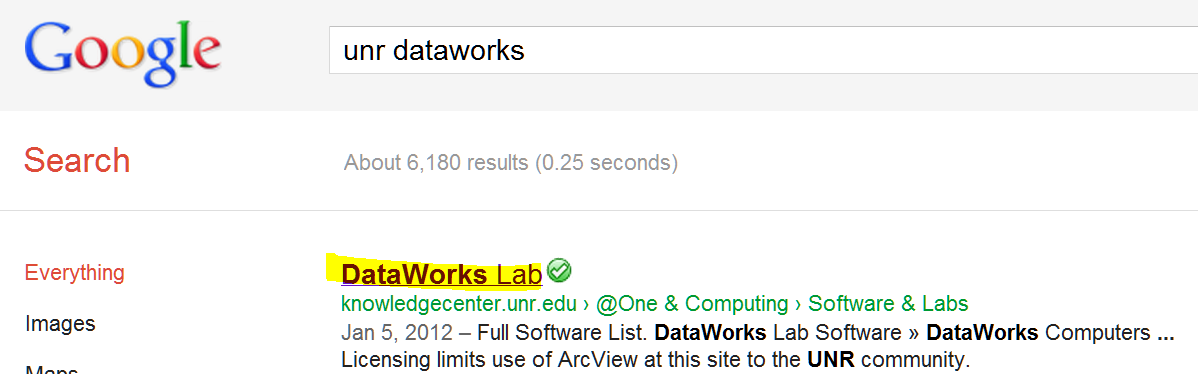
UNR and International Survey Center, February 2012

Version Week\_13\_dev\_a

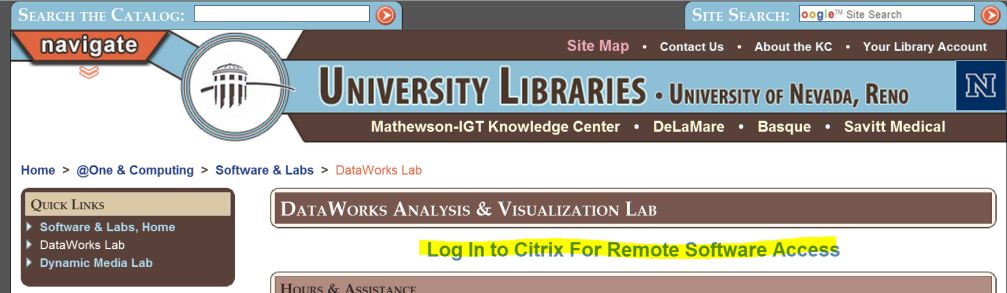
# Section #1. PRELIMINARIES: Get Stata running on UNR Citrix

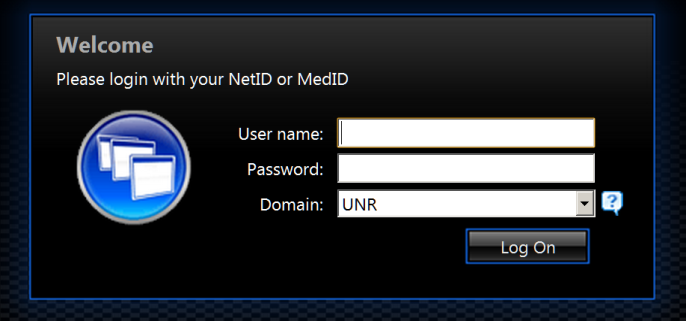
0. If you have a data stick/ thumb drive, stick it in now.

1. Log on to UNR DataWorks Citrix server from your browser. Start by Googling "unr dataworks" and choosing "DataWorks Lab":

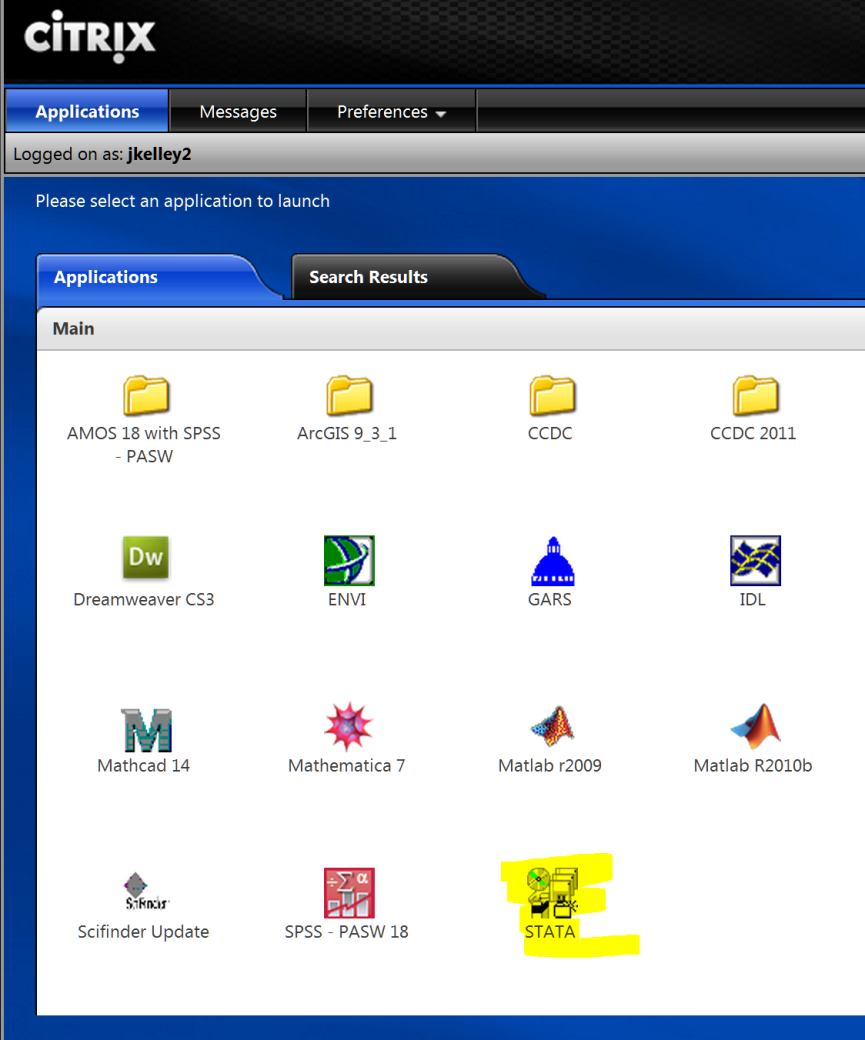


2. Then click through to Citrix and log yourself in with your UNR ID:

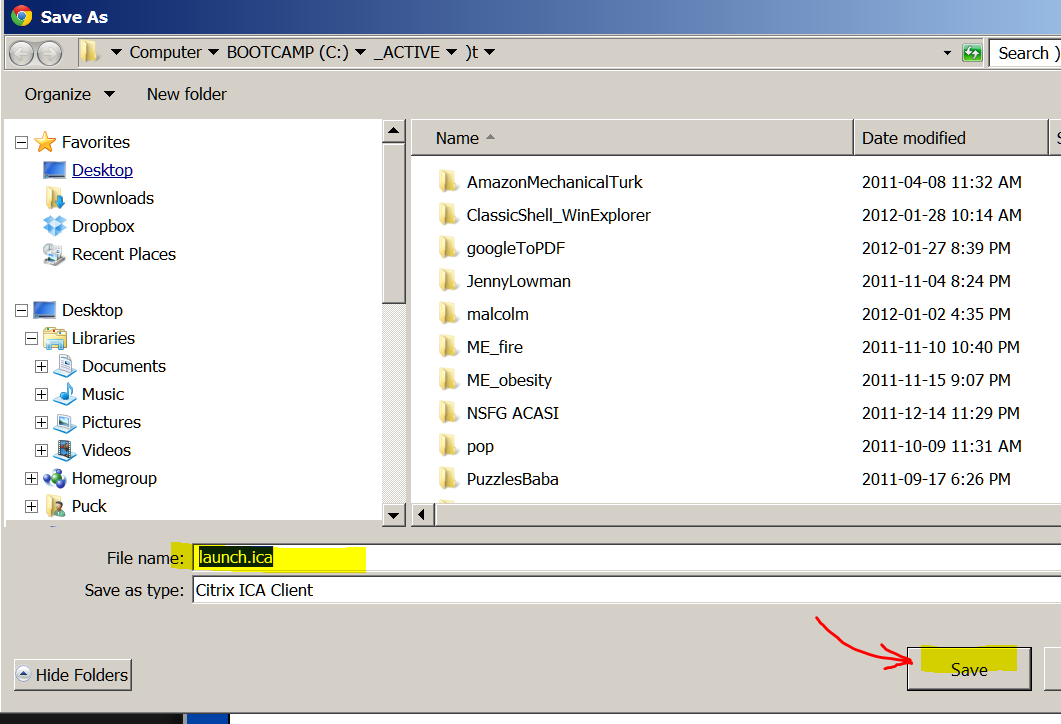




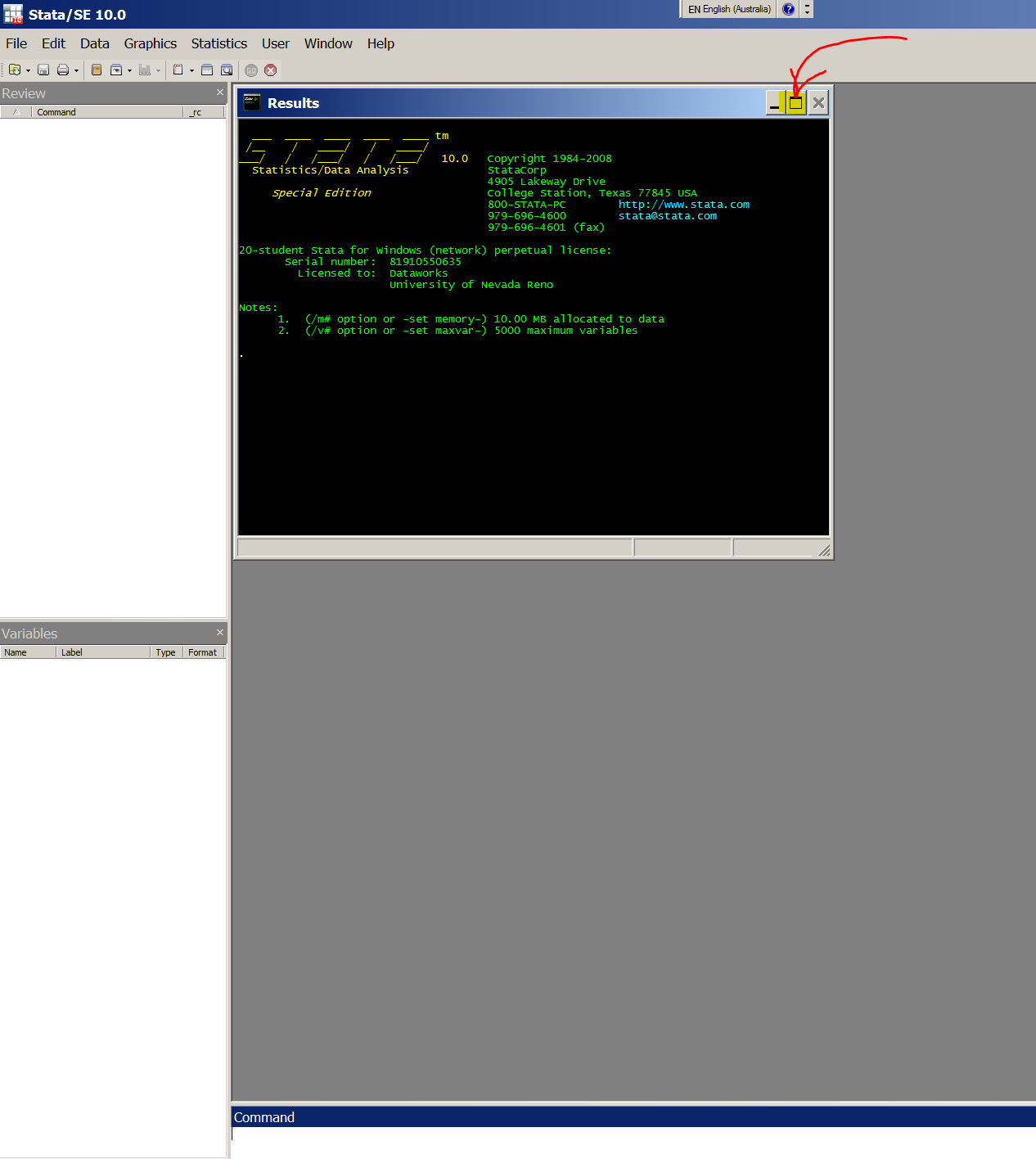
3. Find Stata and double click on it:



* Sometimes Citrix will launch Stata immediately. That is what you want, so wait patiently (Citrix is slow).
* Sometimes Citrix gets your browser to ask "Do you want to run this or to save it?" You want to run it
* Other times (annoyingly) it wants you to save a small shortcut file -- if so, double click on it to get Stata. (In my browser, Google Chrome, after you save, it appears in the lower left hand corner and you double click it directly from there. That works OK, but saving it to a file on your computer does not. Confusing!)



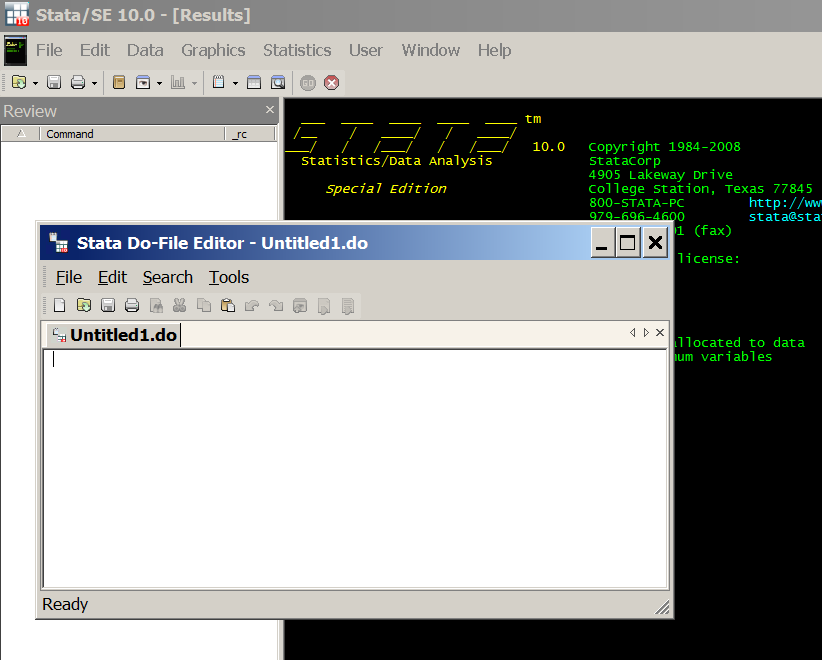
4. Stata then opens. Expand the "Results window" so you can see everything conveniently. Now you are ready to run Stata.



5. In Stata choose:

WINDOW/ DO FILE EDITOR / NEW DO-FILE

to open a "do file" window (so you can paste a whole batch of things in it to run at one go):



### A. For the first time you run things on Citrix you have to do things this way, afterwards another way.

6. Copy-and-paste into the Stata "do file" the following (just to set up Stata conveniently):

clear

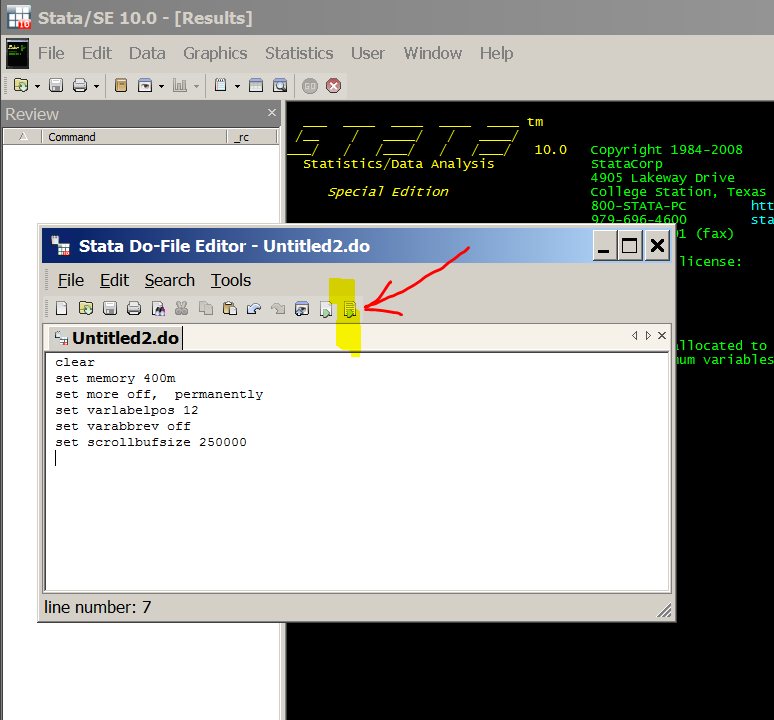
set memory 400m

set more off, permanently

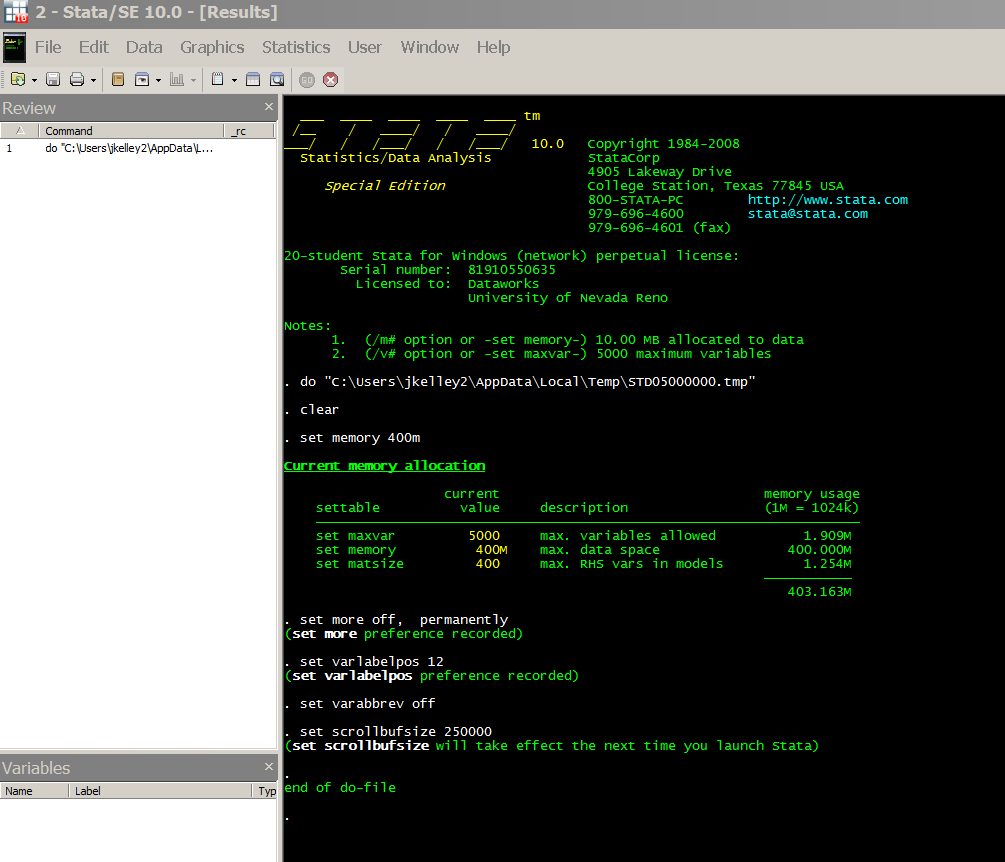
set varlabelpos 12

set varabbrev off

set scrollbufsize 250000



7. run (use the icon to the far right in the do file). You will see something like this: (It may be hiding behind the do-file window; if so, click on the main Stata window to get it in front of the do-file window) :



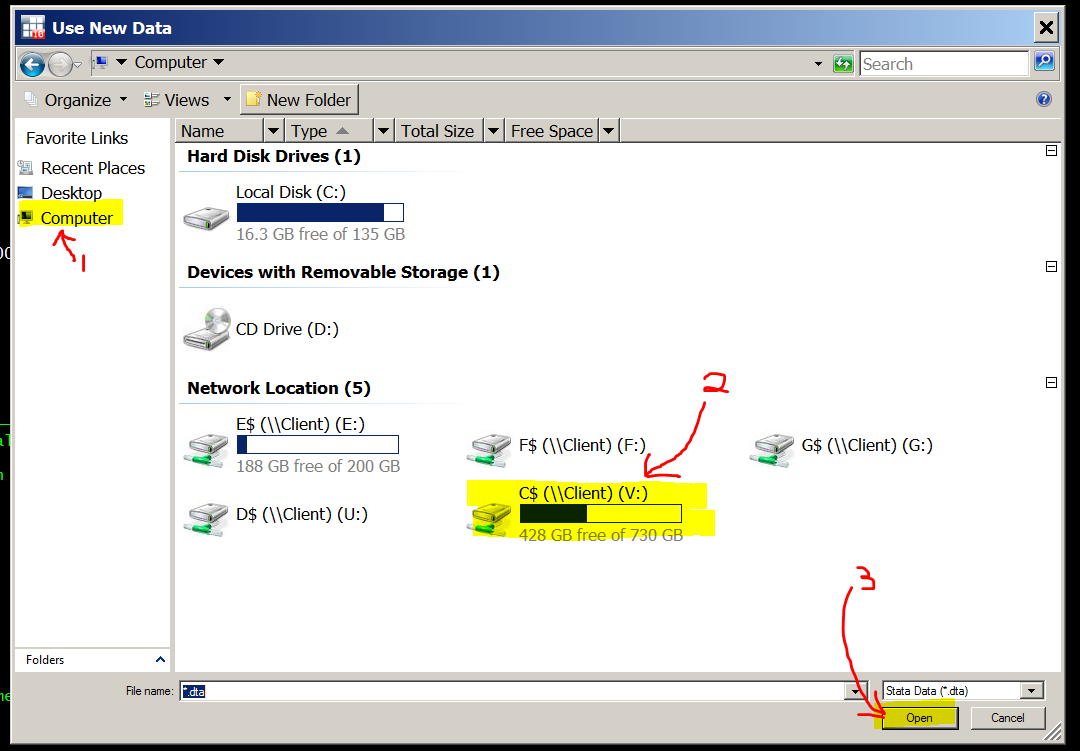
8. Now comes the hard part (but you will only have to do it the first time you run things on Citrix). You need to find the data. In Stata choose:

**FILE/ OPEN**

and navigate to YOUR computer (or data stick or wherever you have the data file).

#### If the file is on your computer (not a data stick) this will happen:

It is often drive "V" and you may have to click on three things in a row to get there (and perhaps expand the window to get enough room to see things):



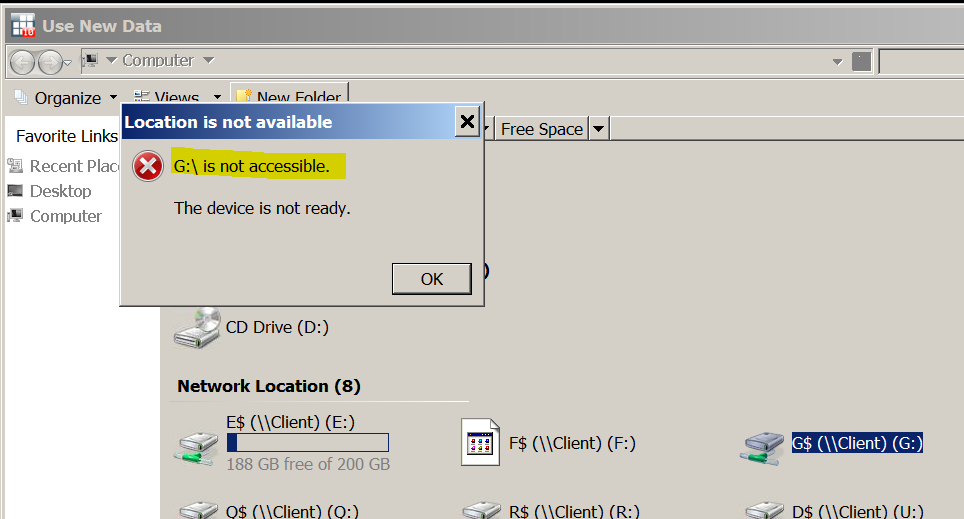
Then navigate to the data file -- it will be where you saved it on your computer. This is where it is on my machine, yours will be different. The file we are after is **IsssUSA\_AnalysisVars\_1.dta** . The ".dta" extension will probably not show up (it is there but the Citrix default is not to show it).

#### If you have the data file on a thumb drive, this is what will happen.

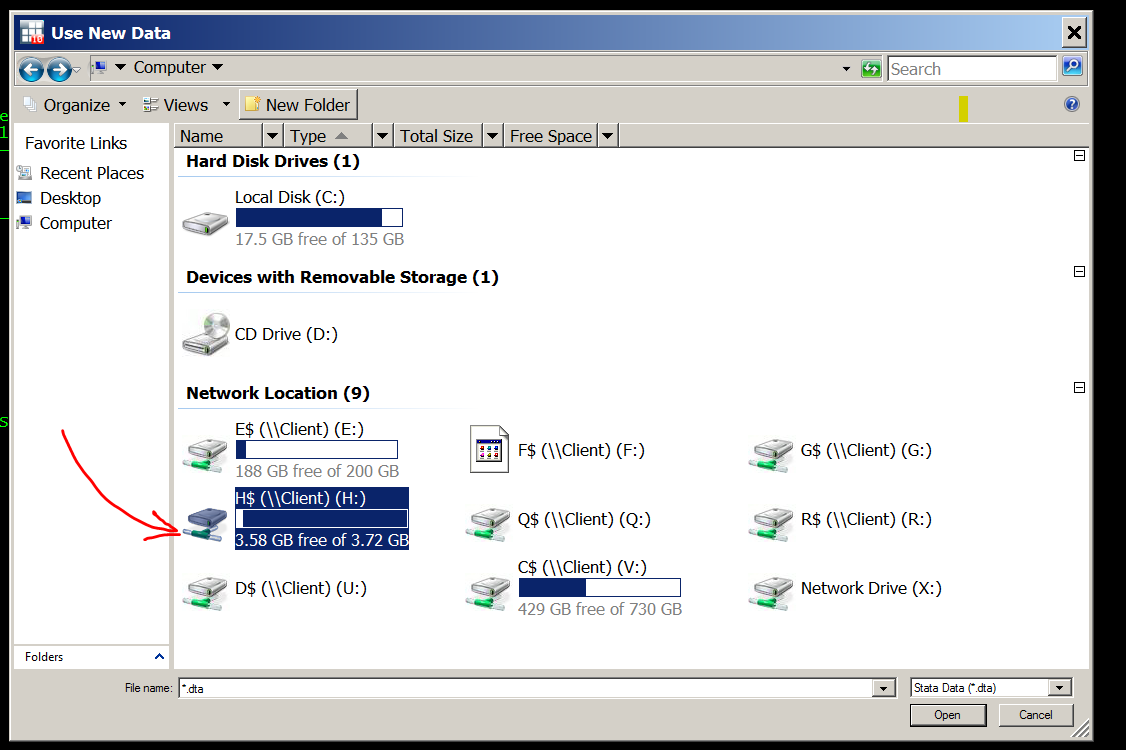
A thumb drive may be convenient for you, it may be a little harder to find.

A Citrix "got you" danger: You should plug in your thumb drive BEFORE opening Stata. If you do it afterwards, Stata may not see it (and you will get **very** frustrated looking for it!)

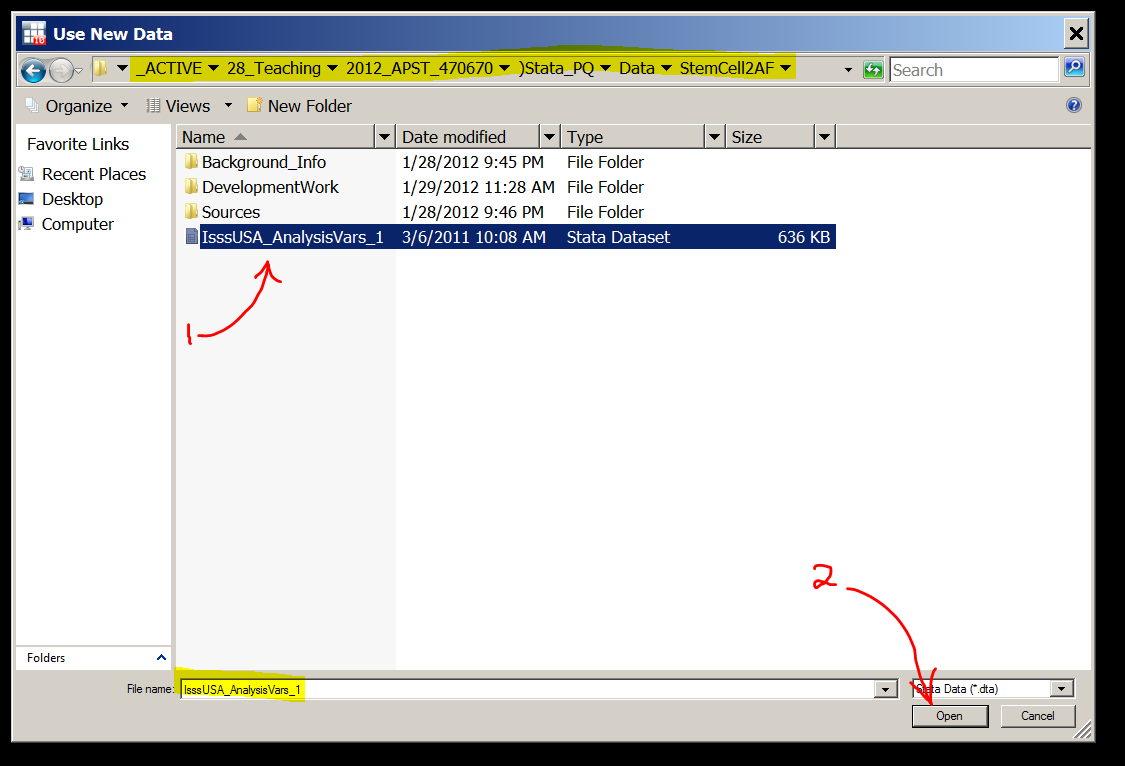
Even then, the thumb drive is hard to find in Citrix because there are lots of other obscure locations listed. Here is an example where I clicked through on the WRONG location, and was told so:



Keep looking until you find the data file. Here is my thumb drive:

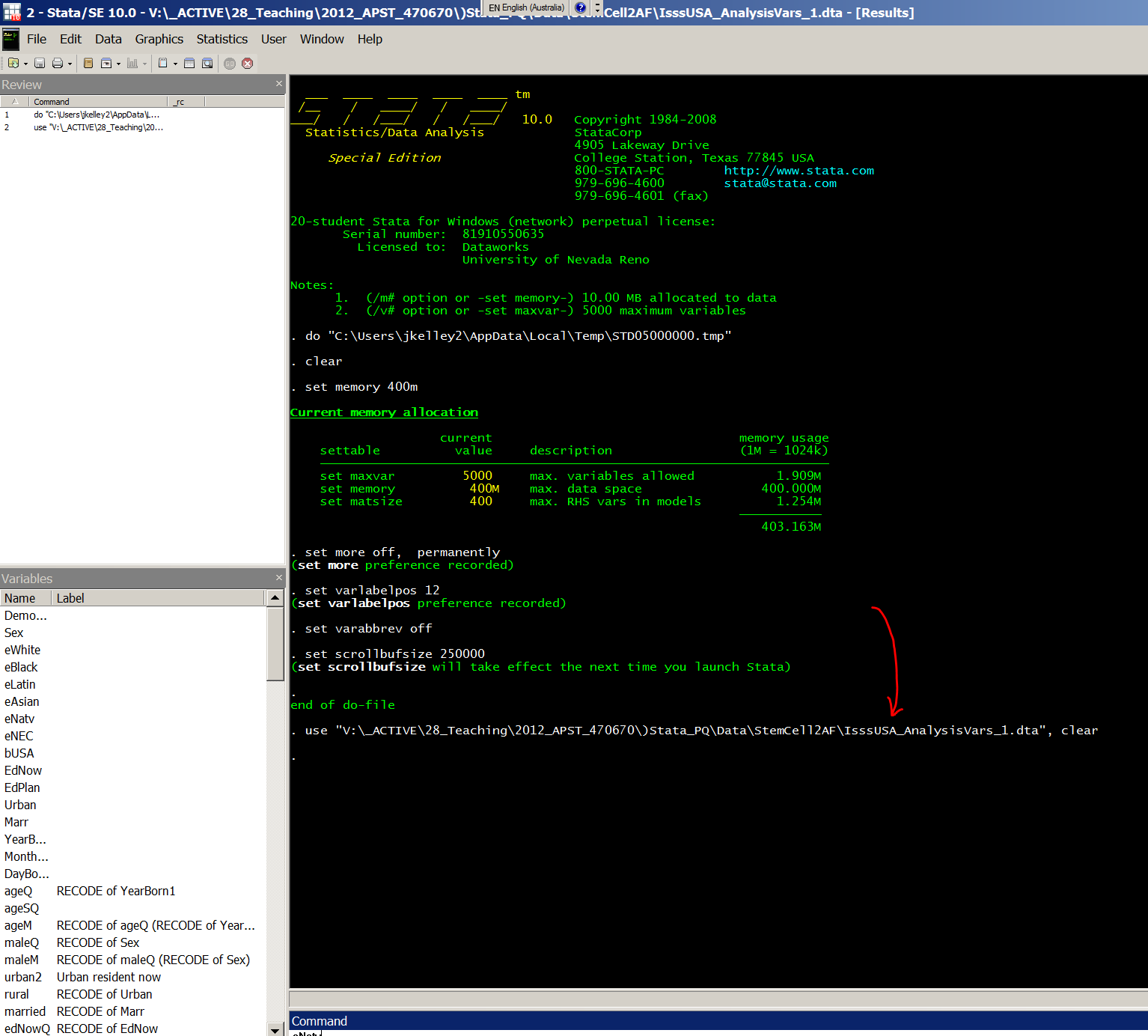


Open the file:



#### For both cases (thumb drive or file on your computer proper):

Stata will then get the data and (conveniently) show the command it used in the "Results" window. This is what things will look like:



9. Copy that line (sweeping it followed by Control-C will get it). Here is the line for my computer (which has a huge file structure, so a very long path):

**use "V:\\_ACTIVE\28\_Teaching\2012\_APST\_470670\)Stata\_PQ\Data\StemCell2AF\IsssUSA\_AnalysisVars\_1.dta", clear**

It may look a bit funny on the screen if there are more than 80 characters, as in my example just above. But that does not bother Stata.

For my thumb drive, it is much shorter:

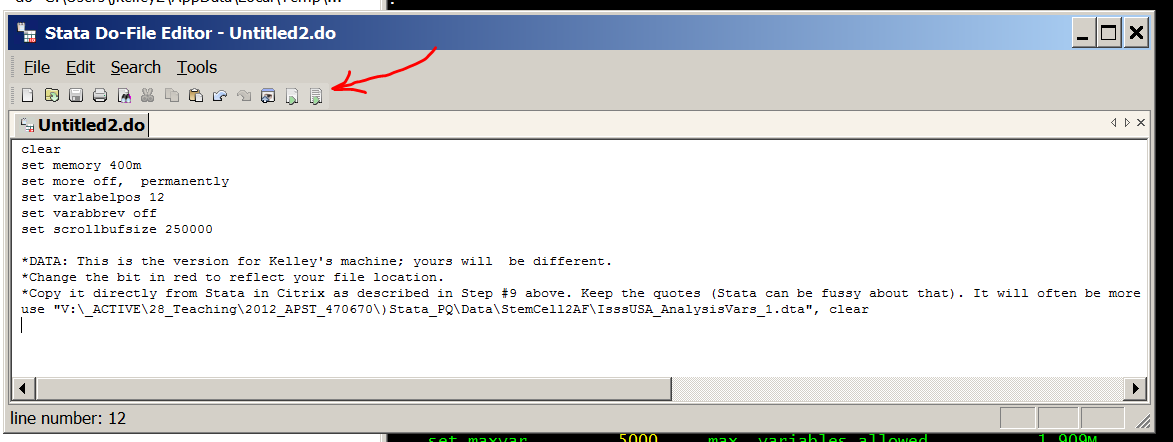
**use "H:\IsssUSA\_AnalysisVars\_1.dta", clear**

Then paste it into the text just below (**in Section #2 of THIS document**) for future use. Do it where it says:

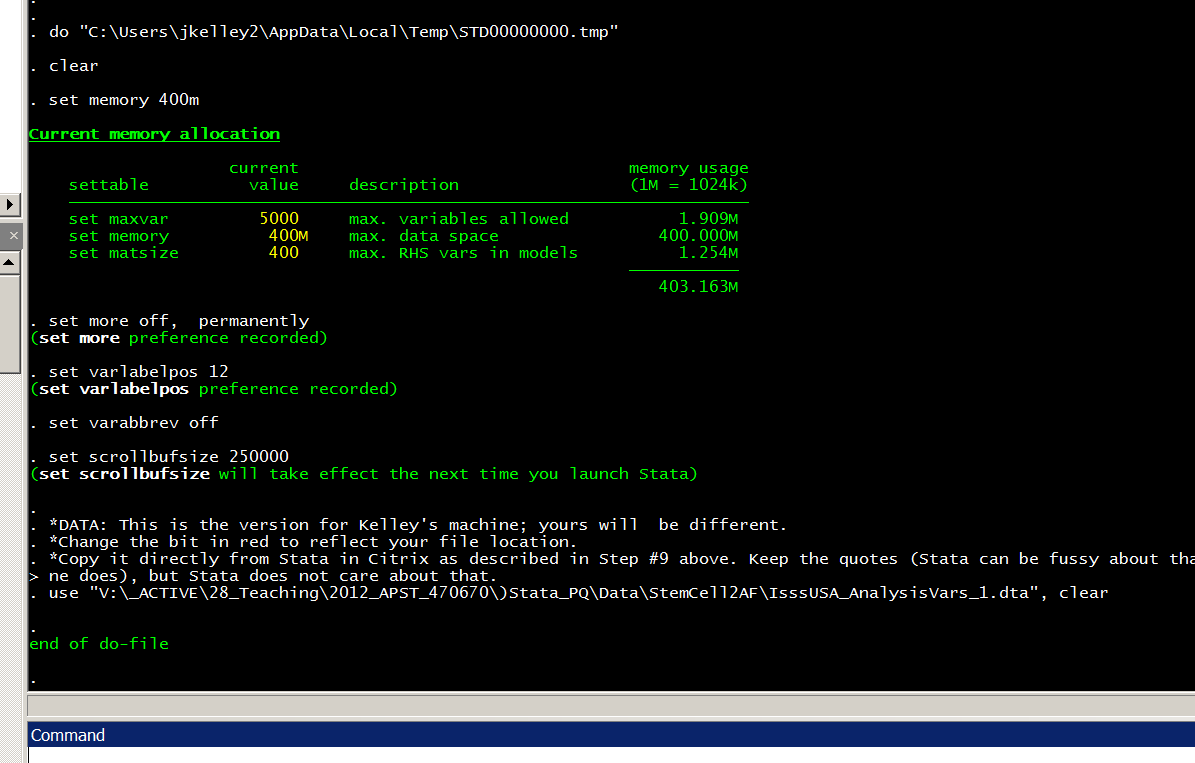
\*Change the bit in red to reflect your file location.

### B. AFTER the first time you run things on Citrix, it is much easier. Do it this way:

You will have pasted the file location on your computer into the text in Section #2 below in this file. Copy and paste all of Section #2 (in the grey box) into the Stata do file and run it:



This is what the Stata "Results" screen will look like.



All is well; you are up and running.

# \*Section #2: USUALLY BEGIN HERE: Standard Stata set up

NORMAL PROCEDURE

(1) Start Stata (see the notes above in Section #1)

(2) Copy and paste all of this (Section #2 in the grey box below) in a Stata do-file. Run it. (See Section #1 above, Part B, for notes on how to do that.) This will get your data file loaded -- often the hardest part!

(3) Then run everything in Section #3 (there won't be anything there for the first week or two, but afterwards there will be). How to do that:

Open a do-file

Erase anything that is in it already

Paste all of Section #3 in the do-file.

Run it.

That gets Stata up and running, with all the basic variables defined & ready to go. You must do this once at the beginning of your session -- you can not pick and choose, but need to run everything in Sections #2 and #3.

(4) Afterwards, run anything you want from Section #4, which has the week-by-week stuff. You can pick and choose here, running only the bits you want and ignoring the rest.

Do not worry about material in hidden text (usually in a blue box like this one) as it will not be copied.

\*NORMALLY BEGIN HERE & RUN EVERYTHING TO THE END OF SECTION #3

\*THEN YOU CAN CONTINUE AT WEEK #2, OR #3 OR #5, ...ETC...

clear

set memory 400m

set more off, permanently

set varlabelpos 12

set varabbrev off

set scrollbufsize 250000

\*DATA: This file location is on our home machine; yours will be different.

\*Change the bit in red to reflect your file location. Done&&&

\*Copy it directly from Stata in Citrix as described in Step #9 above. Keep the quotes (Stata can be fussy about that). It will often be more than 80 characters and look funny on the screen (as mine does), but Stata does not care about that.

\*On ME's machine

\*use "E:\IsssUSA\_AnalysisVars\_1.dta", clear

\*On JK's machine

use "C:\\_ACTIVE\28\_Teaching\2012\_APST\_470670\)Stata\_PQ\Data\StemCell2AF\IsssUSA\_AnalysisVars\_1.dta"

# \*Section #3: variable definitions & other preliminaries

### \*Occupation dummies

tab status

recode status (100=1)(10 / 90=0)(\*=.), gen( prof)

label variable prof "Professional occupation"

tab prof, m

table status, c(m prof freq) m

## \*Week 4 work here... (we came back after first working with previously prepared variables)

### \*Income

\*We included log income in the data file, but forgot income in dollars. So re-create it.

sum lnFamInc

tab lnFamInc

gen famInc = exp( lnFamInc)

replace famInc = round(famInc, 0.5)

label variable famInc "Family income ($1000s)"

tab famInc,m

famInc | Freq. Percent Cum.

------------+-----------------------------------

.5 | 284 12.37 12.37

5 | 44 1.92 14.29

10 | 43 1.87 16.17

15 | 57 2.48 18.65

20 | 94 4.10 22.75

25 | 76 3.31 26.06

30 | 77 3.36 29.41

35 | 95 4.14 33.55

40 | 73 3.18 36.73

45 | 83 3.62 40.35

50 | 90 3.92 44.27

55 | 52 2.27 46.54

60 | 72 3.14 49.67

65 | 62 2.70 52.37

70 | 56 2.44 54.81

75 | 71 3.09 57.91

80 | 57 2.48 60.39

90 | 59 2.57 62.96

100 | 81 3.53 66.49

125 | 73 3.18 69.67

150 | 50 2.18 71.85

200 | 20 0.87 72.72

250 | 12 0.52 73.25

450 | 14 0.61 73.86

. | 600 26.14 100.00

------------+-----------------------------------

Total | 2,295 100.00

Income always has a lot of missing data, but this is more than usual. I don't like it. But there is not much to do about it in this data set (the original data has individual income and employmet information, and something might be done there).

sum famInc

regress famInc status maleQ ageQ edNowQ edNowSq , beta

### \*Gender

\*First, look your variable up in the codebook to see the verbatim question

Sex Are you...

Male

Female

\*Get your frequency distribution using the tab command.

\*Always use the "m" option to get it to show missing data, if any exist

tab1 Sex, m

tab Sex, m

Sex | Freq. Percent Cum.

------------+-----------------------------------

-1 | 147 6.41 6.41

1 | 1,048 45.66 52.07

2 | 1,100 47.93 100.00

------------+-----------------------------------

Total | 2,295 100.00

\* Create a new variable called "female" from the raw variable "Sex".

\* The missing value in the raw variable is "-1", so you will recode that to "." which is STATA's missing data code: cases with . on a variable will be left out of calculations using that variable.

recode Sex (2=1)(1=0)(-1 = .), gen( female)

\*When you have made a new variable ALWAYS "tab" it, to check for errors.

tab female, m

\*If the frequency distribution on your new variable looks ok, next cross-tabulate it with the original variable as a second error check.

tab Sex female,m

. tab female, m

RECODE of |

Sex | Freq. Percent Cum.

------------+-----------------------------------

0 | 1,048 45.66 45.66

1 | 1,100 47.93 93.59

. | 147 6.41 100.00

------------+-----------------------------------

Total | 2,295 100.00

. tab Sex female,m

| RECODE of Sex

Sex | 0 1 . | Total

-----------+---------------------------------+----------

-1 | 0 0 147 | 147

1 | 1,048 0 0 | 1,048

2 | 0 1,100 0 | 1,100

-----------+---------------------------------+----------

Total | 1,048 1,100 147 | 2,295

### \*Religion: Interactions [from Week 3]

gen chXfund = fundmQ \* chGoNowQ

### \*Education: Quadratics [from week 7]

\*get rid of an old version, to avoid confusion

capture drop edNowSQ

gen edNowSq = edNowQ^2

### \*Moral authority

\*Next, we'll work on making some scales of trust in moral authorities. Prior research suggests that some people tend to turn to religious authorities, some to scientific/ medial authorities, and some prefer to judge for themselves. We are making these variables because we will later use regression to analyze their impact on attitudes towards biotechnologies using stem cells. The reason we make multiple-item scales/ indices is that, when this procedure is justified, it gives us better estimates because it reduces random measurement error, and hence brings us close to unbiased, consistent estimates.

\*First check out the exact wording or measurement procedure in the codebook.

In deciding whether it is right to allow these treatments, who would you believe...

Believe!!

Believe

??

Not

Not!!

mrlEthC a. A government Ethics Committee set up to decide what is right?

...

\*Use tab1 when want frequencies for several variables

\*Always use the "m" option to get it to show missing data, if any exist

tab1 mrlEthC mrlMD mrlYou1st mrlPope mrlABC mrlFund mrlNIH mrlYou, m

mrlEthC | Freq. Percent Cum.

------------+-----------------------------------

-1 | 94 4.10 4.10

1 | 246 10.72 14.81

2 | 507 22.09 36.91

3 | 746 32.51 69.41

4 | 339 14.77 84.18

5 | 363 15.82 100.00

------------+-----------------------------------

Total | 2,295 100.00

\*Here, strongly believing is coded 1 and strongly disbelieving is coded 5. Because it is going to be easier to talk/write about "believing increases" rather than "disbelieving decreases", we are going to recode this item to reverse the order of the scores. Moreover, readers will not have a strong intuitive grasp of the metric (1 to 5), so we will map these scores at equal intervals onto a 0 to 100 metric. Readers find 0 to 100 scoring easy to understand because it is similar to something they are very familiar with (%s).

\*When recoding a variable ALWAYS, ALWAYS make it a new variable with a new name. That let's you do the proper error check. And, sooner or later even after error checking you will be part way into your analysis and you will discover that you have made a coding error. If you have written over the top of your original variable it will be a mess. If you just have to correct your new variable, it will be easy. Do yourself a favor and make a habit!

\*Variable naming conventions: If you have a number of related questions, you will often find it is efficient to signal that to yourself and your co-workers with a prefix. In this case, these questions are all about moral authority, so we have used the prefix "mrl". If the original variable name is meaningful (i.e. not v003 etc), then you probably want to retain it inside the new name. Best practice is to keep the old name and add "R" at the end to signify that it is a recoded version. For example, the raw variable mrlEthC is recoded into the new variable mrlEthCR which is ready to use in your analysis. STATA allows long variable names, but try to keep them short, because if you use long ones, STATA does not display them properly in several important statistics.

recode mrlEthC (1=100)(2=75)(3=50)(4=25)(5=0)(-1 = .),gen( mrlEthCR)

\*Now do the standard checks for coding errors in your new variable. Is it ok?

tab mrlEthCR, m

tab mrlEthC mrlEthCR, m

\*Copy your results as tables and paste them in the blue box below. The box is blue, outlined with a dashed line, and the things inside it are in "hidden text". This is a convenient way to keep track of your foundational and intermediate results. It makes it easier to retrace your steps when you find errors and make it easy for people sharing a STATA file to see what their co-workers have done.

X

\*Let's see what it would look like if you accidentally put two original codes into one new code

recode mrlEthC (1=100)(2=75)(3=75)(4=25)(5=0)(-1 = .),gen( mrlEthCW)

tab mrlEthCW, m

tab mrlEthC mrlEthCW, m

\*We would hope to notice that there should be 50's in mrlEthCW but there aren't. Even if we missed that clue, how we we tell from the crosstabulation that something was wrong? Now let's wipe mrlEthCW from our memories & get back to doing things right. You won't need to see mrlEthCW again, so go up and star out those commands. Save your file with a new version key. You'll notice that the file has a rather long name. Someday you, too, will have over 200,000 files on your computer & you will need a good way to keep track of them. The "p" here is my version indicator. Make the next version which you are saving now "r".

\*----Now do the rest. As in building a house, the preparation stage takes a while, can be tedious, and is absolutely essential. Great care and precision are required here!

recode mrlMD mrlYou1st mrlPope mrlABC mrlFund mrlNIH mrlYou (1=100)(2=75)(3=50)(4=25)(5=0)(-1 = .),gen( mrlMDR mrlYou1stR mrlPopeR mrlABCR mrlFundR mrlNIHR mrlYouR)

tab1 mrlMDR mrlYou1stR mrlPopeR mrlABCR mrlFundR mrlNIHR mrlYouR, m

tab mrlMD mrlMDR ,m

\*All ok with mrlMDR?

\*Now you do the checks on mrlYou1stR. Because this is an example and I already checked you do not need to do the rest now. In your own data you need to do this with every variable you are going to use.

#### \*Means, frequencies: understanding points out of 100

\* First, look at the question wording or measurement details for you item and copy thm in here. I've done this one for you, but you SOP should always be to keep your measurement details inside the data preparation for that variable.

In deciding whether it is right to allow these treatments, who would you believe...

100 Believe!!

75 Believe

50 ??

25 Not

0 Not!!

mrlEthC a. A government Ethics Committee set up to decide what is right?

tab mrlMDR ,m

sum mrlMDR

RECODE of |

mrlMD | Freq. Percent Cum.

------------+-----------------------------------

0 | 155 6.75 6.75

25 | 151 6.58 13.33

50 | 687 29.93 43.27

75 | 797 34.73 78.00

100 | 411 17.91 95.90

. | 94 4.10 100.00

------------+-----------------------------------

Total | 2,295 100.00

.

. sum mrlMDR

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

mrlMDR | 2201 63.15311 27.19222 0 100

\* --- If you are going to show a table or graph of the frequency distribution, you usually present the %s with the missing data taken out. To get this table omit the usual "m" at the end of the "tab" command.

tab mrlMDR

\* Paste your output here. Make your own blue box.

\*--- Compare: which has more authority, in people's opinion?

sum mrlMDR mrlNIHR

\*--- All in all, who has most authority and who least?

sum mrlEthCR mrlMDR mrlYou1stR mrlPopeR mrlABCR mrlFundR mrlNIHR mrlYouR

#### \*Scale for religious authority

\*For a scale to be justified: (1) The items in it must have "face validity" -- there must be good reason to think that they all measure whatever the scale is supposed to represent; (2) the items must have high correlations among themselves ("inter-item correlations"); (3) the items in the scale must have similar correlations with important variables outside the scale ("Correlations with criterion variables").

\*Are they all correlated?

corr mrlPopeQ mrlABCQ mrlFundQ

\*Do they have similar correlations with criterion variables?

corr mrlPopeR mrlABCR mrlFundR chGoNowQ ageQ polDem9

corr mrlPopeR mrlABCR mrlFundR chGoNowQ ageQ polDem9

(obs=1989)

| mrlPopeR mrlABCR mrlFundR chGoNowQ ageQ polDem9

-------------+------------------------------------------------------

mrlPopeR | 1.0000

mrlABCR | 0.8187 1.0000

mrlFundR | 0.7693 0.8825 1.0000

chGoNowQ | 0.1456 0.1449 0.1957 1.0000

ageQ | -0.1065 -0.0782 -0.1099 0.1360 1.0000

polDem9 | -0.1250 -0.1167 -0.1600 -0.1890 -0.1031 1.0000

\*The important thing about the correlations with the criterion variables is that they should be similar for all the scale items. For example, all 3 candidate items for this scale have fairly small positive correlations with churchgoing (purple bar). The point is not whether these correlations are large or small, but that they are similar. How about the correlations with another criterion variable, age? Are the correlations of the candidate items with age similar? How about PolDem9?

\*To scale well, variables need to have similar standard deviations. If the standard deviations are similar in the variables as we have created them, then we can use them in scale construction as they are. If not, then we standardize before making the scale.

\*Similar standard deviations?

sum mrlPopeR mrlABCR mrlFundR

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

mrlPopeR | 2196 31.73953 30.76663 0 100

mrlABCR | 2194 33.80811 29.68837 0 100

mrlFundR | 2180 35.12615 30.87444 0 100

\* Standard deviations here are similar, but if yours are not, standardize the variables, then look at the correlations and factor analyses. To standardize your variables, remember APST 270 and z-score them: subtract the mean and then divide by the standard deviation.

\*Make a scale averaging the 3 items. Unweighted (as all have similar std devs)

\*Do this using a new command "egen"

egen mrlRel9 = rowmean( mrlPopeR mrlABCR mrlFundR )

tab mrlRel9

corr mrlRel9 mrlPopeR mrlABCR mrlFundR

mrlRel9 | Freq. Percent Cum.

------------+-----------------------------------

0 | 652 29.61 29.61

8.333333 | 25 1.14 30.74

16.66667 | 84 3.81 34.56

25 | 343 15.58 50.14

33.33333 | 94 4.27 54.41

37.5 | 3 0.14 54.54

41.66667 | 90 4.09 58.63

50 | 501 22.75 81.38

58.33333 | 74 3.36 84.74

62.5 | 2 0.09 84.83

66.66666 | 81 3.68 88.51

75 | 117 5.31 93.82

83.33334 | 33 1.50 95.32

91.66666 | 13 0.59 95.91

100 | 90 4.09 100.00

------------+-----------------------------------

Total | 2,202 100.00

.

. corr mrlRel9 mrlPopeR mrlABCR mrlFundR

(obs=2167)

| mrlRel9 mrlPopeR mrlABCR mrlFundR

-------------+------------------------------------

mrlRel9 | 1.0000

mrlPopeR | 0.9195 1.0000

mrlABCR | 0.9575 0.8192 1.0000

mrlFundR | 0.9414 0.7699 0.8825 1.0000

\*An important indicator of the quality of a scale is its "alpha reliability" (beware -- this is completely different from the alpha in hypothesis testing). This is also called the "Scale reliability coefficient". It ranges from 0 to 1. Over 0.7 is acceptable in most fields; over 0.8 is good; over 0.9 is excellent.

\*Reliability

alpha mrlPopeR mrlABCR mrlFundR, std

Test scale = mean(standardized items)

Average interitem correlation: 0.8234

Number of items in the scale: 3

Scale reliability coefficient: 0.9333

\*The average interitem correlation is the mean correlation between the items in your proposed scale. These should be high, especially if the scale has few items, as here.

\* When you are giving a presentation you should always report the scale alpha and the mean inter-item correlation. For a report or article, you should always present them AND the detailed inter-item correlations and the correlations with criterion variable (if you want to keep the article or report short, you put the correlations in an appendix).

### \*Factor analysis

\*If all is looking well thus far, you are ready to assess systematically the quality of your potential scale.

#### \*Preliminaries

\*correlations

corr mrlEthCR mrlMDR mrlYou1stR mrlPopeR mrlABCR mrlFundR mrlNIHR mrlYouR

\*This was the order in which we got them from the dataset, but notice that they will be much easier to read if we rearrange them so that the candidate items for each of the three scale are groups together.

\*Re-arrange, so easier to read

corr mrlPopeR mrlABCR mrlFundR mrlEthCR mrlMDR mrlNIHR mrlYou1stR mrlYouR

| mrlPopeR mrlABCR mrlFundR mrlEthCR mrlMDR mrlNIHR mrlYo~tR mrlYouR

-------------+------------------------------------------------------------------------

mrlPopeR | 1.0000

mrlABCR | 0.8192 1.0000

mrlFundR | 0.7679 0.8836 1.0000

mrlEthCR | 0.2822 0.3203 0.2706 1.0000

mrlMDR | 0.1384 0.1884 0.1225 0.5476 1.0000

mrlNIHR | 0.1220 0.1851 0.1466 0.5239 0.6620 1.0000

mrlYou1stR | 0.1491 0.1735 0.1773 0.2405 0.3940 0.3361 1.0000

mrlYouR | 0.0844 0.1040 0.1278 0.1941 0.3507 0.3517 0.7221 1.0000

#### \*Factor analysis proper: One factor, the simplest case

\*We are looking for Factor1 having an "eigenvalue" over 1, and nothing else coming close.

factor mrlPopeR mrlABCR mrlFundR

Factor analysis/correlation Number of obs = 2167

Method: principal factors Retained factors = 1

Rotation: (unrotated) Number of params = 3

--------------------------------------------------------------------------

Factor | Eigenvalue Difference Proportion Cumulative

-------------+------------------------------------------------------------

Factor1 | 2.41679 2.45500 1.0539 1.0539

Factor2 | -0.03821 0.04709 -0.0167 1.0372

Factor3 | -0.08530 . -0.0372 1.0000

--------------------------------------------------------------------------

LR test: independent vs. saturated: chi2(3) = 5740.60 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

---------------------------------------

Variable | Factor1 | Uniqueness

-------------+----------+--------------

mrlPopeR | 0.8451 | 0.2859

mrlABCR | 0.9385 | 0.1192

mrlFundR | 0.9066 | 0.1782

---------------------------------------

\*Note that all factor loadings are high

#### \*Factor analysis: scale plus an item that does not belong

\*---Preliminaries: the new item looks bad

corr mrlPopeR mrlABCR mrlFundR ageQ

| mrlPopeR mrlABCR mrlFundR ageQ

-------------+------------------------------------

mrlPopeR | 1.0000

mrlABCR | 0.8208 1.0000

mrlFundR | 0.7742 0.8842 1.0000

ageQ | -0.1132 -0.0899 -0.1204 1.0000

\*---Factor analysis shows its bad also

factor mrlPopeR mrlABCR mrlFundR ageQ

Factor analysis/correlation Number of obs = 2078

Method: principal factors Retained factors = 2

Rotation: (unrotated) Number of params = 6

--------------------------------------------------------------------------

Factor | Eigenvalue Difference Proportion Cumulative

-------------+------------------------------------------------------------

Factor1 | 2.43968 2.42712 1.0502 1.0502

Factor2 | 0.01256 0.05063 0.0054 1.0556

Factor3 | -0.03806 0.05311 -0.0164 1.0392

Factor4 | -0.09117 . -0.0392 1.0000

--------------------------------------------------------------------------

LR test: independent vs. saturated: chi2(6) = 5593.63 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

-------------------------------------------------

Variable | Factor1 Factor2 | Uniqueness

-------------+--------------------+--------------

mrlPopeR | 0.8479 -0.0110 | 0.2810

mrlABCR | 0.9378 0.0301 | 0.1197

mrlFundR | 0.9094 -0.0067 | 0.1729

ageQ | -0.1197 0.1072 | 0.9742

-------------------------------------------------

\*The factor loading for the item that does not belong is very small.

#### \*Factor analysis revealing several concepts

\*The process is very similar to when we are testing for one factor (factors are also sometimes called "dimensions", "Concepts" or "true variables").

\*We have already seen the correlations (above)

\*----- Also need to look at correlations with criterion variables

corr mrlPopeR mrlABCR mrlFundR mrlEthCR mrlMDR mrlNIHR mrlYou1stR mrlYouR maleQ ageQ edNowQ polDem9 ChGoNow

| mrlPopeR mrlABCR mrlFundR mrlEthCR mrlMDR mrlNIHR mrlYo~tR mrlYouR

-------------+------------------------------------------------------------------------

mrlPopeR | 1.0000

mrlABCR | 0.8209 1.0000

mrlFundR | 0.7733 0.8857 1.0000

mrlEthCR | 0.2671 0.3023 0.2503 1.0000

mrlMDR | 0.1220 0.1641 0.0963 0.5284 1.0000

mrlNIHR | 0.1023 0.1621 0.1203 0.5050 0.6444 1.0000

mrlYou1stR | 0.1358 0.1637 0.1733 0.2222 0.3765 0.3204 1.0000

mrlYouR | 0.0721 0.0922 0.1156 0.1717 0.3313 0.3268 0.7136 1.0000

maleQ | 0.0691 0.0572 0.0575 -0.0177 0.0410 -0.0163 -0.0469 -0.0337

ageQ | -0.1120 -0.0906 -0.1162 -0.1234 -0.0238 -0.0558 -0.0778 -0.0723

edNowQ | -0.0381 -0.0318 -0.0585 -0.0070 0.0036 0.0166 0.0336 0.0521

polDem9 | -0.1222 -0.1137 -0.1549 0.2219 0.1929 0.1907 0.0443 0.0480

chGoNowQ | 0.1451 0.1406 0.1927 -0.0703 -0.1712 -0.1554 -0.1082 -0.1051

cathQ | 0.3228 0.1953 0.1486 0.0314 0.0069 0.0265 -0.0200 -0.0126

\*Not perfect, but the balance of the evidence is that the correlations with criterion variable support the hypothesis that the three posited scales exist.

\*---factor analysis

factor mrlPopeR mrlABCR mrlFundR mrlEthCR mrlMDR mrlNIHR mrlYou1stR mrlYouR

actor analysis/correlation Number of obs = 2135

Method: principal factors Retained factors = 3

Rotation: (unrotated) Number of params = 21

--------------------------------------------------------------------------

Factor | Eigenvalue Difference Proportion Cumulative

-------------+------------------------------------------------------------

Factor1 | 3.01094 1.31448 0.6255 0.6255

Factor2 | 1.69647 1.03499 0.3524 0.9779

Factor3 | 0.66148 0.68745 0.1374 1.1153

Factor4 | -0.02597 0.05307 -0.0054 1.1099

Factor5 | -0.07904 0.02888 -0.0164 1.0935

Factor6 | -0.10792 0.03539 -0.0224 1.0711

Factor7 | -0.14331 0.05541 -0.0298 1.0413

Factor8 | -0.19872 . -0.0413 1.0000

--------------------------------------------------------------------------

LR test: independent vs. saturated: chi2(28) = 1.0e+04 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

-----------------------------------------------------------

Variable | Factor1 Factor2 Factor3 | Uniqueness

-------------+------------------------------+--------------

mrlPopeR | 0.7065 -0.4632 0.0260 | 0.2856

mrlABCR | 0.8055 -0.4882 0.0061 | 0.1128

mrlFundR | 0.7599 -0.4909 0.0819 | 0.1749

mrlEthCR | 0.5466 0.2408 -0.3334 | 0.5321

mrlMDR | 0.5360 0.5036 -0.2854 | 0.3776

mrlNIHR | 0.5188 0.4759 -0.2917 | 0.4192

mrlYou1stR | 0.4918 0.4663 0.4263 | 0.3590

mrlYouR | 0.4339 0.4970 0.4412 | 0.3701

-----------------------------------------------------------

\*---That was ugly. We need to rotate. Rotation is the extra step we need to find several factors.

rotate

Factor analysis/correlation Number of obs = 2135

Method: principal factors Retained factors = 3

Rotation: orthogonal varimax (Kaiser off) Number of params = 21

--------------------------------------------------------------------------

Factor | Variance Difference Proportion Cumulative

-------------+------------------------------------------------------------

Factor1 | 2.48672 0.92418 0.5166 0.5166

Factor2 | 1.56253 0.24289 0.3246 0.8412

Factor3 | 1.31964 . 0.2741 1.1153

--------------------------------------------------------------------------

LR test: independent vs. saturated: chi2(28) = 1.0e+04 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

-----------------------------------------------------------

Variable | Factor1 Factor2 Factor3 | Uniqueness

-------------+------------------------------+--------------

mrlPopeR | 0.8408 0.0769 0.0387 | 0.2856

mrlABCR | 0.9323 0.1255 0.0484 | 0.1128

mrlFundR | 0.9029 0.0508 0.0857 | 0.1749

mrlEthCR | 0.2545 0.6279 0.0944 | 0.5321

mrlMDR | 0.0844 0.7366 0.2696 | 0.3776

mrlNIHR | 0.0880 0.7169 0.2431 | 0.4192

mrlYou1stR | 0.1145 0.2198 0.7613 | 0.3590

mrlYouR | 0.0510 0.1981 0.7669 | 0.3701

-----------------------------------------------------------

\*Looking good! 3 eignvalues >1 indicates 3 factors, as we expect. Loadings for the variables that are supposed to measure each factor/ underlying variable have high loadings on that factor and low loadings elsewhere.

\*----- Make scales for the other two concepts

egen mrlMed9 = rowmean( mrlEthCR mrlMDR mrlNIHR)

tab mrlMed9, m

corr mrlMed9 mrlEthCR mrlMDR mrlNIHR

egen mrlYou9 = rowmean( mrlYou1stR mrlYouR )

tab mrlYou9

corr mrlYou9 mrlYou1stR mrlYouR

\*------ Correlations with other variables: a 2nd and better look

corr mrlRel9 mrlMed9 mrlYou9 maleQ ageQ edNowQ polDem9 chGoNowQ, m

Variable | Mean Std. Dev. Min Max

-------------+----------------------------------------------------

mrlRel9 | 33.61376 28.66974 0 100

mrlMed9 | 58.86834 23.71164 0 100

mrlYou9 | 71.03819 23.65904 0 100

maleQ | .4868735 .499947 0 1

ageQ | 46.37375 15.98017 17 87

edNowQ | 14.08329 2.256712 10 22

polDem9 | .5340119 .3304676 0 1

ChGoNow | 3.55895 2.807283 -1 9

| mrlRel9 mrlMed9 mrlYou9

-------------+-----------------------------

mrlRel9 | 1.0000

mrlMed9 | 0.2260 1.0000

mrlYou9 | 0.1449 0.3681 1.0000

maleQ | 0.0678 0.0045 -0.0428

ageQ | -0.1108 -0.0873 -0.0831

edNowQ | -0.0530 0.0046 0.0465

polDem9 | -0.1419 0.2401 0.0491

ChGoNow | 0.1984 -0.1576 -0.1058

#### \*Another factor analysis: Your data

\*[ do two examples with class data]

\*If we have time, we'll do the theory today, but we need time to work on your projects, so we may turn to your projects at this point and do the theory later.

### \*Factor analysis: theory

#### \*Random noise

gen e1 = invnorm(uniform())

gen e2 = invnorm(uniform())

gen e3 = invnorm(uniform())

gen e4 = invnorm(uniform())

corr e1 e2 e3 e4, m

(obs=2295)

Variable | Mean Std. Dev. Min Max

-------------+----------------------------------------------------

e1 | -.0211384 1.027497 -3.562193 2.971157

e2 | .0247844 .9880224 -3.723434 3.47072

e3 | -.0147948 .9930991 -3.19455 3.648272

e4 | -.0048265 .9851684 -3.365304 3.747546

| e1 e2 e3 e4

-------------+------------------------------------

e1 | 1.0000

e2 | -0.0032 1.0000

e3 | 0.0029 -0.0069 1.0000

e4 | 0.0145 -0.0207 0.0035 1.0000

#### \*Item = true score plus random noise

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

polDem9 | 2295 .5320283 .3210737 0 1

gen polQ1 = polDem9 + .30\*e1

gen polQ2 = polDem9 + .35\*e2

gen polQ3 = polDem9 + .25\*e3

gen polQ4 = polDem9 + .30\*e4

#### \*Shows our "usual" pattern

corr polQ1 polQ2 polQ3 polQ4 cThrp9 mrlRel9 mrlMed9 mrlYou9 edNowQ maleQ ageQ

(obs=2094)

| polQ1 polQ2 polQ3 polQ4

-------------+--------------------------------------

polQ1 | 1.0000

polQ2 | 0.5171 1.0000

polQ3 | 0.5928 0.5545 1.0000

polQ4 | 0.5514 0.5011 0.5883 1.0000

cThrp9 | 0.2120 0.2035 0.2289 0.2194

mrlRel9 | -0.1269 -0.0991 -0.1175 -0.1342

mrlMed9 | 0.1818 0.1644 0.1976 0.1794

mrlYou9 | 0.0439 0.0444 0.0327 -0.0028

edNowQ | -0.0201 -0.0201 -0.0043 -0.0462

maleQ | -0.0124 -0.0380 -0.0764 -0.0486

ageQ | -0.0688 -0.0808 -0.0953 -0.0727

#### \*Factor analysis

factor polQ1 polQ2 polQ3 polQ4

(obs=2295)

Factor analysis/correlation Number of obs = 2295

Method: principal factors Retained factors = 1

Rotation: (unrotated) Number of params = 4

--------------------------------------------------------------------------

Factor | Eigenvalue Difference Proportion Cumulative

-------------+------------------------------------------------------------

Factor1 | 2.04050 2.13246 1.2071 1.2071

Factor2 | -0.09196 0.02692 -0.0544 1.1527

Factor3 | -0.11889 0.02029 -0.0703 1.0823

Factor4 | -0.13917 . -0.0823 1.0000

--------------------------------------------------------------------------

LR test: independent vs. saturated: chi2(6) = 3155.56 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

---------------------------------------

Variable | Factor1 | Uniqueness

-------------+----------+--------------

polQ1 | 0.7208 | 0.4805

polQ2 | 0.6621 | 0.5616

polQ3 | 0.7606 | 0.4215

polQ4 | 0.7100 | 0.4959

---------------------------------------

### \*For program Eval4: Subroutine fix\_vars

\*Eval4 is for getting predicted values in complex models.

\*You always need a version of subroutine fix\_vars to keep Eval4 happy, but it can be empty (almmost).

\*If you have interactions, quadratics, etc put them in fix\_vars.

\*Extra things in fix\_vars are harmless. Nor does it matter if you are not actually using interactions and quadratics so long as fix\_vars exists.

\*When you do make interactions, quadratics, etc, just add them to fix\_vars and re-run it -- it will overwrite older versions automatically.

capture program drop fix\_vars /\*Drop previous version, so can replace \*/

program define fix\_vars

/\* compute interactions & quadratics: "quietly replace" \*/

quietly display " "

quietly replace edNowSq = edNowQ^2

/\*interactions, as previously defined but here with "quietly replace" not "gen" \*/

quietly replace chXfund = fundmQ \* chGoNowQ

end

# \*Section #4: ALWAYS RUN UP TO HERE: \*Analysis (week-by-week material)

## \*Week 2 material

### \*Regression with a single dichotomous variable

tab1 maleQ cIPS9

-> tabulation of maleQ

RECODE of |

Sex | Freq. Percent Cum.

------------+-----------------------------------

0 | 1,100 51.21 51.21

1 | 1,048 48.79 100.00

------------+-----------------------------------

Total | 2,148 100.00

-> tabulation of cIPS9

cIPS9 | Freq. Percent Cum.

------------+-----------------------------------

0 | 147 6.59 6.59

8.333333 | 7 0.31 6.90

16.66667 | 12 0.54 7.44

25 | 79 3.54 10.98

33.33333 | 50 2.24 13.22

41.66667 | 67 3.00 16.22

50 | 378 16.94 33.15

58.33333 | 135 6.05 39.20

62.5 | 1 0.04 39.25

66.66666 | 215 9.63 48.88

75 | 376 16.85 65.73

83.33334 | 148 6.63 72.36

87.5 | 2 0.09 72.45

91.66666 | 155 6.94 79.39

100 | 460 20.61 100.00

------------+-----------------------------------

Total | 2,232 100.00

table maleQ , c( m cIPS9 freq)

------------------------------------

RECODE of |

Sex | mean(cIPS9) Freq.

----------+-------------------------

0 | 65.01744 1,100

1 | 69.38077 1,048

------------------------------------

disp 69.38 - 65.017

4.363

regress cIPS9 maleQ

Source | SS df MS Number of obs = 2146

-------------+------------------------------ F( 1, 2144) = 13.57

Model | 10208.2396 1 10208.2396 Prob > F = 0.0002

Residual | 1612475.28 2144 752.08735 R-squared = 0.0063

-------------+------------------------------ Adj R-squared = 0.0058

Total | 1622683.52 2145 756.495813 Root MSE = 27.424

------------------------------------------------------------------------------

cIPS9 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

maleQ | 4.36333 1.184341 3.68 0.000 2.040753 6.685908

\_cons | 65.01744 .8272472 78.59 0.000 63.39515 66.63973

------------------------------------------------------------------------------

ttest cIPS9, by(maleQ)

Two-sample t test with equal variances

------------------------------------------------------------------------------

Group | Obs Mean Std. Err. Std. Dev. [95% Conf. Interval]

---------+--------------------------------------------------------------------

0 | 1099 65.01744 .8484643 28.12758 63.35265 66.68223

1 | 1047 69.38077 .8241062 26.66592 67.76368 70.99786

---------+--------------------------------------------------------------------

combined | 2146 67.14624 .5937292 27.50447 65.9819 68.31059

---------+--------------------------------------------------------------------

diff | -4.36333 1.184341 -6.685908 -2.040753

------------------------------------------------------------------------------

diff = mean(0) - mean(1) t = -3.6842

Ho: diff = 0 degrees of freedom = 2144

Ha: diff < 0 Ha: diff != 0 Ha: diff > 0

Pr(T < t) = 0.0001 Pr(|T| > |t|) = 0.0002 Pr(T > t) = 0.9999

### \*Regression with a single continuous variable

table edNowQ, c(m cIPS9 freq)

------------------------------------

RECODE of |

EdNow | mean(cIPS9) Freq.

----------+-------------------------

10 | 57.93651 42

11 | 74.46236 31

12 | 63.03612 444

13 | 65.65501 729

14 | 68.50394 255

16.5 | 70.70428 453

18 | 73.44377 166

22 | 69.04762 35

------------------------------------

regress cIPS9 edNowQ

Source | SS df MS Number of obs = 2152

-------------+------------------------------ F( 1, 2150) = 24.61

Model | 18409.8391 1 18409.8391 Prob > F = 0.0000

Residual | 1608302.83 2150 748.047829 R-squared = 0.0113

-------------+------------------------------ Adj R-squared = 0.0109

Total | 1626712.67 2151 756.258797 Root MSE = 27.35

------------------------------------------------------------------------------

cIPS9 | Coef. Std. Err. t P>|t| [95% Conf. Interval]

-------------+----------------------------------------------------------------

edNowQ | 1.284678 .2589609 4.96 0.000 .7768386 1.792518

\_cons | 49.04121 3.696598 13.27 0.000 41.79192 56.29049

------------------------------------------------------------------------------

## \*Week 3 material

\* In Week 2 you learned that regression with a single dummy/dichotomous independent variable X1 (in our case “maleQ”) gives you the exact sample difference of means on the dependent variable (in our case, attitudes towards towards using stem cells derived from adult tissue in research and treatment, “cIPS9”): The predicted value on the dependent variable when X1=0 (=female in the example) is the sample mean of cIPS9 for women. The predicted value on the dependent variable when X1=1 (=male in the example) is the sample mean of cIPS9 for men.

\* Let’s start by replicating that analysis for another dichotomous variable, membership in a fundamentalist denomination. Then we will proceed to regression analysis with two predictors.

\* The new predictor here is fundmQ.

\* Always begin by investigating the frequency distribution of your variables.

### \*Regression with a single dichotomous predictor, review

\*Use the tab command to get your frequency distributions. Check for stray codes.

tab1 fundmQ cIPS9

\*Now see how the means on cIPS9 differ between people belonging to fundamentalist denominations and others.

table fundmQ , c( m cIPS9 freq)

\*Use the disp command to calculate the sample difference of means (e.g.disp 10-5)

\*Find the difference of cIPS9 means in the regression coefficient for fundmQ

regress cIPS9 fundmQ , b

Source | SS df MS Number of obs = 2232

-------------+------------------------------ F( 1, 2230) = 59.02

Model | 44661.148 1 44661.148 Prob > F = 0.0000

Residual | 1687543.52 2230 756.745971 R-squared = 0.0258

-------------+------------------------------ Adj R-squared = 0.0253

Total | 1732204.66 2231 776.425219 Root MSE = 27.509

------------------------------------------------------------------------------

cIPS9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -9.80198 1.275921 -7.68 0.000 -.1605703

\_cons | 69.55577 .6938232 100.25 0.000 .

------------------------------------------------------------------------------

### \*Regression with a single quantitative predictor, review

\*Now let’s do a single quantitative predictor: Church attendance.

tab1 chGoNowQ

\*Look at the means of IPS for different values of church attendance.

table chGoNowQ , c( m cIPS9 freq)

\*Now regress cIPS9 on chGoNowQ. What does the regression coefficient on chGoNowQ tell you?

regress cIPS9 chGoNowQ, b

### \*Regression with two predictors

\*What happens when we include both predictors?

regress cIPS9 fundmQ chGoNowQ, b

### \*Regression with an interaction term

\*Does churchgoing have the same effect on iPS attitudes for respondents belonging to fundamentalist denominations and others? To find out we run the regression separately for fundamentalists and for others:

regress cIPS9 chGoNowQ if fundmQ==1 , b

regress cIPS9 chGoNowQ if fundmQ==0 , b

\*Alternatively, we can make an “interaction term” and then enter it into the regression model. This gives a significance test on the difference. It is also more flexible in models with many independent variables, as we will see later.

\*We have copied this command to Section #3 so the variable already exists:

\*gen chXfund = fundmQ \* chGoNowQ

regress cIPS9 fundmQ chGoNowQ chXfund, b

### \*Practice with a new dv

#### \* Change dependent variable

\*Use the tab and table commands to orient yourself around the new dependent variable, attitudes towards “therapeutic cloning”, use of stem cells derived from “spare” embryos from IVF clinics.

\*Run the simple regression equation predicting attitudes towards therapeutic cloning from affiliation with a fundamentalist denomination.

regress cThrp9 fundmQ , b

\*How does the difference of means in this regression equation compare to the difference we found when the dependent variable was cIPS9?

\*Next, do the simple regression just using chGoNowQ as a predictor.

regress cThrp9 chGoNowQ , b

\*How does the effect of church attendance in this regression equation compare to the difference we found when the dependent variable was cIPS9?

\*Now, enter both predictors together.

regress cThrp9 fundmQ chGoNowQ , b

\* How do the effects compare to what you found entering each predictor separately?

\*Is there an interaction effect?

regress cThrp9 fundmQ chGoNowQ chXfund, b

\*Make predicted values in Excel, graph them, and interpret them.

## \*Week 4 (Week #4 stuff is in Section #3, far above. It has to do with defining new variables. )

## \*Week 5 & 6: Predicted values & interactions (using Eval4)

\*When the effects in your regression involve curves or interactions, STATA's "Predict" function does not work correctly, because it works with the means. Instead, we need a "whole population standardization" in which the predicted values from the regression for every case in the sample are calculated, and then the summary simulations are conducted. The program to use here is Eval4. A copy of it is in the Learning Module for this week. You can also use it to produce simple linear predicted values. Moreover, Eval4 produces confidence intervals for the regression line. These are especially useful when you have an interaction, and you want to know WHERE the two groups differ (not just whether they differ on average.

### \*Eval4

\*Run Eval4.doc now (open the file; copy to the Do File Window; run)

\*For help on how to use Eval4: eval4 help

### \*Example #1:

#### \*The dependent variable

Questions on what we call "EUGENIC" CLONING.

The three questions are scattered about in the questionnaire, each together with other questions that involve the same technique -- details in the questionnaire.

[re therapeutic cloning -- which uses the respondent's own cells but in a way not approved in the USA although OK in the UK] ctSkin j. How about using cloned cells to create new skin, to restore someone’s youthful appearance?

[re cloning using "spare" IVf cells, a procedure approved in the USA] cnSkin c. Using cloned cells to create new skin, to restore someone’s youthful appearance?

[re using adult cells, no embryos destroyed] cipsSkin c. To create new skin, to restore someone’s youthful appearance?

Answer categories for all these questions, and scoring:

100 Yes!!

75 Yes

50 ??

25 No

0 No!!

The scale, cEugen9, is the average of the 3 questions.

tab cEugen9, m

cEugen9 | Freq. Percent Cum.

------------+-----------------------------------

0 | 353 15.38 15.38

8.333333 | 93 4.05 19.43

12.5 | 2 0.09 19.52

16.66667 | 154 6.71 26.23

25 | 333 14.51 40.74

33.33333 | 182 7.93 48.67

37.5 | 4 0.17 48.85

41.66667 | 141 6.14 54.99

50 | 343 14.95 69.93

58.33333 | 109 4.75 74.68

62.5 | 2 0.09 74.77

66.66666 | 110 4.79 79.56

75 | 167 7.28 86.84

83.33334 | 57 2.48 89.32

87.5 | 5 0.22 89.54

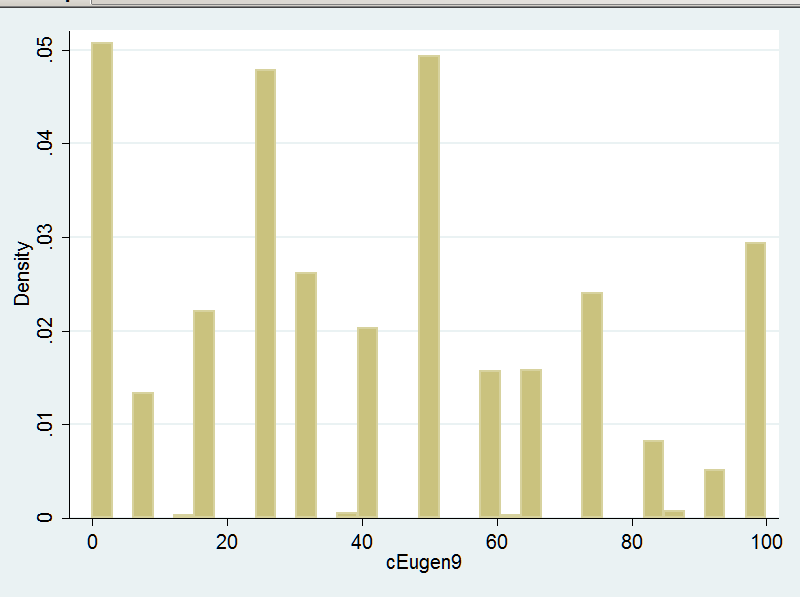
91.66666 | 36 1.57 91.11

100 | 204 8.89 100.00

------------+-----------------------------------

Total | 2,295 100.00

histogram cEugen9



#### \*One independent variable: church going

tab chGoNowQ, m

Church |

going |

(times per |

year) | Freq. Percent Cum.

------------+-----------------------------------

.5 | 638 27.80 27.80

.75 | 343 14.95 42.75

1.5 | 105 4.58 47.32

4 | 300 13.07 60.39

12 | 91 3.97 64.36

30 | 98 4.27 68.63

45 | 121 5.27 73.90

52 | 364 15.86 89.76

140 | 31 1.35 91.11

. | 204 8.89 100.00

------------+-----------------------------------

Total | 2,295 100.00

#### \*Start with an ordinary regression, one independent variable, no interactions

regress cEugen9 chGoNowQ, b

Source | SS df MS Number of obs = 2091

-------------+------------------------------ F( 1, 2089) = 42.74

Model | 38169.3049 1 38169.3049 Prob > F = 0.0000

Residual | 1865410.48 2089 892.968159 R-squared = 0.0201

-------------+------------------------------ Adj R-squared = 0.0196

Total | 1903579.79 2090 910.803727 Root MSE = 29.883

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

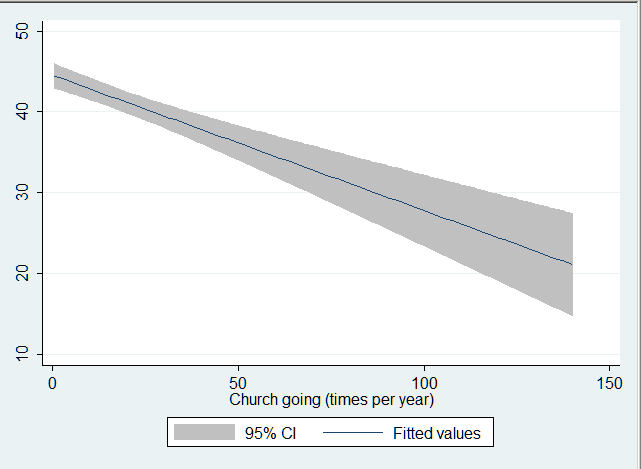
chGoNowQ | -.167011 .025545 -6.54 0.000 -.1416027

\_cons | 44.5182 .7788063 57.16 0.000 .

------------------------------------------------------------------------------

#### \*Graph it in Stata (simple graphs are easy in Stata)

twoway ( lfitci cEugen9 chGoNowQ )



#### \*Predicted values, using Eval4

\*First remind ourselves of Eval's way of doing things: (We just need a simple, basic example):

eval4 help

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

Run a probit, logit, or any linear model (OLS, xtreg, or anything

for which Stata's predict function works properly with option xb).

Choose: linear, logit, probit. Loop over 1 variable, or optionally 2.

Assumes the whole population should be analyzed (otherwise

analyze a work file restricted to the relevant cases).

SIMPLE EXAMPLES

eval4 linear age "18 50 64"

eval4 probit churchGo "ln(0.5) ln(52)"

eval4 logit age "18 64" status "0 30 35 74 100"

SPECIAL CASES: Interactions, quadratics, sets of dummies, recodes:

that is whenever one variable depends by definition on others.

...etc...

\*Now do it for our variables. Its a linear model. Eval4 knows we mean to use the regression we just ran, no need to say more about that. We want predicted values for church attendance; we will go from "never attend" to roughly "once a week", ignoring the very few who go more than that.

eval4 linear chGoNowQ "0 10 20 30 40 50"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

This is a linear model.

I'll call fix\_vars, just to be sure it runs OK (it checks dummies too)

Fix\_vars ran OK

chGoNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 44.52 .7788 42.99 46.04 0 0 1

10 42.85 .6748 41.53 44.17 -1.67 -1.67 .9625

20 41.18 .6593 39.89 42.47 -1.67 -3.34 .925

30 39.51 .7379 38.06 40.95 -1.67 -5.01 .8875

40 37.84 .8859 36.1 39.57 -1.67 -6.68 .8499

50 36.17 1.075 34.06 38.27 -1.67 -8.351 .8124

\*---------------------------------------------------------------------------------\*

\*Compare the Eval4 output to the Stata graph.

#### \*Do it again with another independent variable

\*Lets look at fundamentalists

\*First: who are they? (There is a big literature on this, and also things we can do with our data to sort things out. But for now, lets just take a conventional definition.)

table DnomNow, c(m fundmQ freq)

--------------------------------------------

DnomNow | mean(fundmQ) Freq.

----------------+---------------------------

-1 | 0 192

1. No relig | 0 641

2. AME | 1 7

3. Baptst | 1 273

4. Cath | 0 337

5. C of Christ | 1 74

6. Congreg | 0 14

7. Episcop | 0 40

8. Jewish | 0 64

9. Jehovah Witt | 1 17

10. Hindu | 0 7

11. LDS | 0 19

12. Lutheran | 0 93

13. Method | 0 118

14. Muslim | 0 11

15. Non-demon | 1 225

16. Orthodox | 0 10

17. Pentcost | 1 54

18. Presby | 0 66

19. 7th day | 1 12

20. Pagan | 0 21

--------------------------------------------

tab fundmQ, m

RECODE of |

DnomNow | Freq. Percent Cum.

------------+-----------------------------------

0 | 1,633 71.15 71.15

1 | 662 28.85 100.00

------------+-----------------------------------

Total | 2,295 100.00

#### \*Now the new regression, a graph, and predicted values:

regress cEugen9 fundmQ , b

Source | SS df MS Number of obs = 2295

-------------+------------------------------ F( 1, 2293) = 21.43

Model | 19652.2434 1 19652.2434 Prob > F = 0.0000

Residual | 2102461.45 2293 916.904252 R-squared = 0.0093

-------------+------------------------------ Adj R-squared = 0.0088

Total | 2122113.69 2294 925.071357 Root MSE = 30.28

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

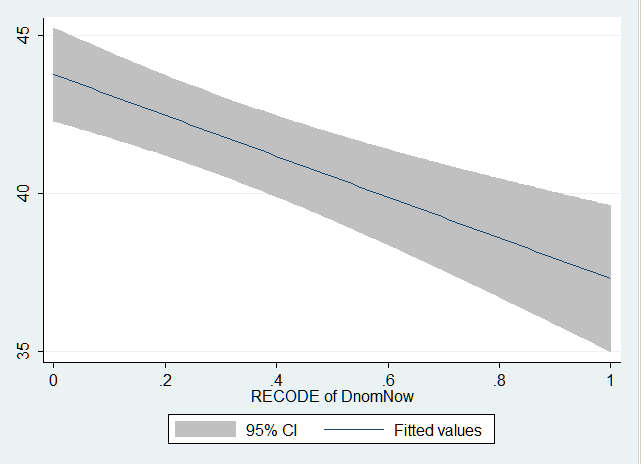
-------------+----------------------------------------------------------------

fundmQ | -6.459149 1.395183 -4.63 0.000 -.0962325

\_cons | 43.76403 .7493227 58.40 0.000 .

------------------------------------------------------------------------------

twoway ( lfitci cEugen9 fundmQ )



eval4 linear fundmQ "0 1"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

This is a linear model.

I'll call fix\_vars, just to be sure it runs OK (it checks dummies too)

Fix\_vars ran OK

fundmQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 43.76 .7493 42.3 45.23 0 0 1

1 37.3 1.177 35 39.61 -6.459 -6.459 .8524

\*---------------------------------------------------------------------------------\*

#### \*Two independent variables together

\*If they are correlated -- i.e. there is overlap between them -- then taking them only one at a time as we have done so far will exaggerate their influence. Are they correlated?

corr fundmQ chGoNowQ cEugen9

(obs=2091)

| fundmQ chGoNowQ cEugen9

-------------+---------------------------

fundmQ | 1.0000

chGoNowQ | 0.2127 1.0000

cEugen9 | -0.1099 -0.1416 1.0000

Answer: they are correlated, but only a little. So the regressions with each separately will have slightly over-estimated their true importance.

\*Now look at both together

regress cEugen9 fundmQ chGoNowQ , b

Source | SS df MS Number of obs = 2091

-------------+------------------------------ F( 2, 2088) = 28.66

Model | 50869.4005 2 25434.7003 Prob > F = 0.0000

Residual | 1852710.39 2088 887.313404 R-squared = 0.0267

-------------+------------------------------ Adj R-squared = 0.0258

Total | 1903579.79 2090 910.803727 Root MSE = 29.788

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -5.492197 1.451714 -3.78 0.000 -.083593

chGoNowQ | -.1460418 .0260602 -5.60 0.000 -.1238237

\_cons | 45.83044 .8502976 53.90 0.000 .

------------------------------------------------------------------------------

\*And for comparison, here again are the single variable results:

regress cEugen9 fundmQ , b

regress cEugen9 chGoNowQ , b

...

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -6.459149 1.395183 -4.63 0.000 -.0962325

\_cons | 43.76403 .7493227 58.40 0.000 .

------------------------------------------------------------------------------

...

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

chGoNowQ | -.167011 .025545 -6.54 0.000 -.1416027

\_cons | 44.5182 .7788063 57.16 0.000 .

------------------------------------------------------------------------------

So the single variable regressions modestly over-stated things: the effect of fundmQ and chGoNowQ is more accurately shown in the multiple regression.

#### \*Extend that idea: control for lots of other variables

regress cEugen9 fundmQ chGoNowQ maleQ ageQ edNowQ rural polDem9 lnFamInc, b

Source | SS df MS Number of obs = 1584

-------------+------------------------------ F( 8, 1575) = 17.85

Model | 123337.016 8 15417.127 Prob > F = 0.0000

Residual | 1360223.54 1575 863.633996 R-squared = 0.0831

-------------+------------------------------ Adj R-squared = 0.0785

Total | 1483560.56 1583 937.182918 Root MSE = 29.388

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -4.835158 1.685356 -2.87 0.004 -.0725654

chGoNowQ | -.1157187 .0305231 -3.79 0.000 -.0969895

maleQ | 9.489746 1.488888 6.37 0.000 .1550321

ageQ | -.1201485 .0488251 -2.46 0.014 -.0615989

edNowQ | .5940966 .3427608 1.73 0.083 .0443058

rural | -1.67962 1.560574 -1.08 0.282 -.0264207

polDem9 | 11.7243 2.321756 5.05 0.000 .1271608

lnFamInc | -1.299957 .417628 -3.11 0.002 -.0781302

\_cons | 36.57025 5.451592 6.71 0.000 .

------------------------------------------------------------------------------

These other controls further reduced the effects, but not by a lot.

Controls for things more closely related to fundamentalism and church attendance (for example, religious beliefs) might well do more.

#### \*Interactions? -- a first look

\*Now another possibility: we speculate that there may be something special about fundamentalists who regularly go to church -- not just that they are less keen on eugenics than folk in other denominations (although that is true) nor just because they go more often to church than other folk (although that too is true), but something more.

\*> Perhaps going to church with other fundamentalists reaffirms and reinforces their beliefs, insulating them from secular pressures in the wider non-fundamentalist worlds.

\*> Or putting the same idea the other way around, perhaps fundamentalists who only rarely go to church are more exposed to secular pressures from the wider world, and that undermines and devalues their beliefs.

\*> Those arguments (and others like them) suggest an INTERACTION.

\*> A word to the wise: interactions are often very interesting, not to mention publishable (see, for example, "National Context, Parental Socialization, and Religious Belief: Results From 15 Nations", Jonathan Kelley and Nan Dirk De Graaf, *American Sociological Review* 1997). But they are rare. Treasure them when you find them!

\*> First, lets explore the possibility by running the regression twice, once for fundamentalists and once for others. (Note the strange double equal sign "==" that Stata requires in an "if" clause.)

regress cEugen9 chGoNowQ if fundmQ==1, b

These are the fundamentalists:

...

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

chGoNowQ | -.0758921 .0423297 -1.79 0.073 -.0712486

\_cons | 38.5949 1.555252 24.82 0.000 .

------------------------------------------------------------------------------

regress cEugen9 chGoNowQ if fundmQ==0, b

Everyone else here:

...

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

chGoNowQ | -.184524 .0328523 -5.62 0.000 -.1455813

\_cons | 46.33085 .8981201 51.59 0.000 .

------------------------------------------------------------------------------

It looks like there is an interaction, with church going mattering more for non-fundamentalists

\*Alternatively, we could do it the other way around, comparing those who rarely go to church with those who go a lot. We find the same interaction, in a different disguise:

regress cEugen9 fundmQ if chGoNowQ <= 2, b

Attend church twice a year or less:

...

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -9.835638 2.34774 -4.19 0.000 -.1262264

\_cons | 47.02314 1.007512 46.67 0.000 .

------------------------------------------------------------------------------

regress cEugen9 fundmQ if chGoNowQ > 2, b

Attend church more than twice a year:

...

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -2.542783 1.792453 -1.42 0.156 -.0407987

\_cons | 39.89848 1.10804 36.01 0.000 .

------------------------------------------------------------------------------

#### \*Interactions -- significance test, using a multiplicative interaction variable

\*There is another way to do this, which is harder to set up but has some advantages. It gives a significance test, and is easily extended to add lots of other controls.

\*First we do the simple version. Begin by making an interaction variable (we actually made it earlier, so we drop the original -- otherwise Stata fusses -- and remake it, just for clarity):

drop chXfund

gen chXfund = fundmQ \* chGoNowQ

\*Here is what we made:

tab chXfund

chXfund | Freq. Percent Cum.

------------+-----------------------------------

0 | 1,459 69.78 69.78

.5 | 76 3.63 73.41

.75 | 101 4.83 78.24

1.5 | 23 1.10 79.34

4 | 98 4.69 84.03

12 | 39 1.87 85.89

30 | 56 2.68 88.57

45 | 54 2.58 91.15

52 | 174 8.32 99.47

140 | 11 0.53 100.00

------------+-----------------------------------

Total | 2,091 100.00

table chGoNowQ fundmQ , c(m chXfund )

----------------------

Church |

going | RECODE of

(times | DnomNow -- this is fundmQ, poorly labeled

per year) | 0 1

----------+-----------

.5 | 0 .5

.75 | 0 .75

1.5 | 0 1.5

4 | 0 4

12 | 0 12

30 | 0 30

45 | 0 45

52 | 0 52

140 | 0 140

----------------------

\*now the test:

regress cEugen9 fundmQ chGoNowQ chXfund , b

Source | SS df MS Number of obs = 2091

-------------+------------------------------ F( 3, 2087) = 20.46

Model | 54396.4075 3 18132.1358 Prob > F = 0.0000

Residual | 1849183.38 2087 886.048578 R-squared = 0.0286

-------------+------------------------------ Adj R-squared = 0.0272

Total | 1903579.79 2090 910.803727 Root MSE = 29.767

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -7.735949 1.83554 -4.21 0.000 -.1177436

chGoNowQ | -.184524 .0324067 -5.69 0.000 -.1564514

chXfund | .1086319 .0544482 2.00 0.046 .0675102

\_cons | 46.33085 .8859367 52.30 0.000 .

------------------------------------------------------------------------------

The difference is significant, but only just (and we have not yet controlled for age, gender, and other variables).

#### \*Predicted values for the interaction using Eval4

\*You need to tell subroutine fix\_vars about the interaction you just created -- Eval4 will need that information.

\*Recall it was created by: gen chXfund = fundmQ \* chGoNowQ

\*You need to put that into fix\_vars, adding it to whatever was already in fix\_vars and changing the "gen" to "quietly replace". Like this:

capture program drop fix\_vars /\*Drop previous version, so can replace \*/

program define fix\_vars

/\* compute interactions & quadratics: "quietly replace" \*/

quietly display " "

/\*interactions, as previously defined but here with "quietly replace" not "gen" \*/

quietly replace chXfund = fundmQ \* chGoNowQ

end

\*Now re-run fix\_vars (pop it into Stata's "do file editor" and run from there). Stata will then show:

. do "C:\Users\Puck\AppData\Local\Temp\STD03000000.tmp"

. capture program drop fix\_vars /\*Drop previous version, so can replace \*/

. program define fix\_vars

1.

. /\* compute interactions & quadratics: "quietly replace" \*/

. quietly display " "

2.

. /\*interactions, as previously defined but here with "quietly replace" not "gen" \*/

. quietly replace chXfund = fundmQ \* chGoNowQ

3.

. end

.

end of do-file

\*Now run Eval4, asking about both fundamentalists and church going. We will re-run the regression first, just for drill, and then eval4:

regress cEugen9 fundmQ chGoNowQ chXfund , b

eval4 linear chGoNowQ "0 10 20 30 40 50" fundmQ "0 1"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

This is a linear model.

I'll call fix\_vars, just to be sure it runs OK (it checks dummies too)

Fix\_vars ran OK

\*---------- fundmQ = 0

chGoNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 46.33 .8859 44.59 48.07 0 0 1

10 44.49 .7853 42.95 46.02 -1.845 -1.845 .9602

20 42.64 .8116 41.05 44.23 -1.845 -3.69 .9203

30 40.8 .9543 38.92 42.67 -1.845 -5.536 .8805

40 38.95 1.172 36.65 41.25 -1.845 -7.381 .8407

50 37.1 1.43 34.3 39.91 -1.845 -9.226 .8009

\*---------- fundmQ = 1

chGoNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 38.59 1.608 35.44 41.75 0 0 1

10 37.84 1.351 35.19 40.48 -.7589 -.7589 .9803

20 37.08 1.203 34.72 39.43 -.7589 -1.518 .9607

30 36.32 1.205 33.96 38.68 -.7589 -2.277 .941

40 35.56 1.357 32.9 38.22 -.7589 -3.036 .9213

50 34.8 1.616 31.63 37.97 -.7589 -3.795 .9017

\*---------------------------------------------------------------------------------\*

>>Note that the bits in green are close to what we got from the two separate regressions (compare them with the slopes in the separate regressions).

>>Church going matters more for non-fundamentalists than for fundamentalists, as we have seen before.

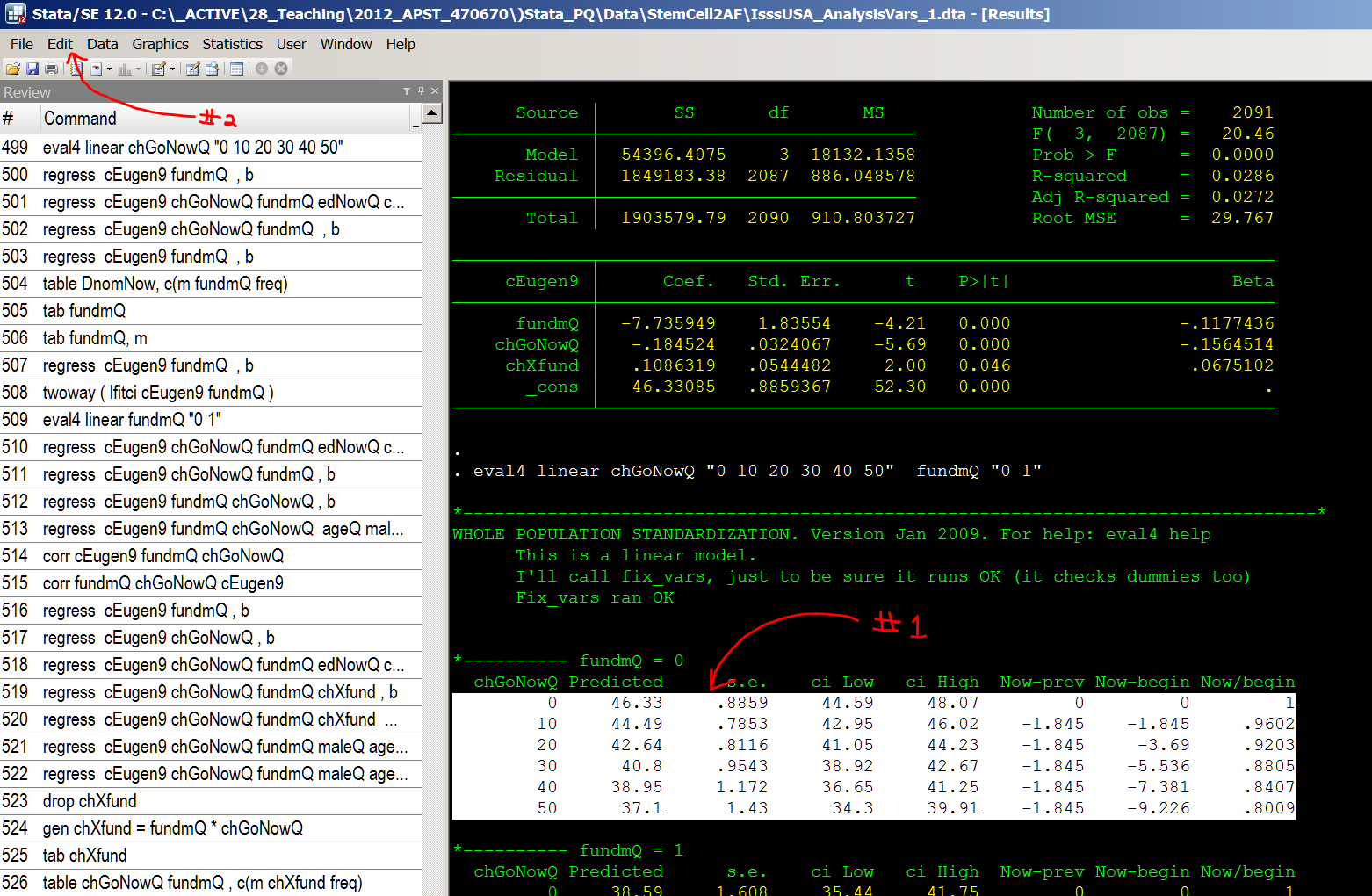
#### \*Graphing the predicted values

\*We will graph the predicted values in Excel (for multiple regressions that is easier than Stata).

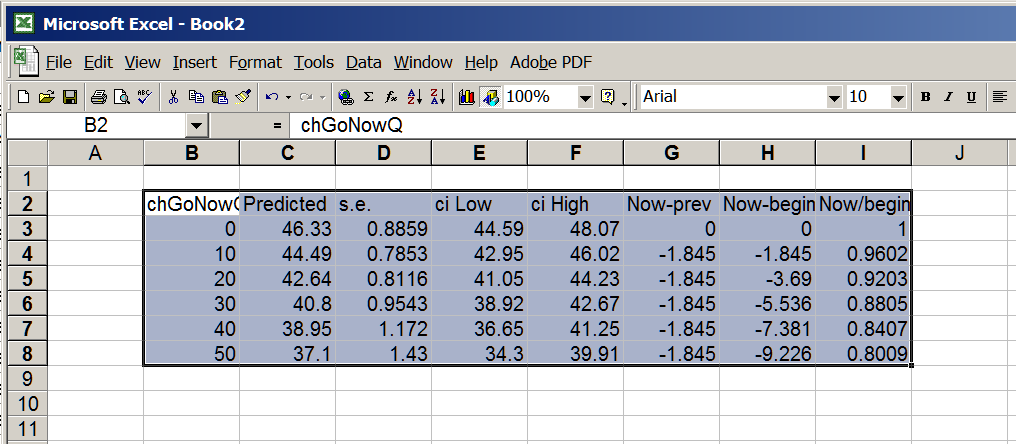
\*First get data for fundamentalists into Excel.

\* #1. Highlight it in Stata

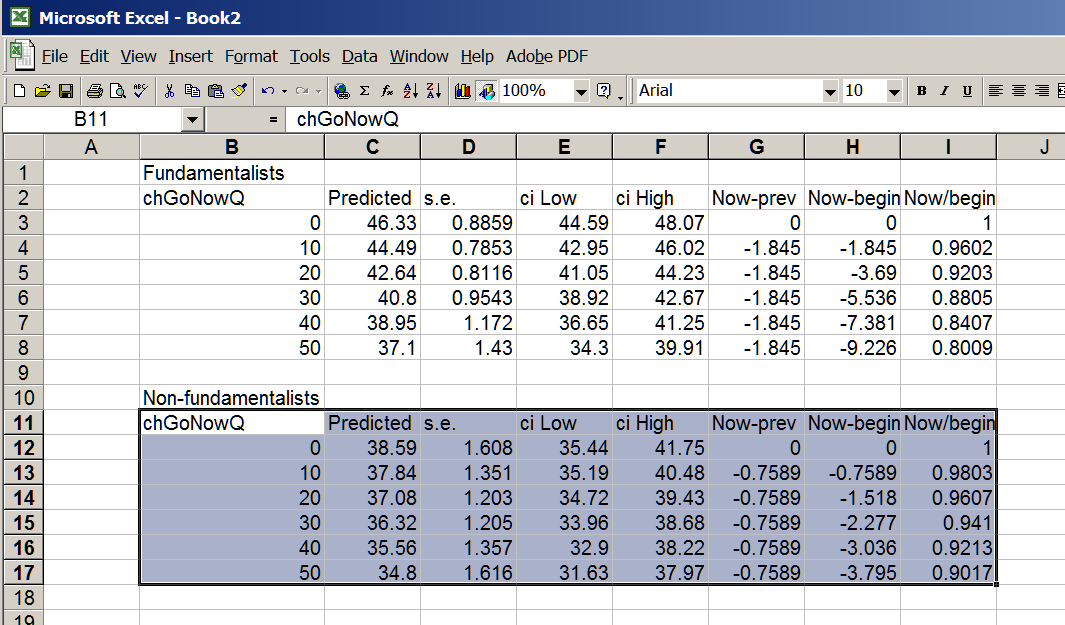
\* #2. EDIT/ COPY TABLE



\* Paste into Excel (note that "COPY TABLE" gets it into separate columns, as desired):



\*Label them "Fundamentalists" (so you don't forget). Then copy and paste in the data for non-fundamentalists:

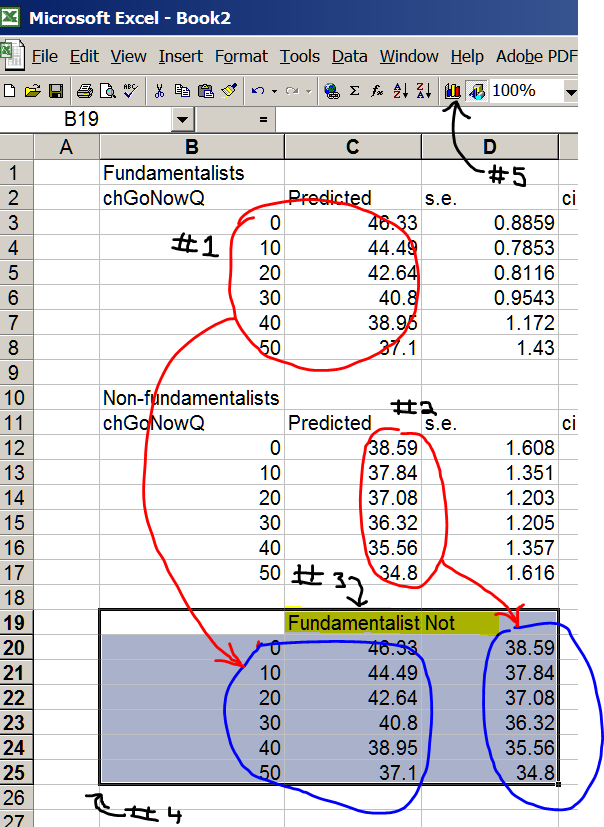


\*Now (1 and 2) copy the predicted values into a form that Excel's graphing routines want, as follows

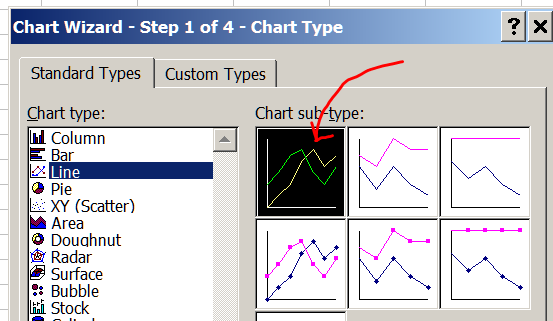
\* (3) label the two columns (in yellow in the screen shot).

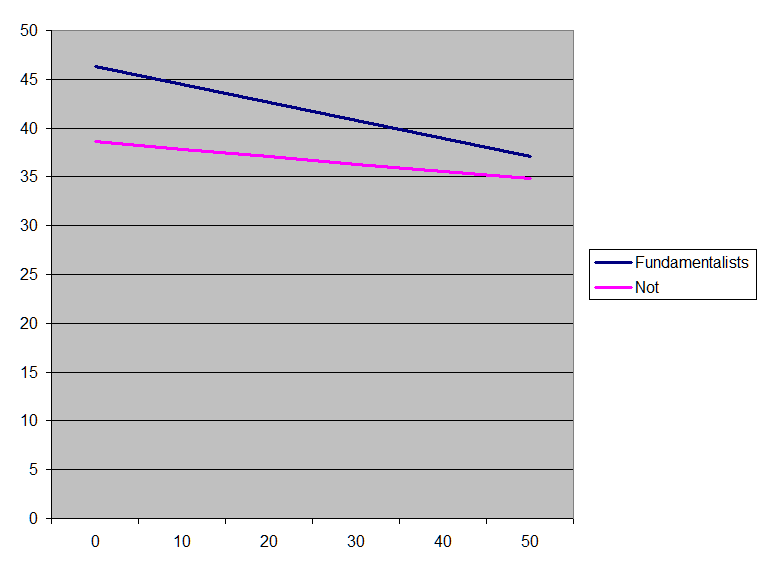
\* (4) Then select the new stuff:

\* (5) Choose the Excel "graph" icon



\*Now choose the line graph shown, and click through to "Finish":





\*You can then edit it to make it clearer

#### \*Interaction, with lots of controls

regress cEugen9 fundmQ chGoNowQ chXfund maleQ ageQ edNowQ rural polDem9 lnFamInc, b

Source | SS df MS Number of obs = 1584

-------------+------------------------------ F( 9, 1574) = 16.85

Model | 130365.913 9 14485.1014 Prob > F = 0.0000

Residual | 1353194.65 1574 859.717056 R-squared = 0.0879

-------------+------------------------------ Adj R-squared = 0.0827

Total | 1483560.56 1583 937.182918 Root MSE = 29.321

------------------------------------------------------------------------------

cEugen9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

fundmQ | -8.563515 2.127853 -4.02 0.000 -.12852

chGoNowQ | -.1781501 .0374722 -4.75 0.000 -.1493163

chXfund | .1767736 .0618232 2.86 0.004 .1086128

maleQ | 9.339498 1.486437 6.28 0.000 .1525775

ageQ | -.1145249 .048754 -2.35 0.019 -.0587158

edNowQ | .6056134 .3420063 1.77 0.077 .0451647

rural | -1.704229 1.557055 -1.09 0.274 -.0268078

polDem9 | 11.6516 2.316624 5.03 0.000 .1263723

lnFamInc | -1.28411 .4167167 -3.08 0.002 -.0771777

\_cons | 37.05379 5.441844 6.81 0.000 .

------------------------------------------------------------------------------

The interaction is still significant (a pleasant surprise, given the general way with interactions)

\*You get predicted values (same as before)

eval4 linear chGoNowQ "0 10 20 30 40 50" fundmQ "0 1"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

This is a linear model.

I'll call fix\_vars, just to be sure it runs OK (it checks dummies too)

Fix\_vars ran OK

\*---------- fundmQ = 0

chGoNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 46.5 2.062 42.46 50.54 0 0 1

10 44.72 2.005 40.79 48.65 -1.782 -1.782 .9617

20 42.93 2.018 38.98 46.89 -1.782 -3.563 .9234

30 41.15 2.1 37.04 45.27 -1.782 -5.345 .8851

40 39.37 2.243 34.98 43.77 -1.781 -7.126 .8467

50 37.59 2.435 32.82 42.36 -1.782 -8.908 .8084

\*---------- fundmQ = 1

chGoNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 37.93 2.589 32.86 43.01 0 0 1

10 37.92 2.381 33.25 42.59 -.01377 -.01377 .9996

20 37.91 2.269 33.46 42.35 -.01376 -.02753 .9993

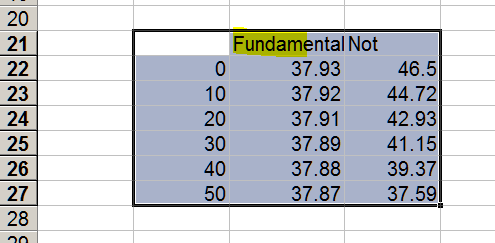
30 37.89 2.268 33.45 42.34 -.01377 -.0413 .9989

40 37.88 2.377 33.22 42.54 -.01376 -.05506 .9985

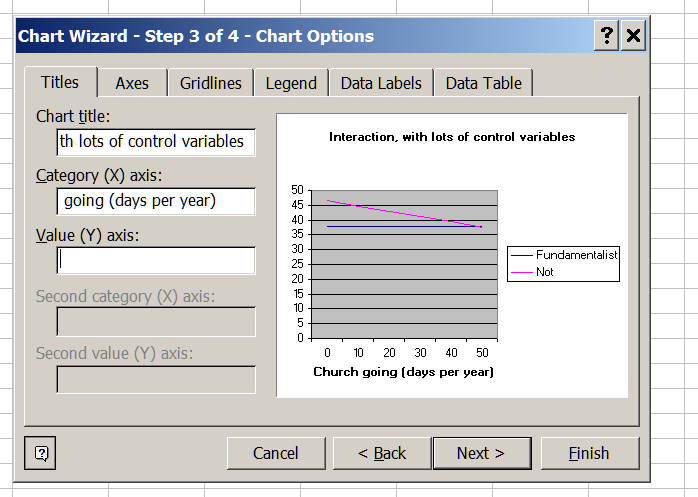
50 37.87 2.581 32.81 42.92 -.01376 -.06882 .9982

\*---------------------------------------------------------------------------------\*

\*Copy them into Excel, and graph (same as before)



\*This time we will pause to do some labeling, and dress things up a bit afterwards as well.



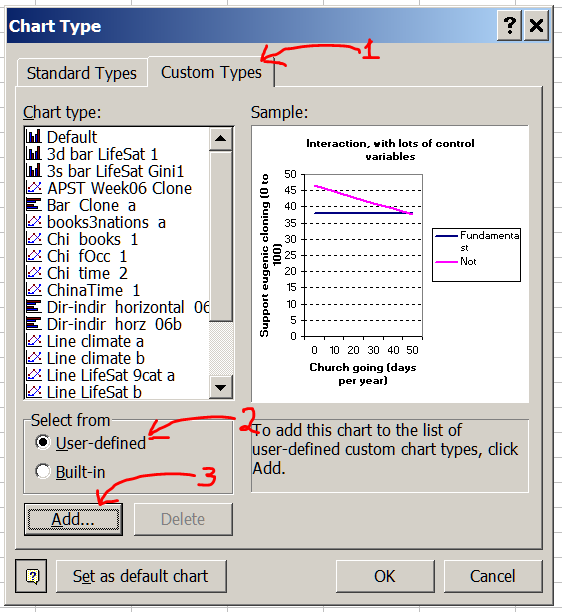
\*The interaction is much like we had before, without controls, but a bit nicer.

\*Describe what we have here, in plain English.

\*Having made a nice, tarted-up graph, you might want to save it and then re-use it on other data. For my (older) version of Excel select the chart and then in the menu (entirely un-intuitively):

\* CHART/ CHART TYPE/ CUSTOM TYPES/ USER DEFINED/ ADD

\* Then give it a name.



## \*Week 6, continued: Curvilinear effects. New example

\*Most effects are linear, at least to a reasonable approximation. So we have to hunt about a bit to find one that is curvilinear. Lets look at education and income: economic theory (and common sense) suggest that going from (say) 12 years of education to 13 won't increase income as much as going from 16 to 17, or from 19 to 20.

#### \*Is the relation linear?

\*First, lets take a quick peek:

table edNowQ, c(m famInc freq)

--------------------------------------

RECODE of |

EdNow | mean(famInc) Freq.

----------+---------------------------

10 | 26.36486 42 -- too few cases; mostly ignore

11 | 20.42 31 -- too few cases; mostly ignore

12 | 35.24776 444

13 | 45.32854 729 -- gain about $10k from year before

14 | 50.74279 255

16.5 | 70.57381 453 -- gain about $20k from before. But that is

2.5 years of ed, so we are gaining about $8k a year

18 | 85.32971 166 -- gain about $15k for 1.5 years of ed, or

something like $10k a year

22 | 136.5484 35 -- gain about $51k for 4 years of ed,

something like $13k a year

--------------------------------------

Its not linear, but a table like this is hard to read.

\*Graph it in Excel -- just like we did for predicted values.

\*In Stata, "COPY TABLE"

\*Then paste it into Excel

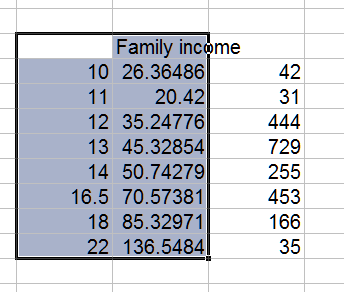
\*Label the column of means (its column "C" in the example)

\*Select the first two rows.

\*Be sure there is NOTHING in the upper left corner -- otherwise Excel gets confused

\*Click the graph icon, as before.

\*If you have saved the previous graph (the one you tarted up) as a named type, use that (give new title and axis labels. You will also need to change the Y-axis scale by hand: click on it, FORMAT AXIS/ SCALE/ enter 150 for the top category)



\*This is what you get:



\* A bit messy! I don't the drop at year 11. There are not a lot of cases there so it might be bad luck, but still I don't like it.

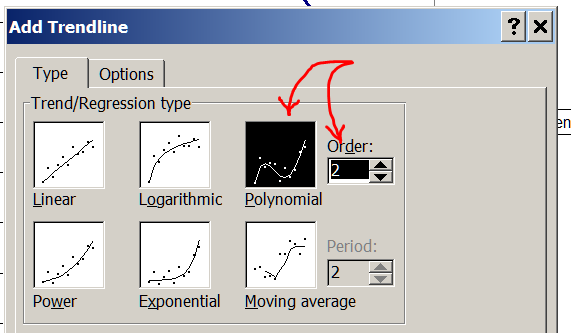
\* To re-iterate a point we have made before: bigger samples are easier to analyze! We have 2000 cases here (not bad) but it would be easier if we had 20,000. Then if the "drop" was still there, we would know it was real, not chance, and so something we have to worry about.

\*An annoyance with Excel is that it treats each education category as an equal interval above the one below -- i.e. it pays no attention to our "years" label. That is OK for 10, 11, 12, 13 and 14. But after that we have larger gaps, so its not quite right. The regression (to come) will be right.

\* Smooth things by adding a trend line.

\* Select the chart and then CHART/ ADD TREND LINE.

\* Then choose type POLYNOMIAL OF ORDER 2.



\*That gives:



\*We assume (slightly reluctantly, in this example) that a smooth curve of this sort is a reasonable summary. The most common choice of curve is a quadratic (a "polynomial of order 2" in Excel jargon). There are other choices, but we will work with that for now. Its usually a good choice.

#### \*Create a quadratic variable (we actually did it earlier; here we repeat)

drop edNowSq

gen edNowSq = edNowQ^2

\*This is what we have created

tab edNowSq, m

edNowSq | Freq. Percent Cum.

------------+-----------------------------------

100 | 42 1.83 1.83

121 | 31 1.35 3.18

144 | 444 19.35 22.53

169 | 729 31.76 54.29

196 | 255 11.11 65.40

272.25 | 453 19.74 85.14

324 | 166 7.23 92.37

484 | 35 1.53 93.90

. | 140 6.10 100.00

------------+-----------------------------------

Total | 2,295 100.00

table edNowQ, c(m edNowSq sd edNowSq freq)

-------------------------------------------------------

RECODE of |

EdNow | mean(edNowSq) sd(edNowSq) Freq.

----------+--------------------------------------------

10 | 100 0 42

11 | 121 0 31

12 | 144 0 444

13 | 169 0 729

14 | 196 0 255

16.5 | 272.25 0 453

18 | 324 0 166

22 | 484 0 35

-------------------------------------------------------

#### \*Is there statistically significant curvature?

\*to be realistic, we will control for some of the "usual suspects" -- age, gender, and occupational status.

regress famInc edNowQ edNowSq status maleQ ageQ , beta

There is significant curvature, but only just

Source | SS df MS Number of obs = 1450

-------------+------------------------------ F( 5, 1444) = 40.68

Model | 600152.377 5 120030.475 Prob > F = 0.0000

Residual | 4260561.65 1444 2950.52746 R-squared = 0.1235

-------------+------------------------------ Adj R-squared = 0.1204

Total | 4860714.02 1449 3354.53004 Root MSE = 54.319

------------------------------------------------------------------------------

famInc | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

edNowQ | -10.52542 6.342226 -1.66 0.097 -.4198902

edNowSq | .5400023 .2057438 2.62 0.009 .6623626

status | .3350793 .0701189 4.78 0.000 .1395924

maleQ | -.0079687 2.875327 -0.00 0.998 -.0000687

ageQ | .1823257 .0932067 1.96 0.051 .0492288

\_cons | 66.60159 47.67987 1.40 0.163 .

------------------------------------------------------------------------------

#### \*graph it

\*To graph it, we first have to tell subroutine fix\_vars (and therefore Eval4) just what quadratic variable we have made.

\*Recall we made it with: gen edNowSq = edNowQ^2

\*So we add that to our existing fix\_vars, and re-run it (via the do-file editor):

capture program drop fix\_vars /\*Drop previous version, so can replace \*/

program define fix\_vars

/\* compute interactions & quadratics: "quietly replace" \*/

quietly display " "

/\*interactions, as previously defined but here with "quietly replace" not "gen" \*/

quietly replace chXfund = fundmQ \* chGoNowQ

quietly replace edNowSq = edNowQ^2

end

. do "C:\Users\Puck\AppData\Local\Temp\STD03000000.tmp"

. capture program drop fix\_vars /\*Drop previous version, so can replace \*/

. program define fix\_vars

1.

. /\* compute interactions & quadratics: "quietly replace" \*/

. quietly display " "

2.

. /\*interactions, as previously defined but here with "quietly replace" not "gen" \*/

. quietly replace chXfund = fundmQ \* chGoNowQ

3.

. quietly replace edNowSq = edNowQ^2

4.

. end

\*Re-run the regression (just to be pedantic) and then run Eval4

regress famInc edNowQ edNowSq status maleQ ageQ , beta

eval4 linear edNowQ "10 11 12 13 14 15 16 17 18 19 20"

Source | SS df MS Number of obs = 1450

-------------+------------------------------ F( 5, 1444) = 40.68

Model | 600152.377 5 120030.475 Prob > F = 0.0000

Residual | 4260561.65 1444 2950.52746 R-squared = 0.1235

-------------+------------------------------ Adj R-squared = 0.1204

Total | 4860714.02 1449 3354.53004 Root MSE = 54.319

------------------------------------------------------------------------------

famInc | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

edNowQ | -10.52542 6.342226 -1.66 0.097 -.4198902

edNowSq | .5400023 .2057438 2.62 0.009 .6623626

status | .3350793 .0701189 4.78 0.000 .1395924

maleQ | -.0079687 2.875327 -0.00 0.998 -.0000687

ageQ | .1823257 .0932067 1.96 0.051 .0492288

\_cons | 66.60159 47.67987 1.40 0.163 .

------------------------------------------------------------------------------

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

This is a linear model.

I'll call fix\_vars, just to be sure it runs OK (it checks dummies too)

Fix\_vars ran OK -- Eval is happy with our revised fix\_vars

edNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

10 41.44 6.21 29.27 53.61 0 0 1

11 42.25 4.536 33.36 51.14 .8146 .8146 1.02

12 44.15 3.483 37.32 50.98 1.895 2.709 1.065

13 47.12 3.073 41.1 53.15 2.975 5.684 1.137

14 51.18 3.097 45.11 57.25 4.055 9.739 1.235

15 56.31 3.263 49.92 62.71 5.135 14.87 1.359

16 62.53 3.431 55.8 69.25 6.215 21.09 1.509

17 69.82 3.628 62.71 76.93 7.295 28.38 1.685

18 78.2 4.01 70.34 86.06 8.375 36.76 1.887

19 87.65 4.774 78.29 97.01 9.455 46.21 2.115

20 98.19 6.045 86.34 110 10.53 56.75 2.369

\*---------------------------------------------------------------------------------\*

Around high school graduation, a year of education is worth a couple of thousand dollars in future income. Around the end of graduate school, its more like ten thousand.

\*Graph it in Excel in the usual way (we have also added a linear trend line to show how curve and linear compare -- note where the discrepancies are)



## \*Week 7: Logistic and Probit regressions

\* When you have a dichotomous dependent/response variable, these are the methods to use. Which one you choose is mainly a matter of convenience in your field. With extreme splits neither does very well unless your theory has led you to exactly the right set of predictors, but the probit tends to do a little better. (Check out Chapter 7 in the Hamilton "Regression with graphics" text.)

\*This approach to learning logistic and probit regression will involve dichotomous variables with different splits, and we'll also do comparisons to OLS to see what the gains are.

### \*Case #1: Dichotomous dependent/response variable, but not an extreme split

#### \*Analyze with conventional OLS first, just to see what is up

regress married chGoNowQ maleM ageM edNow2M rural polDem9, b

Source | SS df MS Number of obs = 2054

-------------+------------------------------ F( 6, 2047) = 20.39

Model | 28.8623622 6 4.8103937 Prob > F = 0.0000

Residual | 482.999858 2047 .235954987 R-squared = 0.0564

-------------+------------------------------ Adj R-squared = 0.0536

Total | 511.86222 2053 .249324023 Root MSE = .48575

------------------------------------------------------------------------------

married | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

chGoNowQ | .0014473 .0004363 3.32 0.001 .0735864

maleM | -.0547296 .0215475 -2.54 0.011 -.0546363

ageM | .0055162 .0006871 8.03 0.000 .1749014

edNow2M | .006194 .0047415 1.31 0.192 .0284918

rural | .0312748 .02263 1.38 0.167 .0300856

polDem9 | -.124615 .033389 -3.73 0.000 -.0826324

\_cons | .1840771 .0822769 2.24 0.025 .

------------------------------------------------------------------------------

\*This gets predicted values for each case -- convenient!

predict olsHat1

list married olsHat1 ageM in 1/10

+---------------------------+

| married olsHat1 ageM |

|---------------------------|

1. | 0 .26539 28 |

2. | 0 .3827111 31 |

3. | 1 .4091021 49 |

4. | 1 .3939637 42 |

5. | 0 .3503894 34 |

|---------------------------|

6. | 0 .6229985 72 |

7. | 0 .4059684 30 |

8. | 0 .5455852 36 |

9. | 1 .4950664 49 |

10. | 1 .467046 43 |

+---------------------------+

\*Lets look at another way of finding R-squared

corr married olsHat1

| married olsHat1

-------------+------------------

married | 1.0000

olsHat1 | 0.2375 1.0000

display .2375^2

.05640625 -- compare with the OLS above

\*Now graph it: Run eval4, "copy table" and paste into Excel, make graph:

eval4 linear ageM "20 30 40 50 60 70"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

This is a linear model.

I'll call fix\_vars, just to be sure it runs OK (it checks dummies too)

Fix\_vars ran OK

ageM Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

20 .3255 .03157 .2636 .3874 0 0 1

30 .3806 .02813 .3255 .4358 .05516 .05516 1.169

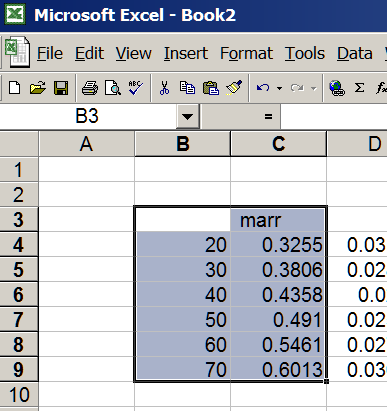
40 .4358 .0261 .3846 .487 .05516 .1103 1.339

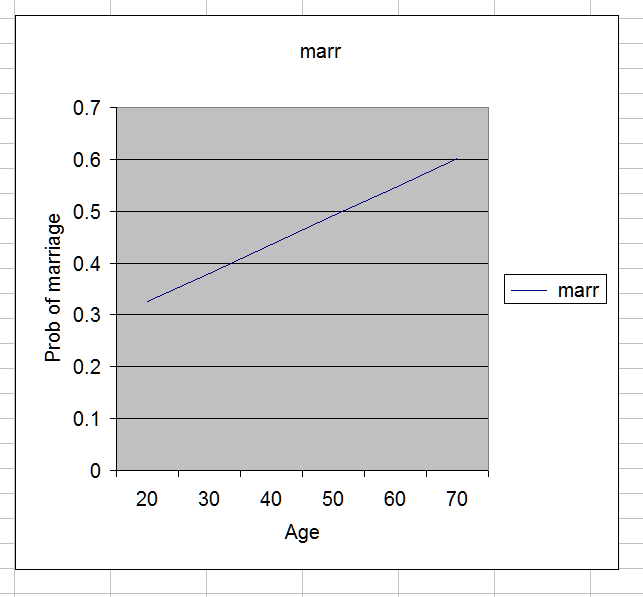
50 .491 .02585 .4403 .5416 .05516 .1655 1.508

60 .5461 .02742 .4924 .5999 .05516 .2206 1.678

70 .6013 .03051 .5415 .6611 .05516 .2758 1.847

\*---------------------------------------------------------------------------------\*





#### \*Any curvilinearity?

\*Lets take a quick look at the means:

gen age10 = round( ageM, 10)

. tab age10

age10 | Freq. Percent Cum.

------------+-----------------------------------

20 | 220 9.59 9.59 -- this is really 15 to 25

30 | 364 15.86 25.45 -- and this 25 to 35 or so

40 | 413 18.00 43.44

50 | 605 26.36 69.80

60 | 324 14.12 83.92

70 | 297 12.94 96.86

80 | 66 2.88 99.74

90 | 6 0.26 100.00 -- not enough cases

------------+-----------------------------------

Total | 2,295 100.00

table age10, c(m married freq)

----------------------------------------

age10 | mean(married) Freq.

----------+-----------------------------

20 | .1409091 220

30 | .4120879 364

40 | .5012106 413

50 | .3966942 605-- odd, out of sequence.I don't like that!

60 | .5555556 324

70 | .5690235 297

80 | .5151515 66

90 | .3333333 6

----------------------------------------

\*Graph it in Excel. Add a trend line in Excel (polynomial, degree 2, ie our usual quadratic)



##### \*Deal with the curvature systematically (as we did last week)

\*First, we are going to roughly center age on the way to squaring it, because age itself can take on pretty substantial values, so their squares are very large. Rescoring those values as differences from a number near the mean before squaring them makes the squared values smaller, so there is less rounding error in calculations using them.

gen ageMSq = (ageM -40)^2

\*now tell fix\_vars about it:

\*No need to edit out last week's interaction chXfund -- leave it in case we want it later

\*Run the revised fix\_vars from the Stata do-file editor.

capture program drop fix\_vars /\*Drop previous version, so can replace \*/

program define fix\_vars

/\* compute interactions & quadratics: "quietly replace" \*/

quietly display " "

quietly replace edNowSq = edNowQ^2

quietly replace ageMSq = (ageM -40)^2

/\*interactions, as previously defined but here with "quietly replace" not "gen" \*/

quietly replace chXfund = fundmQ \* chGoNowQ

end

\*Now our regression, with a proper test of the curvature:

regress married chGoNowQ maleM ageM ageMSq edNow2M rural polDem9, b

Source | SS df MS Number of obs = 2054

-------------+------------------------------ F( 7, 2046) = 23.54

Model | 38.1464912 7 5.44949874 Prob > F = 0.0000

Residual | 473.715729 2046 .231532614 R-squared = 0.0745

-- look how much bigger R-squared is

-------------+------------------------------ Adj R-squared = 0.0714

Total | 511.86222 2053 .249324023 Root MSE = .48118

------------------------------------------------------------------------------

married | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

chGoNowQ | .0014517 .0004322 3.36 0.001 .073808

maleM | -.0533131 .0213458 -2.50 0.013 -.0532222

ageM | .0092809 .0009037 10.27 0.000 .2942701

ageMSq | -.0002572 .0000406 -6.33 0.000 -.1807948

-- this is the right sig test

edNow2M | .0081668 .0047072 1.73 0.083 .0375663

rural | .0348956 .0224242 1.56 0.120 .0335688

polDem9 | -.1367353 .0331299 -4.13 0.000 -.0906694

\_cons | .0603565 .0838113 0.72 0.472 .

------------------------------------------------------------------------------

\*Now for predicted values: noticeably better than what we had before

predict olsHat2

list married olsHat1 olsHat2 ageM in 1/10

corr married olsHat1 olsHat2 ageM

+--------------------------------------+

| married olsHat1 olsHat2 ageM |

|--------------------------------------|

1. | 0 .26539 .2260525 28 |

2. | 0 .3827111 .3868366 31 |

3. | 1 .4091021 .4739023 49 |

4. | 1 .3939637 .462032 42 |

5. | 0 .3503894 .3624285 34 |

|--------------------------------------|

6. | 0 .6229985 .5217903 72 |

7. | 0 .4059684 .3946553 30 |

8. | 0 .5455852 .5664732 36 |

9. | 1 .4950664 .5610106 49 |

10. | 1 .467046 .5259852 43 |

+--------------------------------------+

. corr married olsHat1 olsHat2 ageM

| married olsHat1 olsHat2 ageM

-------------+------------------------------------

married | 1.0000

olsHat1 | 0.2375 1.0000

olsHat2 | 0.2730 0.8698 1.0000

ageM | 0.1917 0.8071 0.7021 1.0000

##### \*Graph them in Excel, in the usual way. Up to age 80 this time.

eval4 linear ageM "20 30 40 50 60 70 80"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

This is a linear model.

I'll call fix\_vars, just to be sure it runs OK (it checks dummies too)

Fix\_vars ran OK

ageM Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

20 .1981 .0373 .125 .2712 0 0 1

30 .3681 .02796 .3133 .4229 .17 .17 1.858

40 .4866 .0271 .4335 .5397 .1185 .2885 2.456

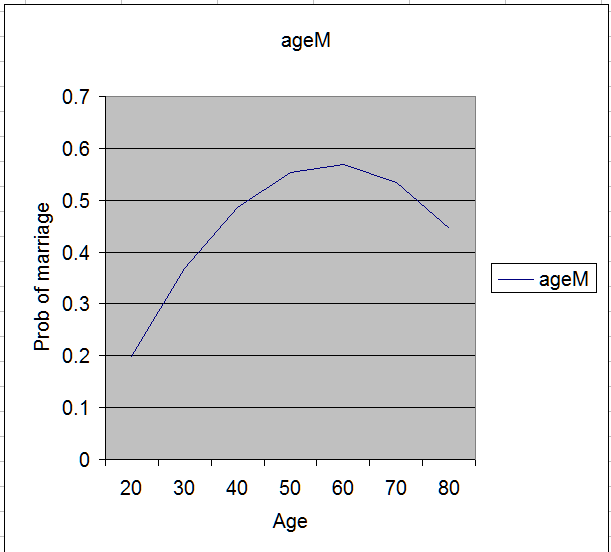
50 .5537 .0275 .4998 .6076 .06709 .3556 2.795

60 .5694 .02743 .5156 .6231 .01566 .3712 2.874

70 .5336 .03211 .4707 .5965 -.03577 .3355 2.693

80 .4464 .04786 .3526 .5402 -.0872 .2483 2.253

\*---------------------------------------------------------------------------------\*



#### \*Now a probit regression

\*Now the probit regression, starting without the age squared term (just as we started the OLS without it):

probit married chGoNowQ maleM ageM edNow2M rural polDem9

It looks quite different than the OLS, but is actually quite similar in the end

Iteration 0: log likelihood = -1420.447

Iteration 1: log likelihood = -1361.2785

Iteration 2: log likelihood = -1361.2095

Iteration 3: log likelihood = -1361.2095

Probit regression Number of obs = 2054

LR chi2(6) = 118.48

Prob > chi2 = 0.0000

Log likelihood = -1361.2095 Pseudo R2 = 0.0417

------------------------------------------------------------------------------

married | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

chGoNowQ | .003755 .0011375 3.30 0.001 .0015255 .0059845

maleM | -.1437437 .0566409 -2.54 0.011 -.2547579 -.0327295

ageM | .0144116 .0018248 7.90 0.000 .010835 .0179882

edNow2M | .0160778 .0124471 1.29 0.196 -.0083181 .0404736

rural | .0807265 .0594317 1.36 0.174 -.0357575 .1972106

polDem9 | -.3216707 .0872925 -3.68 0.000 -.4927609 -.1505805

\_cons | -.825758 .2161993 -3.82 0.000 -1.249501 -.4020151

------------------------------------------------------------------------------

\*This gets predicted values for each case, just as for OLS -- convenient! Note the "pr" option, which asks for predicted probabilities (the default is something else)

predict probHat1, pr

list married olsHat1 olsHat2 probHat1 ageM in 1/10

Not actually very different from our earlier OLS estimates

+-------------------------------------------------+

| married olsHat1 olsHat2 probHat1 ageM |

|-------------------------------------------------|

1. | 0 .26539 .2260525 .2701299 28 |

2. | 0 .3827111 .3868366 .3787908 31 |

3. | 1 .4091021 .4739023 .4061627 49 |

4. | 1 .3939637 .462032 .3908046 42 |

5. | 0 .3503894 .3624285 .3485191 34 |

|-------------------------------------------------|

6. | 0 .6229985 .5217903 .6262043 72 |

7. | 0 .4059684 .3946553 .4026269 30 |

8. | 0 .5455852 .5664732 .5467135 36 |

9. | 1 .4950664 .5610106 .4948178 49 |

10. | 1 .467046 .5259852 .4655033 43 |

+-------------------------------------------------+

#### \*Any extra curvilinearity in the probit?

probit married chGoNowQ maleM ageM ageMSq edNow2M rural polDem9

Lots!

Iteration 0: log likelihood = -1420.447

Iteration 1: log likelihood = -1340.2063

Iteration 2: log likelihood = -1339.9431

Iteration 3: log likelihood = -1339.9431

Probit regression Number of obs = 2054

LR chi2(7) = 161.01

Prob > chi2 = 0.0000

Log likelihood = -1339.9431 Pseudo R2 = 0.0567

------------------------------------------------------------------------------

married | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

chGoNowQ | .0037754 .0011396 3.31 0.001 .0015419 .0060089

maleM | -.1454243 .0570418 -2.55 0.011 -.2572243 -.0336244

ageM | .0256922 .0025779 9.97 0.000 .0206395 .0307449

ageMSq | -.0007221 .0001118 -6.46 0.000 -.0009412 -.0005031

edNow2M | .0212936 .0125598 1.70 0.090 -.0033231 .0459103

rural | .0882706 .0599048 1.47 0.141 -.0291407 .2056819

polDem9 | -.3596551 .0880474 -4.08 0.000 -.5322248 -.1870853

\_cons | -1.203333 .2261728 -5.32 0.000 -1.646624 -.7600428

------------------------------------------------------------------------------

\*Graph it using Eval4 and Excel, as before.

eval4 probit ageM "20 30 40 50 60 70 80"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION.

ageM Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

20 .2046 .02927 .1519 .2667 0 0 1

30 .3591 .02782 .3058 .4149 .1545 .1545 1.755

40 .4848 .02801 .4299 .5397 .1258 .2802 2.37

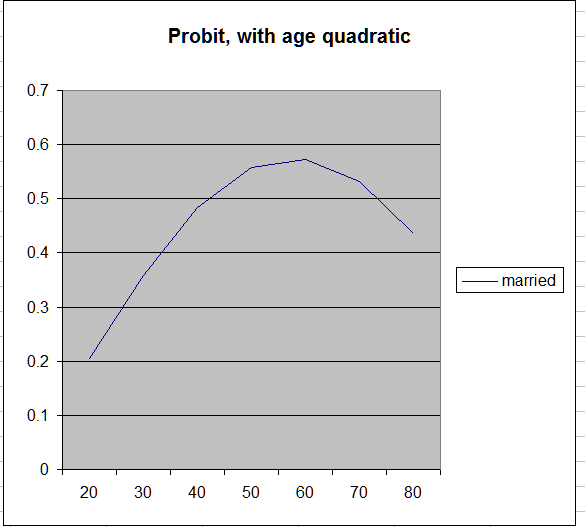
50 .5569 .0281 .5011 .6113 .07211 .3523 2.722

60 .5725 .02784 .5171 .6262 .01556 .3679 2.798

70 .5321 .03294 .467 .5961 -.04041 .3275 2.601

80 .4351 .04858 .3417 .5322 -.09698 .2305 2.127

\*---------------------------------------------------------------------------------\*



\*In case you thought this was pretty familiar, lets compare it systematically with OLS. Get the predicted value, in the usual way:

predict probHat2, pr

list married olsHat1 olsHat2 probHat1 probHat2 ageM in 1/10

+------------------------------------------------------------+

| married olsHat1 olsHat2 probHat1 probHat2 ageM |

|------------------------------------------------------------|

1. | 0 .26539 .2260525 .2701299 .2279259 28 |

2. | 0 .3827111 .3868366 .3787908 .3756584 31 |

3. | 1 .4091021 .4739023 .4061627 .4741499 49 |

4. | 1 .3939637 .462032 .3908046 .4591648 42 |

5. | 0 .3503894 .3624285 .3485191 .3558584 34 |

|------------------------------------------------------------|

6. | 0 .6229985 .5217903 .6262043 .52037 72 |

7. | 0 .4059684 .3946553 .4026269 .3834852 30 |

8. | 0 .5455852 .5664732 .5467135 .5655384 36 |

9. | 1 .4950664 .5610106 .4948178 .567229 49 |

10. | 1 .467046 .5259852 .4655033 .5292059 43 |

+------------------------------------------------------------+

corr married olsHat1 olsHat2 probHat1 probHat2 ageM

| married olsHat1 olsHat2 probHat1 probHat2 ageM

-------------+------------------------------------------------------

married | 1.0000

olsHat1 | 0.2375 1.0000

olsHat2 | 0.2730 0.8698 1.0000

probHat1 | 0.2369 0.9997 0.8687 1.0000

probHat2 | 0.2700 0.8658 0.9987 0.8654 1.0000

ageM | 0.1917 0.8071 0.7021 0.8095 0.7035 1.0000

-- and what, pray tell, does r=.998 imply?

\*Here is the difference graphically -- it ie nothing!

#### \*Now a Logistic regression

\*here is the one without age squared; we won't bother looking at it

logit married chGoNowQ maleM ageM edNow2M rural polDem9

predict logitHat1, pr

\*Now add age squared

logit married chGoNowQ maleM ageM ageMSq edNow2M rural polDem9

Iteration 0: log likelihood = -1420.447

Iteration 1: log likelihood = -1340.72

Iteration 2: log likelihood = -1340.5272

Iteration 3: log likelihood = -1340.5272

Logistic regression Number of obs = 2054

LR chi2(7) = 159.84

Prob > chi2 = 0.0000

Log likelihood = -1340.5272 Pseudo R2 = 0.0563

------------------------------------------------------------------------------

married | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

chGoNowQ | .0062209 .0018869 3.30 0.001 .0025227 .0099191

maleM | -.229121 .0924879 -2.48 0.013 -.410394 -.0478481

ageM | .041087 .004233 9.71 0.000 .0327906 .0493835

ageMSq | -.0011544 .0001819 -6.35 0.000 -.0015109 -.0007979

edNow2M | .0355592 .0204543 1.74 0.082 -.0045306 .0756489

rural | .1464454 .0970351 1.51 0.131 -.0437399 .3366308

polDem9 | -.5806996 .1426165 -4.07 0.000 -.8602228 -.3011765

\_cons | -1.948962 .3708395 -5.26 0.000 -2.675794 -1.22213

------------------------------------------------------------------------------

\*save the predicted value for later

predict logitHat2, pr

eval4 logit ageM "20 30 40 50 60 70 80"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION.

ageM Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

20 .2117 .02837 .1612 .2724 0 0 1

30 .3604 .02792 .3072 .4167 .1487 .1487 1.703

40 .485 .02832 .4294 .5404 .1246 .2733 2.291

50 .5568 .0283 .5004 .6113 .07183 .3451 2.63

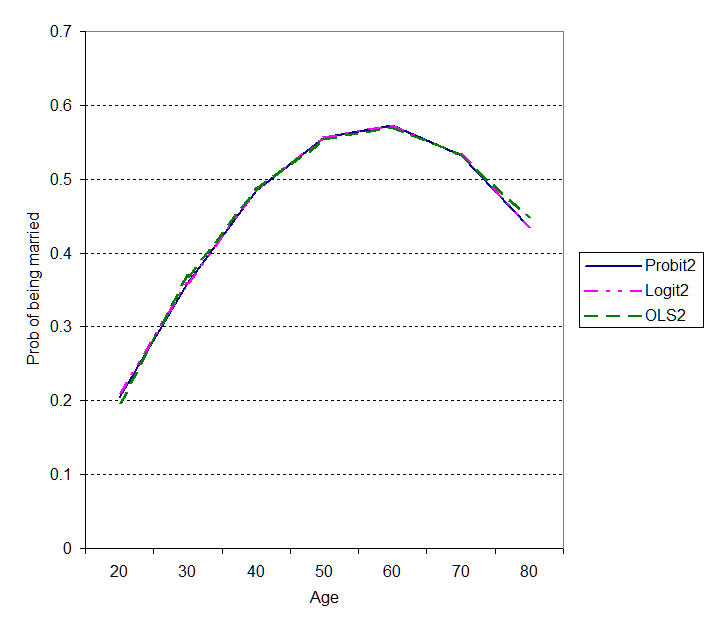
60 .5723 .028 .5164 .6261 .0155 .3606 2.704

70 .5322 .03317 .4665 .5965 -.04013 .3205 2.514

80 .4357 .0487 .3425 .5334 -.09648 .224 2.058

\*---------------------------------------------------------------------------------\*

\*Graph in Excel, as before. Here are the graphs for OLS, probit, and logit (all with age squared). It takes very sharp eyes to see any difference!



### \*Case #2: Dichotomous dependent/response variable, with an extreme split (90% vs 10% or so)

#### \*Dependent/response variable defined: works in a professional job

\*\*tab status

\*\*recode status (100=1)(10 / 90=0)(\*=.), gen( prof)

\*\*tab prof, m

\*\*table status, c(m prof freq) m

Kelleys Worldwide |

Status Score | Freq. Percent Cum.

------------------------+-----------------------------------

10. Farmer | 9 0.48 0.48

14. Unskilled worker | 98 5.26 5.74

24. Semi-skilled worker | 94 5.04 10.78

26. Service (hi & low) | 195 10.46 21.24

37. Skilled worker | 183 9.82 31.06

42. Sales (hi & low) | 238 12.77 43.83

49. Clerical (hi & low) | 372 19.96 63.79

70. Technical | 311 16.68 80.47

75. Administrative | 167 8.96 89.43

100. Professional | 197 10.57 100.00

------------------------+-----------------------------------

Total | 1,864 100.00

RECODE of |

status |

(Kelleys |

Worldwide |

Status |

Score) | Freq. Percent Cum.

------------+-----------------------------------

0 | 1,667 72.64 72.64

1 | 197 8.58 81.22

. | 431 18.78 100.00 -- lots of folk do not work

------------+-----------------------------------

Total | 2,295 100.00

------------------------------------------------

Kelleys Worldwide |

Status Score | mean(prof) Freq.

------------------------+-----------------------

10. Farmer | 0 9

14. Unskilled worker | 0 98

24. Semi-skilled worker | 0 94

26. Service (hi & low) | 0 195

37. Skilled worker | 0 183

42. Sales (hi & low) | 0 238

49. Clerical (hi & low) | 0 372

70. Technical | 0 311

75. Administrative | 0 167

100. Professional | 1 197

------------------------------------------------

#### \*Who gets professional jobs: OLS estimates (no quadratic in ed)

\*Here we use edNowQ, a slightly different version of education. The version we were using earlier paid some attention to people's educational plans. So for example, if someone was in college, of the usual age, and said they planned to finish college, we believed them and coded them into the graduate category. That is probably right for re attitudes, interest, life-style, etc. But for the analysis of current job that we are to do here, we need their actual current education, not their plans (eg you need to have your teaching certificate in hand before taking a teaching job). That is in edNowQ.

\*The original questions:

Education now, and your future plans...

Now:

10th grade or less

11th grade

GED

12th grade

Some college, no degree

Associate degree

BA, BS

MA, MS, LLB

MD

PhD

Plans:

(No more, I'm finished.)

High school diploma/ GED

Associate degree

BA, BS

MA, MS, LLB

MD

PhD

\*What we are using here:

tab edNowQ, m

RECODE of |

EdNow | Freq. Percent Cum.

------------+-----------------------------------

10 | 42 1.83 1.83

11 | 31 1.35 3.18

12 | 444 19.35 22.53

13 | 729 31.76 54.29

14 | 255 11.11 65.40

16.5 | 453 19.74 85.14

18 | 166 7.23 92.37

22 | 35 1.53 93.90

. | 140 6.10 100.00

------------+-----------------------------------

regress prof married chGoNowQ maleM ageM edNowQ rural polDem9, beta

Source | SS df MS Number of obs = 1776

-------------+------------------------------ F( 7, 1768) = 59.48

Model | 31.5896453 7 4.51280647 Prob > F = 0.0000

Residual | 134.139521 1768 .07587077 R-squared = 0.1906

-------------+------------------------------ Adj R-squared = 0.1874

Total | 165.729167 1775 .093368545 Root MSE = .27545

------------------------------------------------------------------------------

prof | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

married | .0022735 .0135153 0.17 0.866 .0037143

chGoNowQ | .0004646 .0002679 1.73 0.083 .0387635

maleM | .0211198 .013142 1.61 0.108 .03448

ageM | .0005405 .0004339 1.25 0.213 .0276229

edNowQ | .0569776 .0029539 19.29 0.000 .424771

rural | .0237321 .0138308 1.72 0.086 .0373366

polDem9 | .0005352 .0203232 0.03 0.979 .0005859

\_cons | -.760925 .0475275 -16.01 0.000 .

------------------------------------------------------------------------------

predict olsHat3

eval4 probit edNowQ "10 12 14 16 18 20 22"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION.

edNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

10 .4451 .008379 .4287 .4615 0 0 1

12 .4903 .007307 .476 .5047 .04528 .04528 1.102

14 .5358 .006793 .5224 .549 .0454 .09068 1.204

16 .5807 .006988 .5669 .5943 .04494 .1356 1.305

18 .6246 .007751 .6093 .6397 .04391 .1795 1.403

20 .667 .00879 .6495 .684 .04235 .2219 1.499

22 .7073 .009854 .6877 .7263 .04032 .2622 1.589

\*---------------------------------------------------------------------------------\*

#### \*Who gets professional jobs: probit estimates

probit prof married chGoNowQ maleM ageM edNowQ rural polDem9

Iteration 0: log likelihood = -593.4376

Iteration 1: log likelihood = -446.90323

Iteration 2: log likelihood = -440.26127

Iteration 3: log likelihood = -440.23862

Iteration 4: log likelihood = -440.23862

Probit regression Number of obs = 1776

LR chi2(7) = 306.40

Prob > chi2 = 0.0000

Log likelihood = -440.23862 Pseudo R2 = 0.2582

------------------------------------------------------------------------------

prof | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

married | .0671814 .0951551 0.71 0.480 -.1193191 .2536819

chGoNowQ | .0021291 .0016976 1.25 0.210 -.0011981 .0054564

maleM | .1205948 .0947907 1.27 0.203 -.0651915 .3063811

ageM | .0043075 .0031777 1.36 0.175 -.0019206 .0105356

edNowQ | .3079578 .0206867 14.89 0.000 .2674125 .348503

rural | .1526669 .0996606 1.53 0.126 -.0426643 .3479981

polDem9 | -.0606992 .1381297 -0.44 0.660 -.3314283 .21003

\_cons | -6.324576 .3728632 -16.96 0.000 -7.055374 -5.593777

------------------------------------------------------------------------------

predict probHat3, pr

eval4 probit edNowQ "10 12 14 16 18 20 22"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION.

edNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

10 .002093 .00129 .0006924 .005748 0 0 1

12 .01207 .004807 .005574 .02442 .009973 .009973 5.764

14 .04986 .0133 .02909 .08124 .0378 .04777 23.82

16 .1498 .02821 .1013 .2119 .09995 .1477 71.57

18 .3343 .04646 .248 .4301 .1844 .3322 159.7

20 .5718 .05733 .4563 .681 .2376 .5697 273.2

22 .7853 .05074 .6724 .8713 .2135 .7832 375.2

\*---------------------------------------------------------------------------------\*

#### \*...logit estimates

logit prof married chGoNowQ maleM ageM edNowQ rural polDem9

eval4 logit edNowQ "10 12 14 16 18 20 22"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION.

edNowQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

10 .004608 .001784 .002289 .009283 0 0 1

12 .01498 .004621 .008438 .02655 .01037 .01037 3.251

14 .0475 .01169 .02964 .07547 .03252 .04289 10.31

16 .1399 .02778 .09409 .203 .09239 .1353 30.36

18 .3445 .05316 .2475 .4559 .2046 .3399 74.76

20 .6282 .06374 .4946 .7444 .2837 .6236 136.3

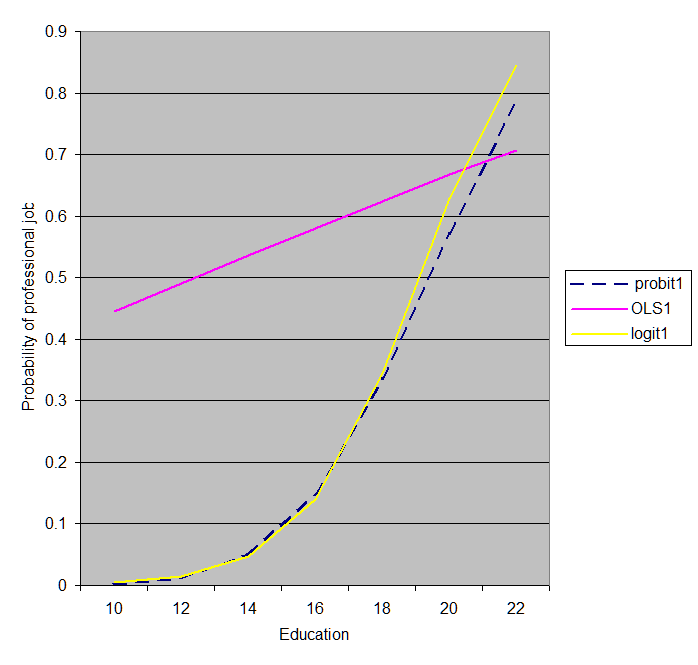
22 .8453 .04497 .7376 .9139 .2171 .8407 183.4

\*---------------------------------------------------------------------------------\*

predict lgtHat3, pr

#### \*Graph all. Note that logit and probit have a built-in curve.

\*Probit and logit are very similar, OLS is way off.



. corr prof probHat3 lgtHat3 olsHat3

(obs=1776)

| prof probHat3 lgtHat3 olsHat3

-------------+------------------------------------

prof | 1.0000

probHat3 | 0.4679 1.0000

lgtHat3 | 0.4696 0.9978 1.0000

olsHat3 | 0.4366 0.9311 0.9120 1.0000

Probit & logit clearly better here, OLS worse. But this way of putting it (with correlations) rather flatters OLS -- the graph is clearer.

### \*Case #2, cont: Dichotomous with an extreme split, but allow curvilinearity in educ.

#### \*OLS: lots of curvilinearity

regress prof married chGoNowQ maleM ageM edNowQ edNowSq rural polDem9, beta

Source | SS df MS Number of obs = 1776

-------------+------------------------------ F( 8, 1767) = 58.21

Model | 34.5654007 8 4.32067508 Prob > F = 0.0000

Residual | 131.163766 1767 .074229636 R-squared = 0.2086

-------------+------------------------------ Adj R-squared = 0.2050

Total | 165.729167 1775 .093368545 Root MSE = .27245

------------------------------------------------------------------------------

prof | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

married | .0044101 .0133726 0.33 0.742 .007205

chGoNowQ | .000441 .000265 1.66 0.096 .0367968

maleM | .0159943 .0130242 1.23 0.220 .0261122

ageM | .0006647 .0004296 1.55 0.122 .0339735

edNowQ | -.1289223 .0295059 -4.37 0.000 -.9611224

edNowSq | .0060852 .0009611 6.33 0.000 1.390788

rural | .0209353 .0136875 1.53 0.126 .0329366

polDem9 | -.0041814 .0201159 -0.21 0.835 -.0045777

\_cons | .6205234 .223192 2.78 0.005 .

------------------------------------------------------------------------------

predict olsHat4

eval4 linear edNowQ "10 12 14 16 18 20 22"

#### \*Probit: but no obvious curvilinearity here

probit prof married chGoNowQ maleM ageM edNowQ edNowSq rural polDem9

Iteration 0: log likelihood = -593.4376

Iteration 1: log likelihood = -447.42328

Iteration 2: log likelihood = -440.17369

Iteration 3: log likelihood = -440.10196

Iteration 4: log likelihood = -440.10196

Probit regression Number of obs = 1776

LR chi2(8) = 306.67

Prob > chi2 = 0.0000

Log likelihood = -440.10196 Pseudo R2 = 0.2584

------------------------------------------------------------------------------

prof | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

married | .0664265 .0952429 0.70 0.486 -.1202462 .2530992

chGoNowQ | .0021374 .0016971 1.26 0.208 -.0011888 .0054637

maleM | .1235261 .0950494 1.30 0.194 -.0627674 .3098196

ageM | .0041962 .0031898 1.32 0.188 -.0020557 .0104482

edNowQ | .4112855 .1988069 2.07 0.039 .0216311 .8009399

edNowSq | -.003224 .0061627 -0.52 0.601 -.0153027 .0088548

rural | .1554566 .0999853 1.55 0.120 -.040511 .3514241

polDem9 | -.0598454 .1382005 -0.43 0.665 -.3307133 .2110226

\_cons | -7.132624 1.592886 -4.48 0.000 -10.25462 -4.010624

------------------------------------------------------------------------------

predict prbtHat4

eval4 probit edNowQ "10 12 14 16 18 20 22"

#### \*Logit, continued. Some curvilinearity for ed here.

logit prof married chGoNowQ maleM ageM edNowQ edNowSq rural polDem9

Logistic regression Number of obs = 1776

LR chi2(8) = 311.76

Prob > chi2 = 0.0000

Log likelihood = -437.55663 Pseudo R2 = 0.2627

------------------------------------------------------------------------------

prof | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

married | .1012172 .1779149 0.57 0.569 -.2474895 .449924

chGoNowQ | .0045561 .0030532 1.49 0.136 -.0014281 .0105403

maleM | .1923622 .1773966 1.08 0.278 -.1553288 .5400531

ageM | .0095601 .0059691 1.60 0.109 -.0021392 .0212593

edNowQ | 1.47562 .4030231 3.66 0.000 .685709 2.26553

edNowSq | -.027009 .012125 -2.23 0.026 -.0507736 -.0032444

rural | .2832371 .1885153 1.50 0.133 -.0862461 .6527203

polDem9 | -.0373163 .2570127 -0.15 0.885 -.5410519 .4664194

\_cons | -19.17244 3.351953 -5.72 0.000 -25.74215 -12.60273

------------------------------------------------------------------------------

predict logitHat4

eval4 logit edNowQ "10 12 14 16 18 20 22"

#### \*Compare and graph

>> Very similar, but OLS better than before (now that it has curvature in education) but still not quite as good as logit or probit.

>> Maybe: logit a tiny bit better than probit here??

corr prof probHat3 lgtHat3 olsHat3 prbtHat4 logitHat4 olsHat4

(obs=1776)

| prof probHat3 lgtHat3 olsHat3 prbtHat4 logitH~4 olsHat4

-------------+---------------------------------------------------------------

prof | 1.0000

probHat3 | 0.4679 1.0000

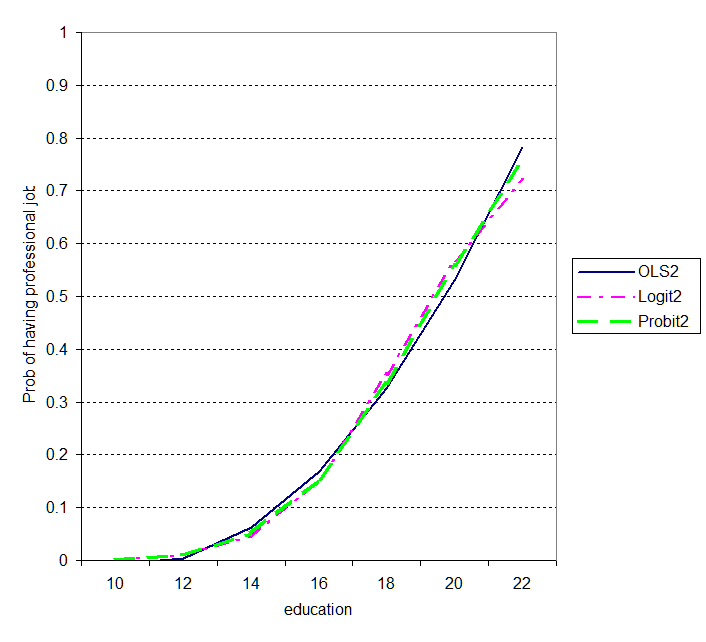
lgtHat3 | 0.4696 0.9978 1.0000

olsHat3 | 0.4366 0.9311 0.9120 1.0000

prbtHat4 | 0.4693 0.9996 0.9959 0.9381 1.0000

logitHat4 | 0.4764 0.9961 0.9927 0.9390 0.9977 1.0000

olsHat4 | 0.4567 0.9890 0.9826 0.9560 0.9890 0.9822 1.0000



## \*Week 8/9: Logit/probit continued

### \*Case #3: Vote

ptyVote

Who did you vote for in the last presidential election?

1 McCain

2 Did not vote, strongly preferred McCain

3 Did not vote, preferred McCain

4 Did not vote, had no preference

5 Did not vote, preferred Obama

6 Did not vote, strongly preferred Obama

7 Obama

tab ptyVote

ptyVote | Freq. Percent Cum.

------------+-----------------------------------

-1 | 151 6.58 6.58

1 | 708 30.85 37.43

2 | 69 3.01 40.44

3 | 63 2.75 43.18

4 | 213 9.28 52.46

5 | 126 5.49 57.95

6 | 118 5.14 63.09

7 | 847 36.91 100.00

------------+-----------------------------------

Total | 2,295 100.00

recode ptyVote (1 2 3 4= 0)(5 6 7=1)(else = .), gen(Obama )

label variable Obama "Preferred Obama (vs McCain or no preference)"

tab Obama, m

table ptyVote, c(m Obama freq )

#### \*Explore with OLS

regress Obama status married chGoNowQ maleM ageM edNowQ rural , beta

Source | SS df MS Number of obs = 1407

-------------+------------------------------ F( 8, 1398) = 8.65

Model | 16.5843216 8 2.0730402 Prob > F = 0.0000

Residual | 335.164079 1398 .239745407 R-squared = 0.0471

-------------+------------------------------ Adj R-squared = 0.0417

Total | 351.748401 1406 .250176672 Root MSE = .48964

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

status | -.0007272 .0006434 -1.13 0.259 -.0350575

-- surprising -- looks like occupation does not matter??

married | -.0313881 .0314299 -1.00 0.318 -.0311037

chGoNowQ | -.0030503 .0005113 -5.97 0.000 -.1604617

maleM | -.038294 .0263976 -1.45 0.147 -.0381426

ageM | -.0014369 .0008923 -1.61 0.108 -.0445313

edNowQ | .0114913 .0067933 1.69 0.091 .0531745

rural | -.0963936 .0274297 -3.51 0.000 -.0928691

lnFamInc | -.0080053 .0083732 -0.96 0.339 -.0293814

\_cons | .5927204 .0914003 6.48 0.000 .

------------------------------------------------------------------------------

#### \*Explore occupational status

tab status, m

Kelleys Worldwide |

Status Score | Freq. Percent Cum.

------------------------+-----------------------------------

10. Farmer | 9 0.39 0.39

14. Unskilled worker | 98 4.27 4.66

24. Semi-skilled worker | 94 4.10 8.76

26. Service (hi & low) | 195 8.50 17.25

37. Skilled worker | 183 7.97 25.23

42. Sales (hi & low) | 238 10.37 35.60

49. Clerical (hi & low) | 372 16.21 51.81

70. Technical | 311 13.55 65.36

75. Administrative | 167 7.28 72.64

100. Professional | 197 8.58 81.22

. | 431 18.78 100.00

------------------------+-----------------------------------

Total | 2,295 100.00

table status, c(m Obama freq)

Nothing here

--------------------------------------------------

Kelleys Worldwide |

Status Score | mean(Obama) Freq.

------------------------+-------------------------

10. Farmer | .4444444 9

14. Unskilled worker | .5773196 98

24. Semi-skilled worker | .4782609 94

26. Service (hi & low) | .5803109 195

37. Skilled worker | .489011 183

42. Sales (hi & low) | .517094 238

49. Clerical (hi & low) | .4932249 372

70. Technical | .5032258 311

75. Administrative | .4666667 167

100. Professional | .5025381 197

--------------------------------------------------

#### \*Look at religion again, as its promising

regress Obama married chGoNowQ fundmQ maleM ageM edNowQ rural status , beta

Source | SS df MS Number of obs = 1761

-------------+------------------------------ F( 8, 1752) = 13.13

Model | 24.894303 8 3.11178787 Prob > F = 0.0000

Residual | 415.201097 1752 .236986928 R-squared = 0.0566

-------------+------------------------------ Adj R-squared = 0.0523

Total | 440.0954 1760 .250054205 Root MSE = .48681

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

married | -.0392952 .0240634 -1.63 0.103 -.0392348

chGoNowQ | -.0027358 .0004775 -5.73 0.000 -.1396264

fundmQ | -.108006 .0264308 -4.09 0.000 -.0985007

maleM | -.058478 .0233353 -2.51 0.012 -.0583409

ageM | -.0014977 .0007727 -1.94 0.053 -.0468042

edNowQ | .005228 .0060891 0.86 0.391 .023857

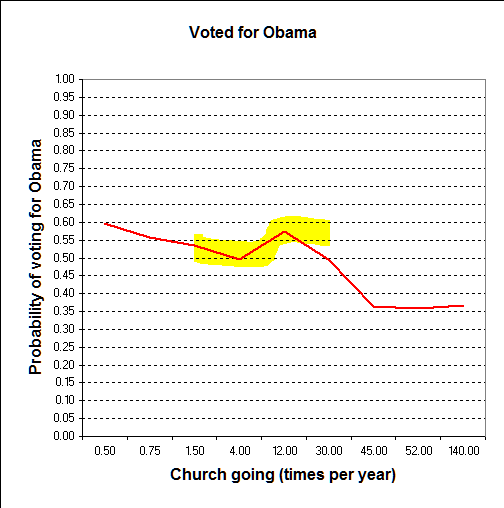
rural | -.0930365 .0243788 -3.82 0.000 -.0895144

status | -.0005337 .0005748 -0.93 0.353 -.0255669

\_cons | .6918279 .0837391 8.26 0.000 .

------------------------------------------------------------------------------

table chGoNowQ , c(m Obama freq)



\*Lets try a log

gen lnChGo = ln( chGoNowQ)

tab lnChGo, m

table chGoNowQ, c(m lnChGo sd lnChGo freq)

----------------------------------------------------

Church |

going |

(times |

per year) | mean(lnChGo) sd(lnChGo) Freq.

----------+-----------------------------------------

.5 | -.6931472 0 638

.75 | -.2876821 0 343

1.5 | .4054651 0 105

4 | 1.386294 0 300

12 | 2.484907 0 91

30 | 3.401197 0 98

45 | 3.806663 0 121

52 | 3.951244 0 364

140 | 4.941642 0 31

----------------------------------------------------

\*compare

regress Obama married chGoNowQ fundmQ maleM ageM edNowQ rural status , beta

regress Obama married lnChGo fundmQ maleM ageM edNowQ rural status , beta

Source | SS df MS Number of obs = 1761

-------------+------------------------------ F( 8, 1752) = 13.13

Model | 24.894303 8 3.11178787 Prob > F = 0.0000

Residual | 415.201097 1752 .236986928 R-squared = 0.0566

-------------+------------------------------ Adj R-squared = 0.0523

Total | 440.0954 1760 .250054205 Root MSE = .48681

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

married | -.0392952 .0240634 -1.63 0.103 -.0392348

chGoNowQ | -.0027358 .0004775 -5.73 0.000 -.1396264

fundmQ | -.108006 .0264308 -4.09 0.000 -.0985007

maleM | -.058478 .0233353 -2.51 0.012 -.0583409

ageM | -.0014977 .0007727 -1.94 0.053 -.0468042

edNowQ | .005228 .0060891 0.86 0.391 .023857

rural | -.0930365 .0243788 -3.82 0.000 -.0895144

status | -.0005337 .0005748 -0.93 0.353 -.0255669

\_cons | .6918279 .0837391 8.26 0.000 .

------------------------------------------------------------------------------

.

ln version just a tiny shade better.

Also it slightly reduces the fundamentalist effect.

. regress Obama married lnChGo fundmQ maleM ageM edNowQ rural status , beta

Source | SS df MS Number of obs = 1761

-------------+------------------------------ F( 8, 1752) = 13.42

Model | 25.4191591 8 3.17739488 Prob > F = 0.0000

Residual | 414.676241 1752 .236687352 R-squared = 0.0578

-------------+------------------------------ Adj R-squared = 0.0535

Total | 440.0954 1760 .250054205 Root MSE = .48651

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

married | -.0335144 .0241244 -1.39 0.165 -.0334629

lnChGo | -.0385657 .0065108 -5.92 0.000 -.1480154

fundmQ | -.094198 .0269749 -3.49 0.000 -.0859079

maleM | -.0630865 .0233263 -2.70 0.007 -.0629385

ageM | -.0014095 .0007736 -1.82 0.069 -.0440465

edNowQ | .0049805 .006079 0.82 0.413 .0227277

rural | -.0942817 .0243622 -3.87 0.000 -.0907125

status | -.0005477 .0005743 -0.95 0.340 -.0262372

\_cons | .68954 .0836963 8.24 0.000 .

------------------------------------------------------------------------------

predict olsHat

#### \*Now the probit version

probit Obama married lnChGo fundmQ maleM ageM edNowQ rural status

Probit regression Number of obs = 1761

LR chi2(8) = 103.79

Prob > chi2 = 0.0000

Log likelihood = -1168.4303 Pseudo R2 = 0.0425

------------------------------------------------------------------------------

Obama | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

married | -.0876795 .0631056 -1.39 0.165 -.2113643 .0360053

lnChGo | -.0994758 .0170512 -5.83 0.000 -.1328956 -.0660561

fundmQ | -.2439378 .0705319 -3.46 0.001 -.3821778 -.1056978

maleM | -.1634613 .0611786 -2.67 0.008 -.2833692 -.0435533

ageM | -.0037083 .0020261 -1.83 0.067 -.0076794 .0002628

edNowQ | .0130634 .0159141 0.82 0.412 -.0181276 .0442545

rural | -.2448727 .0637538 -3.84 0.000 -.3698279 -.1199175

status | -.00143 .0015035 -0.95 0.342 -.0043768 .0015167

\_cons | .4915773 .2190722 2.24 0.025 .0622037 .9209509

------------------------------------------------------------------------------

. eval4 probit fundmQ "0 1"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

fundmQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 .5371 .03362 .4708 .6026 0 0 1

1 .4433 .03759 .3706 .518 -.09377 -.09377 .8254

\*---------------------------------------------------------------------------------\*

-- compare with OLS.

this is a 1st difference

eval4 probit maleM "0 1"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

maleM Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 .5415 .03422 .474 .6082 0 0 1

1 .4793 .03398 .4132 .5464 -.06224 -.06224 .8851

\*---------------------------------------------------------------------------------\*

eval4 probit lnChGo " ln(0.5) ln(52)"

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

lnChGo Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

-.6931 .5828 .03489 .5133 .6501 0 0 1

3.951 .4038 .03697 .3331 .478 -.179 -.179 .6929

\*---------------------------------------------------------------------------------\*

this is a 1st difference

\*Or we can graph the whole range

eval4 probit lnChGo " ln(0.5) ln(2.5) ln(12) ln(30) ln(52) ln(140)"

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

lnChGo Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

-.6931 .5828 .03489 .5133 .6501 0 0 1

.9163 .5209 .03337 .4553 .5861 -.06196 -.06196 .8937

2.485 .46 .0343 .3934 .5278 -.0609 -.1229 .7892

3.401 .4247 .03582 .3558 .4962 -.03523 -.1581 .7288

3.951 .4038 .03697 .3331 .478 -.02089 -.179 .6929

4.942 .3669 .0393 .2926 .4467 -.03689 -.2159 .6296

\*---------------------------------------------------------------------------------\*

\*and OLS for comparison

regress Obama married lnChGo fundmQ maleM ageM edNowQ rural status, beta

eval4 linear lnChGo " ln(0.5) ln(2.5) ln(12) ln(30) ln(52) ln(140)"

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

lnChGo Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

-.6931 .5828 .03494 .5143 .6512 0 0 1

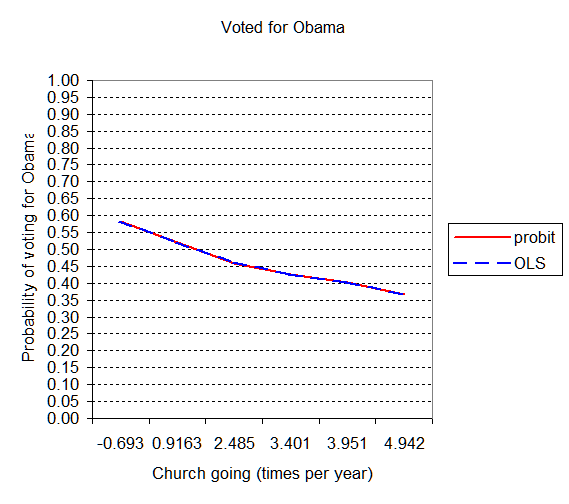
.9163 .5207 .03281 .4564 .585 -.06207 -.06207 .8935

2.485 .4602 .03389 .3938 .5266 -.06049 -.1226 .7897

3.401 .4249 .03586 .3546 .4951 -.03534 -.1579 .729

3.951 .4036 .03744 .3302 .477 -.02121 -.1791 .6926

4.942 .3654 .04093 .2852 .4457 -.0382 -.2173 .6271



#### \*Now a logit

logit Obama married lnChGo fundmQ maleM ageM edNowQ rural status

eval4 logit lnChGo " ln(0.5) ln(2.5) ln(12) ln(30) ln(52) ln(140)"

Logistic regression Number of obs = 1761

LR chi2(8) = 104.00

Prob > chi2 = 0.0000

Log likelihood = -1168.3221 Pseudo R2 = 0.0426

------------------------------------------------------------------------------

Obama | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

married | -.1418587 .1017754 -1.39 0.163 -.3413348 .0576175

lnChGo | -.1603161 .0275064 -5.83 0.000 -.2142276 -.1064046

fundmQ | -.3951919 .1138619 -3.47 0.001 -.6183571 -.1720266

maleM | -.2674947 .0989968 -2.70 0.007 -.4615248 -.0734646

ageM | -.0059722 .0032728 -1.82 0.068 -.0123868 .0004424

edNowQ | .0210661 .025743 0.82 0.413 -.0293892 .0715215

rural | -.397007 .1031344 -3.85 0.000 -.5991467 -.1948673

status | -.0023311 .0024365 -0.96 0.339 -.0071066 .0024444

\_cons | .7956253 .3549391 2.24 0.025 .0999574 1.491293

------------------------------------------------------------------------------

\*---------------------------------------------------------------------------------\*

WHOLE POPULATION STANDARDIZATION. Version Jan 2009. For help: eval4 help

lnChGo Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

-.6931 .5827 .03488 .5129 .6497 0 0 1

.9163 .5206 .03353 .4547 .5862 -.06204 -.06204 .8935

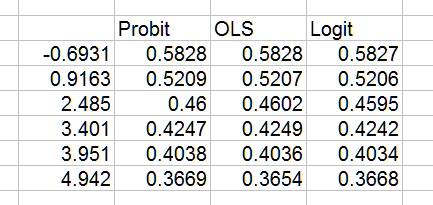
2.485 .4595 .03443 .3929 .5278 -.06111 -.1232 .7886

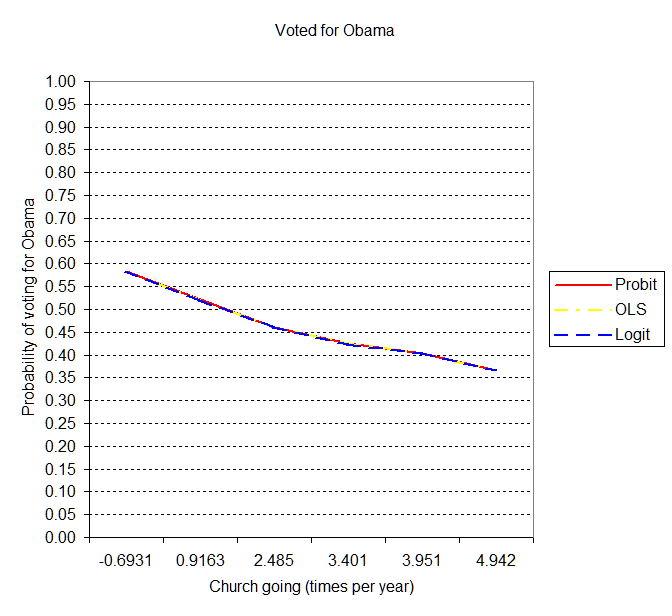
3.401 .4242 .03585 .3556 .4961 -.03527 -.1584 .7281

3.951 .4034 .0369 .3332 .4778 -.02084 -.1793 .6923

4.942 .3668 .03897 .2937 .4464 -.03665 -.2159 .6294

\*---------------------------------------------------------------------------------\*





#### \*"logistic" version rather than "logit"

logit Obama married lnChGo fundmQ maleM ageM edNowQ rural status

logistic Obama married lnChGo fundmQ maleM ageM edNowQ rural status

Logistic regression Number of obs = 1761

LR chi2(8) = 104.00

Prob > chi2 = 0.0000

Log likelihood = -1168.3221 Pseudo R2 = 0.0426

------------------------------------------------------------------------------

Obama | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

married | -.1418587 .1017754 -1.39 0.163 -.3413348 .0576175

lnChGo | -.1603161 .0275064 -5.83 0.000 -.2142276 -.1064046

fundmQ | -.3951919 .1138619 -3.47 0.001 -.6183571 -.1720266

maleM | -.2674947 .0989968 -2.70 0.007 -.4615248 -.0734646

ageM | -.0059722 .0032728 -1.82 0.068 -.0123868 .0004424

edNowQ | .0210661 .025743 0.82 0.413 -.0293892 .0715215

rural | -.397007 .1031344 -3.85 0.000 -.5991467 -.1948673

status | -.0023311 .0024365 -0.96 0.339 -.0071066 .0024444

\_cons | .7956253 .3549391 2.24 0.025 .0999574 1.491293

------------------------------------------------------------------------------

.

. logistic Obama married lnChGo fundmQ maleM ageM edNowQ rural status

Logistic regression Number of obs = 1761

LR chi2(8) = 104.00

Prob > chi2 = 0.0000

Log likelihood = -1168.3221 Pseudo R2 = 0.0426

------------------------------------------------------------------------------

Obama | Odds Ratio Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

married | .8677439 .088315 -1.39 0.163 .7108209 1.05931

lnChGo | .8518745 .023432 -5.83 0.000 .8071647 .8990608

fundmQ | .6735508 .0766918 -3.47 0.001 .538829 .8419567

maleM | .7652944 .0757617 -2.70 0.007 .6303218 .9291691

ageM | .9940456 .0032533 -1.82 0.068 .9876896 1.000442

edNowQ | 1.02129 .0262911 0.82 0.413 .9710384 1.074141

rural | .6723293 .0693403 -3.85 0.000 .5492801 .8229438

status | .9976716 .0024309 -0.96 0.339 .9929186 1.002447

\_cons | 2.215826 .7864833 2.24 0.025 1.105124 4.442837

------------------------------------------------------------------------------

To get the odds ratios, start with the predicted values:

fundmQ Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 .5369 .03365 .4706 .6025 0 0 1

1 .4427 .03758 .3703 .5176 -.09425 -.09425 .8245

\*---------------------------------------------------------------------------------\*

\*Now divide each predicted value by its compliment. this is the "odds"

. display .5369/(1 - .5369)

1.1593608

. display .4427/ (1 - .4427)

.79436569

. \*odds ratio is the ratio of one to the other

. display .79436 / 1.1593

.68520659

## \*Week 10: Multi-level analysis

### \*Data for this: Use a special file from the World Inequality Study

\*\*this is where the file is on Kelley's computer; yours will differ

\*\*cd "C:\\_ACTIVE\28\_Teaching\2012\_APST\_470670\)Stata\_PQ\Data\WIS\_extract\_XTREG"

use WIS\_extract1.dta

Data are from WIS, extracted at "\*Data for ME's regression class ^^2012\_03\_28" in file "C:\\_ACTIVE\00\_Projects\_Major\2006\_WelfState\_WLFP\_Semyonov\SPSS\Semyonov\_SPSS\_21b\_Ozy.doc"

at location ^^2012\_03\_28 .

#### \*Description of the data

VARIABLES:

>> ed is education in years

>> faocc4t is father’s occupational status, in Kelley’s worldwide scores which range from a low of zero (farm laborers) to a high of 100 (higher professionals)

>> gdp\_ppUSA is gross domestic product per capita (GDP)at parity purchasing power, as of 1995. It is scored so that the USA has a score of 1.0 and other nations are scored as a proportion of the US figure. For example, Australian GDP is about 77% of US GDP.

>> exRedq is a dummy variable, scored 1 for formerly Communist nations and zero for other countries.

\*This is what the data look like case-by-case (showing only the first 3 cases in each nation). Note how the country characteristics (GDP and exRed [and edMean, a special variable we discuss later]) are the same for each person in the country.

+-----------------------------------------------------------------------+

| countryx cID ed edMean faocc4t gdp\_pp~A exRedq |

|-----------------------------------------------------------------------|

1. | 566. niger 1 6.000 3.045194 32. Routin 0.030 0.000 |

2. | 566. niger 2 0.000 3.045194 10. Farmer 0.030 0.000 |

3. | 566. niger 3 6.000 3.045194 10. Farmer 0.030 0.000 |

501. | 356. ind: 1 0.000 3.054264 51. Higher 0.070 0.000 |

502. | 356. ind: 2 6.000 3.054264 10. Farmer 0.070 0.000 |

503. | 356. ind: 3 9.000 3.054264 10. Farmer 0.070 0.000 |

1001. | 156. chi: 1 0.000 7.504814 0. Farm wo 0.080 1.000 |

1002. | 156. chi: 2 9.000 7.504814 70. Techni 0.080 1.000 |

1003. | 156. chi: 3 6.000 7.504814 0. Farm wo 0.080 1.000 |

...etc...

17501. | 756. swz: 1 6.000 10.42014 32. Routin 0.920 0.000 |

17502. | 756. swz: 2 9.000 10.42014 10. Farmer 0.920 0.000 |

17503. | 756. swz: 3 9.000 10.42014 14. Unskil 0.920 0.000 |

18001. | 840. usa: 1 12.000 13.12176 24. Semi-s 1.000 0.000 |

18002. | 840. usa: 2 17.000 13.12176 75. Admini 1.000 0.000 |

18003. | 840. usa: 3 14.000 13.12176 100. Profe 1.000 0.000 |

+-----------------------------------------------------------------------+

sum

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

countryx | 18500 416.4865 247.7072 36 840

ed | 18500 9.777405 4.247361 0 20

gdp\_ppUSA | 18500 .5558973 .2786955 .03 1

exRedq | 18500 .2702703 .4441114 0 1

faocc4t | 18500 31.27708 24.77927 0 100

edMean | 18500 10.2035 2.45343 3.045194 13.12176

cID | 18500 250.5 144.3412 1 500

### \*A problem that multi-level analysis trys to sovle

\*Suppose we have interviewed just one person in each of our 39 nations

\*The same logic would be one rabbit in each of our 39 study sites

\*or one butterfly in each of 39 back yards

\*or one grad student in each of 39 universities

\*Suppose also that we want to study not only individual level variables (like father's occupation) but characteristics of the CONTEXT as well. Here we look at the country's GNP and whether it was formerly Communiste

\*The same logic would apply if you were studying rabbits in each of 39 study sites, and wanted to know if the density of vegitation at the study site mattered (or how exposed to preditors the site was; or the site's elevation, or the site's mean annual rainfall, etc)

\*The same logic applies to a study of butterflis in each of 39 back yards.

\*Or to grad students in each of 39 universities, where you might want to know the effect of university tuitiion levels, or faculty quality, or nearness of ski resorts).

\*In our example we have this (sorted by country GNP, just for clarity):

list countryx cID ed faocc4t gdp\_ppUSA exRedq if cID==1, sep(0)

+------------------------------------------------------+

| countryx cID ed faocc4t gdp\_pp~A exRedq |

|------------------------------------------------------|

| 566. niger 1 6.000 32 0.030 0.000 |

| 356. ind: 1 0.000 51 0.070 0.000 |

| 156. chi: 1 0.000 0 0.080 1.000 |

| 608. phl: 1 5.000 10 0.130 0.000 |

| 428. lva: 1 20.000 32 0.180 1.000 |

| 100. blg: 1 11.000 29 0.200 1.000 |

| 76. bzl: b 1 5.000 15 0.240 0.000 |

| 616. pol: 1 10.000 26 0.240 1.000 |

| 643. rus: 1 8.000 21 0.260 1.000 |

| 152. chl: 1 1.000 32 0.270 0.000 |

| 703. svk: 1 12.000 37 0.310 1.000 |

| 348. hun: 1 13.000 100 0.340 1.000 |

| 203. cze: 1 9.000 24 0.440 1.000 |

| 705. svn: 1 16.000 43 0.470 1.000 |

| 620. prt 1 6.000 10 0.490 0.000 |

| 724. esp: 1 6.000 0 0.550 0.000 |

| 196. cyp: 1 16.000 51 0.600 0.000 |

| 376. isr: 1 8.000 51 0.620 0.000 |

| 554. nzl: 1 13.000 100 0.630 0.000 |

| 372. irl: 1 16.000 47 0.630 0.000 |

| 828. nir: 1 9.000 28 0.640 0.000 |

| 246. fin: 1 6.000 10 0.680 0.000 |

| 278. g\_e: 1 8.000 100 0.700 1.000 |

| 826. gbr: 1 11.000 56 0.710 0.000 |

| 752. swe: 1 6.000 37 0.730 0.000 |

| 380. ita: 1 13.000 14 0.740 0.000 |

| 250. fra: 1 12.000 32 0.750 0.000 |

| 528. nl: 1 9.000 10 0.750 0.000 |

| 36. oz: a 1 9.000 10 0.770 0.000 |

| 40. aut: a 1 10.000 24 0.800 0.000 |

| 56. bel: b 1 6.000 24 0.810 0.000 |

| 710. sa: 1 10.000 37 0.810 0.000 |

| 280. g\_w: 1 8.000 24 0.820 0.000 |

| 124. can: 1 9.000 24 0.830 0.000 |

| 208. den: 1 6.000 14 0.840 0.000 |

| 392. jpn: 1 6.000 0 0.840 0.000 |

| 578. nor: 1 6.000 10 0.900 0.000 |

| 756. swz: 1 6.000 32 0.920 0.000 |

| 840. usa: 1 12.000 24 1.000 0.000 |

+------------------------------------------------------+ …etc…

#### \*Regression analysis with 1 case per country (ie one case per context)

\*In these next few analyses we use edMean (the mean education of the country) instead of individual ed (which is what we logically should use). We do it only to avoid unnecessary complications (to do with which 37 people are chosen) and will revert to proper individual education data in a moment.

regress edMean faocc4t gdp\_ppUSA exRedq if cID==1, beta

Source | SS df MS Number of obs = 37

-------------+------------------------------ F( 3, 33) = 11.31

Model | 112.882222 3 37.6274074 Prob > F = 0.0000

Residual | 109.820557 33 3.32789566 R-squared = 0.5069

-------------+------------------------------ Adj R-squared = 0.4620

Total | 222.702779 36 6.18618831 Root MSE = 1.8243

------------------------------------------------------------------------------

edMean | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0028785 .012148 0.24 0.814 .0295374

gdp\_ppUSA | 6.532758 1.248056 5.23 0.000 .7441518

exRedq | 3.745894 .7986921 4.69 0.000 .6780687

\_cons | 5.473454 .9226697 5.93 0.000 .

------------------------------------------------------------------------------

\*Look at the standard errors and t-tests, particularly for the country level variables GNP and formerly Communist.

### \*Regression analyses with 2 cases per country

\*The same logic would be 2 rabbits in each of our 39 study sites

\*or 2 butterflys in each of 39 back yards

\*or 2 students in each of 39 universities

This is what the data look like – still the same 39 countries, but two different people in each country:

+------------------------------------------------------+

| countryx cID ed faocc4t gdp\_pp~A exRedq |

|------------------------------------------------------|

| 566. niger 2 0.000 10 0.030 0.000 |

| 566. niger 1 6.000 32 0.030 0.000 | -- same

For everyone in the country

| 356. ind: 2 6.000 10 0.070 0.000 |

| 356. ind: 1 0.000 51 0.070 0.000 |

| 156. chi: 1 0.000 0 0.080 1.000 |

| 156. chi: 2 9.000 70 0.080 1.000 |

...etc...

| 756. swz: 2 9.000 10 0.920 0.000 |

| 756. swz: 1 6.000 32 0.920 0.000 |

| 840. usa: 2 17.000 75 1.000 0.000 |

| 840. usa: 1 12.000 24 1.000 0.000 |

+------------------------------------------------------+

\*Note how the standard error for gdp and exRed go down, and the t-tests go up. That is not right: we still have just the same 39 nations, no new ones.

\*We do have twice as many people as before, so twice as much data on father’s occupation. So the standard error on that should go down in theory).

regress edMean faocc4t gdp\_ppUSA exRedq if cID<=2, beta

Source | SS df MS Number of obs = 74

-------------+------------------------------ F( 3, 70) = 25.16

Model | 231.09089 3 77.0302967 Prob > F = 0.0000

Residual | 214.314668 70 3.06163812 R-squared = 0.5188

-------------+------------------------------ Adj R-squared = 0.4982

Total | 445.405558 73 6.101446 Root MSE = 1.7498

------------------------------------------------------------------------------

edMean | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0106121 .0077774 1.36 0.177 .1165531

gdp\_ppUSA | 6.438731 .8490233 7.58 0.000 .7334411

exRedq | 3.603552 .5476544 6.58 0.000 .6523024

\_cons | 5.30197 .6183812 8.57 0.000 .

------------------------------------------------------------------------------

### \*Regression analysis with more cases per country

#### \*Regression analysis with 10 cases per country

\*Note how yet again the standard error for gdp and exRed go down, and the t-tests go up. That is not right: we still have just the same 39 nations, no new ones.

regress edMean faocc4t gdp\_ppUSA exRedq if cID<=10, beta

Source | SS df MS Number of obs = 370

-------------+------------------------------ F( 3, 366) = 132.22

Model | 1158.25976 3 386.086586 Prob > F = 0.0000

Residual | 1068.76803 366 2.92013124 R-squared = 0.5201

-------------+------------------------------ Adj R-squared = 0.5162

Total | 2227.02779 369 6.03530567 Root MSE = 1.7088

------------------------------------------------------------------------------

edMean | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0114113 .0034851 3.27 0.001 .119365

gdp\_ppUSA | 6.426158 .3710345 17.32 0.000 .7320089

exRedq | 3.705475 .2334521 15.87 0.000 .6707521

\_cons | 5.270419 .2718891 19.38 0.000 .

------------------------------------------------------------------------------

#### \*Regression analysis with 50 cases per country

\*Note how yet again the standard error for gdp and exRed go down, and the t-tests go up. That is not right: we still have just the same 39 nations, no new ones.

regress edMean faocc4t gdp\_ppUSA exRedq if cID<=50, beta

regress edMean faocc4t gdp\_ppUSA exRedq if cID<=50, beta

Source | SS df MS Number of obs = 1850

-------------+------------------------------ F( 3, 1846) = 662.45

Model | 5772.87975 3 1924.29325 Prob > F = 0.0000

Residual | 5362.25921 1846 2.90479914 R-squared = 0.5184

-------------+------------------------------ Adj R-squared = 0.5177

Total | 11135.139 1849 6.0222493 Root MSE = 1.7043

------------------------------------------------------------------------------

edMean | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0111353 .0016013 6.95 0.000 .1141118

gdp\_ppUSA | 6.343231 .1672898 37.92 0.000 .7224084

exRedq | 3.721007 .1039541 35.79 0.000 .6735637

\_cons | 5.319962 .1198497 44.39 0.000 .

------------------------------------------------------------------------------

#### \*Regression analysis with 500 cases per country

\*Note how yet again the standard error for gdp and exRed go down, and the t-tests go up. That is not right: we still have just the same 39 nations, no new ones.

regress edMean faocc4t gdp\_ppUSA exRedq if cID<=500, beta

Source | SS df MS Number of obs = 18500

-------------+------------------------------ F( 3, 18496) = 6371.81

Model | 56592.622 3 18864.2073 Prob > F = 0.0000

Residual | 54758.7676 18496 2.96057351 R-squared = 0.5082

-------------+------------------------------ Adj R-squared = 0.5082

Total | 111351.39 18499 6.0193194 Root MSE = 1.7206

------------------------------------------------------------------------------

edMean | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0070636 .0005178 13.64 0.000 .0713413

gdp\_ppUSA | 6.419241 .053382 120.25 0.000 .7291888

exRedq | 3.740117 .0331003 112.99 0.000 .677023

\_cons | 5.403289 .0384444 140.55 0.000 .

------------------------------------------------------------------------------

### \*Multi-level analysis to fix the problem

\*The problem comes about because OLS assumes each case is independent

* So in our example it assumes that there 18,500 independent readings on gnp and exRed. That is a lot of countries! So OLS thinks there is a lot more precision that there really is with out 39 countries.
* OLS makes the right assumption about independent readings of education and father's occupation -- we do really have 36,232 separate people, so we narrow down the effects of father's occupation quite well.

\*What we need is an analysis that takes into account that we have only the 39 nations but, at the same time, that we have many thousands of individual people (or rabbits, or butterflys).

\*Multi-level models do that.

\*We will use the simplest (and most usual) model, done by the program XTREG in Stata (and done by many other similar programs, and also known by many other names in the literature!). The whole area is new and the nomemclature is quite varied.

#### \*A multi-level : simple example without explicite context variables

\*Here is where we tell XTREG that country is the "context" variable here (also called the "group" variable)

\*Other examples of context would be rabbits in each of our 39 study sites

\*or 39 back yards for our butterflys

\*or 39 universities for our students

\*the classic example in the technical literature 39 separate classrooms in a school.

\*[Minor point: We are back to the proper individual level education varible not.]

\*We start with a very simple model where are just interested in individual effects and our 37 contexts are essentially a nusance factor we need to control, not a main focus of interest.

xtreg ed faocc4t , i(countryx)

Random-effects GLS regression Number of obs = 18500

Group variable: countryx Number of groups = 37

R-sq: within = 0.1556 Obs per group: min = 500

between = 0.4260 avg = 500.0

overall = 0.1782 max = 500

Wald chi2(1) = 3411.70

corr(u\_i, X) = 0 (assumed) Prob > chi2 = 0.0000

------------------------------------------------------------------------------

ed | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

faocc4t | .0561593 .0009615 58.41 0.000 .0542748 .0580437

\_cons | 8.020908 .3374172 23.77 0.000 7.359582 8.682233

-------------+----------------------------------------------------------------

sigma\_u | 2.0389009

sigma\_e | 3.0725403

rho | .30572385 (fraction of variance due to u\_i)

------------------------------------------------------------------------------

\*Here is the corresponding analys in OLS

regress ed faocc4t , beta

Source | SS df MS Number of obs = 18500

-------------+------------------------------ F( 1, 18498) = 4011.29

Model | 59471.457 1 59471.457 Prob > F = 0.0000

Residual | 274251.898 18498 14.8260298 R-squared = 0.1782

-------------+------------------------------ Adj R-squared = 0.1782

Total | 333723.355 18499 18.0400754 Root MSE = 3.8505

------------------------------------------------------------------------------

ed | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0723589 .0011425 63.33 0.000 .4221444

\_cons | 7.514232 .0455883 164.83 0.000 .

------------------------------------------------------------------------------

\*The results are very similar, but not quite identical.

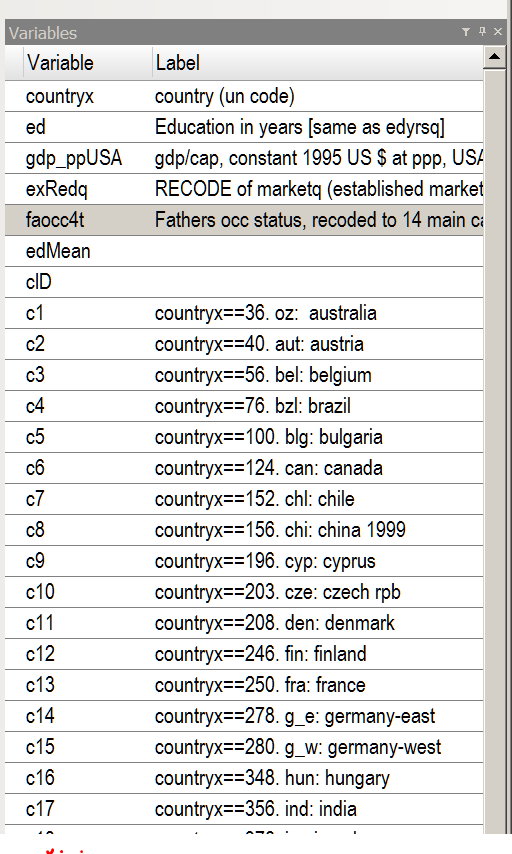
\*A main reason for that is that multi-level models like XTREG actually include hidden dummy variables, one for each country. They are what the "sigma\_u" and "rho" in the XTREG output are talking about -- saying how much the context matters.

#### \*Multi-level models compared to OLS with dummy context variables

\*Lets make dummy variables for the contexts:

tabulate countryx, gen( c )

\*Here are the first few....



\*Now the OLS using them. As you may recall, we need to leave one dummy out of the regression, to serve as the reference category. The USA is a good choice for that. So run that, and also get predicted values in the usual way:

regress ed faocc4t c1 - c36, beta

predict dumOLS

corr dumOLS faocc4t ed

display .6787^2

Source | SS df MS Number of obs = 18500

-------------+------------------------------ F( 37, 18462) = 456.44

Model | 159432.776 37 4308.99395 Prob > F = 0.0000

Residual | 174290.579 18462 9.4405037 R-squared = 0.4777

-------------+------------------------------ Adj R-squared = 0.4767

Total | 333723.355 18499 18.0400754 Root MSE = 3.0725

------------------------------------------------------------------------------

ed | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0560774 .0009614 58.33 0.000 .3271579

c1 | -2.638072 .1943276 -13.58 0.000 -.100723

c2 | -3.340939 .1943894 -17.19 0.000 -.1275589

c3 | -2.146136 .1949336 -11.01 0.000 -.0819406

c4 | -5.094428 .1944732 -26.20 0.000 -.194508

c5 | -.3853276 .1943246 -1.98 0.047 -.014712

c6 | -.1041987 .1943711 -0.54 0.592 -.0039784

c7 | -1.464446 .19448 -7.53 0.000 -.0559133

c8 | -3.821649 .1952182 -19.58 0.000 -.1459126

c9 | .6730601 .1943342 3.46 0.001 .0256978

c10 | 1.006243 .1943246 5.18 0.000 .0384189

c11 | -3.971768 .1943459 -20.44 0.000 -.1516442

c12 | -4.016245 .1944202 -20.66 0.000 -.1533424

c13 | .5062681 .1946454 2.60 0.009 .0193296

c14 | -1.296528 .1943328 -6.67 0.000 -.0495022

c15 | -2.561389 .1943342 -13.18 0.000 -.0977952

c16 | .2421259 .1943873 1.25 0.213 .0092445

c17 | -6.834983 .1943461 -35.17 0.000 -.2609633

c18 | -3.181857 .1945303 -16.36 0.000 -.121485

c19 | -1.522393 .1943248 -7.83 0.000 -.0581258

c20 | -1.366054 .1944682 -7.02 0.000 -.0521567

c21 | .943495 .1944458 4.85 0.000 .0360231

c22 | -2.024816 .1943783 -10.42 0.000 -.0773086

c23 | .1684275 .1944732 0.87 0.386 .0064307

c24 | -9.622365 .195043 -49.33 0.000 -.3673871

c25 | -3.559565 .1943331 -18.32 0.000 -.1359061

c26 | -4.727303 .1949343 -24.25 0.000 -.180491

c27 | -.6448636 .1943986 -3.32 0.001 -.0246212

c28 | -4.208124 .1943424 -21.65 0.000 -.1606684

c29 | .99883 .1943656 5.14 0.000 .0381359

c30 | 1.170987 .19433 6.03 0.000 .0447089

c31 | -.2045773 .1943261 -1.05 0.292 -.0078109

c32 | -.2162671 .1949364 -1.11 0.267 -.0082572

c33 | -.4673214 .1943355 -2.40 0.016 -.0178426

c34 | -3.490888 .1943521 -17.96 0.000 -.133284

c35 | -2.699639 .1943358 -13.89 0.000 -.1030737

c36 | -1.449891 .1944886 -7.45 0.000 -.0553576

\_cons | 9.951916 .1407413 70.71 0.000 .

------------------------------------------------------------------------------

. corr dumOLS faocc4t ed

(obs=18500)

| dumOLS faocc4t ed

-------------+---------------------------

dumOLS | 1.0000

faocc4t | 0.6108 1.0000

ed | 0.6912 0.4221 1.0000

. display .6912^2

.47775744 --compare that to the regression R-square for the OLS.

\*Lets go back to the multi-level XTREG again, and compare it with the OLS with dummy variab les.

\*note the lower r-squared: this is just the "xb" part of the prediction,like OLS without dummys

xtreg ed faocc4t , i(countryx)

predict dumMLxb, xb

\*Here is the alternative predicted value which includes the hidden dummys:

predict dumMLxbu, xbu

Random-effects GLS regression Number of obs = 18500

Group variable: countryx Number of groups = 37

R-sq: within = 0.1556 Obs per group: min = 500

between = 0.4260 avg = 500.0

overall = 0.1782 max = 500

Wald chi2(1) = 3411.70

corr(u\_i, X) = 0 (assumed) Prob > chi2 = 0.0000

------------------------------------------------------------------------------

ed | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

faocc4t | .0561593 .0009615 58.41 0.000 .0542748 .0580437

\_cons | 8.020908 .3374172 23.77 0.000 7.359582 8.682233

-------------+----------------------------------------------------------------

sigma\_u | 2.0389009

sigma\_e | 3.0725403

rho | .30572385 (fraction of variance due to u\_i)

------------------------------------------------------------------------------

.

. predict dumMLxb, xb

\*Here is the alternative predicted value which includes the hidden dummys:

predict dumMLxbu, xbu

\*Now compare multi-level R2 with OLS model with dummy variables

corr dumOLS dumMLxb dumMLxbu faocc4t ed , m

(obs=18500)

Variable | Mean Std. Dev. Min Max

-------------+----------------------------------------------------

dumOLS | 9.777405 2.93572 .3295507 16.73065

dumMLxb | 9.777405 1.391585 8.020908 13.63683

dumMLxbu | 9.777405 2.927156 .3631632 16.7221

faocc4t | 31.277 24.779 0.000 100.000

ed | 9.777 4.247 0.000 20.000

| dumOLS dumMLxb dumMLxbu faocc4t ed

-------------+---------------------------------------------

dumOLS | 1.0000

dumMLxb | 0.6108 1.0000

dumMLxbu | 1.0000 0.6125 1.0000

faocc4t | 0.6108 1.0000 0.6125 1.0000

ed | 0.6912 0.4221 0.6912 0.4221 1.0000

\*Now look at the R-squared from XTREG, the version that takes into account the hidden dummys. Compare it to the OLS with dummys.

disp .6912^2

.47775744

### \*Multi-level models including variables describing the context

\*Here is the obvious OLS version -- we just treat the context like any other variable.

\*As we noted before, the same logic would apply if you were studying rabbits in each of 39 study sites, and wanted to know if the density of vegitation at the study site mattered (or how exposed to preditors the site was; or the site's elevation, or the site's mean annual rainfall, etc). You would just make explicit variables for density of vegitation (etc), attach them to the rabbit data (like we attached country GNP to each person in our data), and include them in the analysis along with individual level data.

\*The same logic applies to a study of butterflis in each of 39 back yards.

\*Or to grad students in each of 39 universities, where you might want to know the effect of university tuitiion levels, or faculty quality, or nearness of ski resorts).

regress ed faocc4t gdp\_ppUSA exRedq, beta

Source | SS df MS Number of obs = 18500

-------------+------------------------------ F( 3, 18496) = 2796.68

Model | 104141.436 3 34713.8119 Prob > F = 0.0000

Residual | 229581.92 18496 12.4125173 R-squared = 0.3121

-------------+------------------------------ Adj R-squared = 0.3119

Total | 333723.355 18499 18.0400754 Root MSE = 3.5231

------------------------------------------------------------------------------

ed | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0660016 .0010602 62.25 0.000 .3850557

gdp\_ppUSA | 4.775986 .1093042 43.69 0.000 .3133819

exRedq | 3.904958 .0677757 57.62 0.000 .4083091

\_cons | 4.002717 .0787181 50.85 0.000 .

------------------------------------------------------------------------------

\*The main problem with OLS is that the estimates of the standard error (and therefdore the t-tests) for the contextual variables are way off. XTREG is closer to the truth:

xtreg ed faocc4t gdp\_ppUSA exRedq, i(countryx)

Random-effects GLS regression Number of obs = 18500

Group variable: countryx Number of groups = 37

R-sq: within = 0.1561 Obs per group: min = 500

between = 0.5504 avg = 500.0

overall = 0.3065 max = 500

Wald chi2(3) = 3467.48

corr(u\_i, X) = 0 (assumed) Prob > chi2 = 0.0000

------------------------------------------------------------------------------

ed | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

faocc4t | .0560413 .0009616 58.28 0.000 .0541567 .057926

gdp\_ppUSA | 5.785671 1.059447 5.46 0.000 3.709193 7.862148

exRedq | 4.20978 .6788872 6.20 0.000 2.879186 5.540375

\_cons | 3.670579 .7455203 4.92 0.000 2.209386 5.131772

-------------+----------------------------------------------------------------

sigma\_u | 1.5853612

sigma\_e | 3.0709379

rho | .210429 (fraction of variance due to u\_i)

------------------------------------------------------------------------------

## \*Week 11: Is equal interval scoring correct?

\*We resume working with our stem cell data, not the World Inequality Data we used last week. So run Sections #2 and #3 first (as usual in earlier weeks). Then continue here.

### \*Example #1: Curvilinear effects

\*We have already seen that some effects are curvilinear. That means that they are NOT equal interval. Here is one of our previous examples:

\*This is the effect of education on family income:



\*(1) going from 10 years of education to 11 years does not increase income at all, either in the raw data (the black line) or in the curve (dashed line) which is pretty much the predicted value from using both education in years and its square (as we saw previously)!

\*(2) going from 11 to 12 does increases it a little

\*(3) going from (say) 14 to 15 increases it even more

\*(4) going from 19 to 20 is even better

\*So here we would NOT want to score education in years (an equal interval scoring) alone. But by including education squared, i.e. education^2, we are OK. That allows for unequal intervals in a reasonable and predictable way.

### \*Example #2: Party and vote

#### \*Vote

\*First we will look at who voted for Obama and who for McCain, ignoring those who did not vote. Keep it simple!

ptyVote Who did you vote for in the last presidential election?

McCain

-1 no answer

1 McCain

2 Did not vote, strongly preferred McCain

3 Did not vote, preferred McCain

4 Did not vote, had no preference

5 Did not vote, preferred Obama

6 Did not vote, strongly preferred Obama

7 Obama

ptyVote | Freq. Percent Cum.

------------+-----------------------------------

-1 | 151 6.58 6.58

1 | 708 30.85 37.43

2 | 69 3.01 40.44

3 | 63 2.75 43.18

4 | 213 9.28 52.46

5 | 126 5.49 57.95

6 | 118 5.14 63.09

7 | 847 36.91 100.00

------------+-----------------------------------

Total | 2,295 100.00

recode ptyVote (7=1)(1=0)(else = .), gen(Obama)

\*Check that we have done what we wanted to do

tab Obama, m

tab ptyVote Obama, m

#### \*Party ID

ptyYou In politics, do you usually think of yourself as a...

-1 no answer

1 Strong Republican

2 Republican

3 Independent, leaning Republican

4 Independent

5 Independent, leaning Democrat

6 Democrat

7 Strong Democrat

8 [ Other party ]

tabulation of ptyYou

ptyYou | Freq. Percent Cum.

------------+-----------------------------------

-1 | 194 8.45 8.45

1 | 152 6.62 15.08

2 | 343 14.95 30.02

3 | 204 8.89 38.91

4 | 415 18.08 56.99

5 | 256 11.15 68.15

6 | 442 19.26 87.41

7 | 168 7.32 94.73

8 | 121 5.27 100.00

------------+-----------------------------------

Total | 2,295 100.00

\*the issue is how to deal with party ID in a sensible way. The complications have especially to do with "strong" versus "normal" strength of identification with a party, and what to do with independents, and especially those "leaning" one way or another

#### \*Conventional Equal interval scoring of Party ID

recode ptyYou (1=0)(2=17)(3=33)(4=50)(5=67)(6=83)(7=100)( else = .), gen( ptyDem)

tab ptyDem, m

tab ptyYou ptyDem, m

##### \*An alternate way of getting there

\*This is an alternative way of getting the same 0 to 100 scoring. We mention it here because it is an easier way of doing things later on.

\*Think back to elementary arithmetic: (1) we get rid of missing data and the odds and sods -- here "other" parties; (2) we rescore things to the bottom is zero (in the numerator); (3) we make the range from 0 to 1 (in the denominator); then (just for convenience -- it does not really matter) (4) we multiply by 100 so the range is from 0 to 100, rather than 0 to one.

recode ptyYou (-1 8 = .), gen( ptyDem2)

tab ptyDem2, m

tab ptyYou ptyDem2, m

\*Now for the serious step: get it into the range 0 to 1.

This is what we have so far: In politics, do you usually think of yourself as a...

1 Strong Republican -- the LOW score

2 Republican

3 Independent, leaning Republican

4 Independent

5 Independent, leaning Democrat

6 Democrat

7 Strong Democrat -- the HIGH score

This is the logic:

new score = (old score - LOW) / ( HIGH - LOW)

replace ptyDem2 = (ptyDem2 - 1) / (7 - 1)

tab ptyDem2

\*For convenience: Convert to range 0 to 100, and round

replace ptyDem2 = ptyDem2 \* 100

tab ptyDem2

replace ptyDem2 = round( ptyDem2, 1)

tab ptyDem2

\*Now see how it compares to our original ptyDem obtained via a recode:

corr ptyDem ptyDem2, m

The two versions are the same

(obs=1980)

Variable | Mean Std. Dev. Min Max

-------------+----------------------------------------------------

ptyDem | 52.50051 29.95217 0 100

ptyDem2 | 52.50051 29.95217 0 100

| ptyDem ptyDem2

-------------+------------------

ptyDem | 1.0000

ptyDem2 | 1.0000 1.0000 -- they are the same

#### \*Is the equal interval assumption justified? Criterion: effects on Vote?

table ptyDem, c(m Obama freq)

\*Graph it in Excel, in the usual way (and add a linear trend line in red):



The increases are NOT all the same. So its not equal interval

table ptyDem, c(m Obama freq)

------------------------------------

RECODE of |

ptyYou | mean(Obama) Freq.

----------+-------------------------

0 | .0211268 152

17 | .1134752 343 -- increases .11 - .02 = .09

33 | .1446541 204 -- increases .14 - .11 = .03

50 | .5689655 415 -- increases .57 - .14 = .43

67 | .8894472 256 -- increases .89 - .57 = .32

83 | .8912387 442 -- increases .89 - .80 = .00

100 | .9798658 168 -- increases .98 - .89 = .09

------------------------------------

Categories:

Equal Original

intvl: qnr scores:

ptyDem ptyYou

0 1 Strong Republican -- the LOW score

17 2 Republican

33 3 Independent, leaning Republican

50 4 Independent

67 5 Independent, leaning Democrat

83 6 Democrat

100 7 Strong Democrat -- the HIGH score

#### \*Effect proportional scoring

\*Score it according to the means. We do this starting with the original questionnaire scoring, in ptyYou

\*We could also do it with ptyDem, but using the original is simpler. (There is also a tricky bit: Stata is clumsy about rounding and recodes of computed variables sometimes fall afoul of that. So if we use ptyDem we have to be sure to ROUND the answers (as we did). Tedious!)

recode ptyYou (1=.02)(2= .11)(3=.14)(4=.57)(5=.889)(6=.891)(7=.98)(else = .), gen( ptyEp )

tab ptyEp

RECODE of |

ptyYou | Freq. Percent Cum.

------------+-----------------------------------

.02 | 152 7.68 7.68

.11 | 343 17.32 25.00

.14 | 204 10.30 35.30

.57 | 415 20.96 56.26

.889 | 256 12.93 69.19

.891 | 442 22.32 91.52

.98 | 168 8.48 100.00

------------+-----------------------------------

Total | 1,980 100.00

\*Now for clarity, rescore this so that it ranges from exactly 0 to exactly 100. This is easy using the arithmetical tricks we just described:

replace ptyEp = ( ptyEp - .02) / (.98 - .02)

replace ptyEp = ptyEp \* 100

replace ptyEp = round( ptyEp, 1)

tab ptyEp

RECODE of |

ptyYou | Freq. Percent Cum.

------------+-----------------------------------

0 | 152 7.68 7.68

9 | 343 17.32 25.00

13 | 204 10.30 35.30

57 | 415 20.96 56.26

91 | 698 35.25 91.52

100 | 168 8.48 100.00

------------+-----------------------------------

Total | 1,980 100.00

##### \*Compare the equal interval and effect proportional versions

table ptyYou, c(m ptyDem m ptyEp m Obama freq)

------------------------------------------------------------------

ptyYou | mean(ptyDem) mean(ptyEp) mean(Obama) Freq.

----------+-------------------------------------------------------

-1 | .6875 194

1 | 0 0 .0211268 152

2 | 17 9 .1134752 343

3 | 33 13 .1446541 204

4 | 50 57 .5689655 415

5 | 67 91 .8894472 256

6 | 83 91 .8912387 442

7 | 100 100 .9798658 168

8 | .5862069 121

------------------------------------------------------------------

>>The means and standard deviations are not quite the same

>>The two versions are very highly correlated (r=.96)

>>The effect proportional version is slightly better for predicting Obama vote

corr ptyDem ptyEp Obama , m

(obs=1494)

Variable | Mean Std. Dev. Min Max

-------------+----------------------------------------------------

ptyDem | 51.77175 31.63156 0 100

ptyEp | 54.18942 39.20928 0 100

Obama | .54083 .498497 0 1

| ptyDem ptyEp Obama

-------------+---------------------------

ptyDem | 1.0000

ptyEp | 0.9637 1.0000

Obama | 0.7228 0.7497 1.0000

\*Regressions using the two versions are very similar.

\*But the effect proportional version is slightly better.

EQUAL INTERVAL

regress Obama ptyDem famInc female status chGoNow2Q edNow2Q, b

Source | SS df MS Number of obs = 1012

-------------+------------------------------ F( 6, 1005) = 196.32

Model | 135.790218 6 22.631703 Prob > F = 0.0000

Residual | 115.857015 1005 .115280612 R-squared = 0.5396

-------------+------------------------------ Adj R-squared = 0.5369

Total | 251.647233 1011 .248909232 Root MSE = .33953

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

ptyDem | .0111973 .0003535 31.68 0.000 .7063943

famInc | -.0000822 .000194 -0.42 0.672 -.0095321

female | -.0060539 .0214296 -0.28 0.778 -.0060667

status | -.0004447 .0005107 -0.87 0.384 -.0217674

chGoNow2Q | -.0220379 .0056923 -3.87 0.000 -.086351

edNow2Q | .0074293 .0054596 1.36 0.174 .0343775

\_cons | -.0842363 .0736873 -1.14 0.253 .

------------------------------------------------------------------------------

EFFECT PROPORTIONAL

. regress Obama ptyEp famInc female status chGoNow2Q edNow2Q, b

Source | SS df MS Number of obs = 1012

-------------+------------------------------ F( 6, 1005) = 234.47

Model | 146.786333 6 24.4643888 Prob > F = 0.0000

Residual | 104.8609 1005 .104339204 R-squared = 0.5833

-------------+------------------------------ Adj R-squared = 0.5808

Total | 251.647233 1011 .248909232 Root MSE = .32302

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

ptyEp | .0094183 .0002703 34.84 0.000 .74183

famInc | -.0000692 .0001845 -0.37 0.708 -.0080216

female | -.0157304 .0204044 -0.77 0.441 -.0157638

status | -.0000449 .0004863 -0.09 0.926 -.0021981

chGoNow2Q | -.0185406 .0054255 -3.42 0.001 -.0726477

edNow2Q | .0040271 .0051978 0.77 0.439 .0186346

\_cons | .0110412 .0693061 0.16 0.873 .

------------------------------------------------------------------------------

#### \*Digression: One rough and ready of dealing with missing data

\*The mean Obama vote we just looked at gives a rough and ready way of treating missing data (the -1 codes in the questionnaire).

\*And also a very rough and ready way of dealing with minor parties (the "8" codes in the questionnaire).

\*Here are the means again:

table ptyYou, c(m ptyDem m ptyEp m Obama freq)

------------------------------------------------------------------

ptyYou | mean(ptyDem) mean(ptyEp) mean(Obama) Freq.

----------+-------------------------------------------------------

-1 | .6875 194

1 | 0 0 .0211268 152

2 | 17 9 .1134752 343

3 | 33 13 .1446541 204

4 | 50 57 .5689655 415

5 | 67 91 .8894472 256

6 | 83 91 .8912387 442

7 | 100 100 .9798658 168

8 | .5862069 121

------------------------------------------------------------------

It puts them both on the Obama side of neutral:

table ptyYou, c(m ptyDem m ptyEp m Obama freq)

------------------------------------------------------------------

ptyYou | mean(ptyDem) mean(ptyEp) mean(Obama) Freq.

----------+-------------------------------------------------------

-1 | .6875 194

1 | 0 0 .0211268 152

2 | 17 9 .1134752 343

3 | 33 13 .1446541 204

4 | 50 57 .5689655 415

8 | .5862069 121

-1 | .6875 194

5 | 67 91 .8894472 256

6 | 83 91 .8912387 442

7 | 100 100 .9798658 168

8 | .5862069 121

------------------------------------------------------------------

\*Scoring them this way is OK for analysis of Obama voting, but not necessarily for other analyses.

\*Moreover, we should really add in some random error.

\*We will deal more systematically with missing data some other time.

#### \*Dummy variable alternative to effect proportional scoring

Categories:

Equal Original

intvl: qnr scores:

ptyDem ptyYou

0 1 Strong Republican -- the LOW score

17 2 Republican

33 3 Independent, leaning Republican

50 4 Independent

67 5 Independent, leaning Democrat

83 6 Democrat

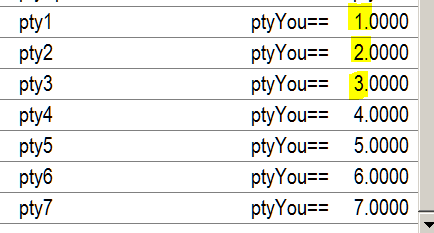
100 7 Strong Democrat -- the HIGH score

\*A Stata trick for generating a whole set of (poorly named) dummy variables:

\*The "if" clause excludes -1 answers and 8 answers , so the dummies are all defined as missing for them.

tab ptyYou if ptyYou != -1 & ptyYou != 8 , gen( pty)

\*The new variables:



ptyYou== |

1.0000 | Freq. Percent Cum.

------------+-----------------------------------

0 | 1,828 79.65 79.65

1 | 152 6.62 86.27

. | 315 13.73 100.00

------------+-----------------------------------

Total | 2,295 100.00

\*A regression using the dummies.

\*We take pty1 as the "omitted" (or "reference") category, so everything will be relative to strong Republicans

\*What we choose as the "omitted" category does not matter in one sense (the predicted values remain the same) but does matter in another (the actual coefficients differ, although *differences* between coefficients remain the same). In the big picture it does not matter, except for being a fertile source of confusion!

regress Obama famInc female status chGoNow2Q edNow2Q pty2 pty3 pty4 pty5 pty6 pty7, b

predict wDummys

Source | SS df MS Number of obs = 1012

-------------+------------------------------ F( 11, 1000) = 128.12

Model | 147.199862 11 13.3818057 Prob > F = 0.0000

Residual | 104.447371 1000 .104447371 R-squared = 0.5849

-------------+------------------------------ Adj R-squared = 0.5804

Total | 251.647233 1011 .248909232 Root MSE = .32318

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

famInc | -.0000744 .000185 -0.40 0.688 -.0086211

female | -.0175991 .0205684 -0.86 0.392 -.0176364

status | -.0000947 .0004908 -0.19 0.847 -.0046347

chGoNow2Q | -.019407 .0054503 -3.56 0.000 -.0760424

edNow2Q | .0043382 .0052497 0.83 0.409 .020074

pty2 | .1041071 .0406313 2.56 0.011 .0831356

pty3 | .0767022 .0457433 1.68 0.094 .0474917

pty4 | .5047533 .0433283 11.65 0.000 .35966

pty5 | .8599569 .0439483 19.57 0.000 .5971851

pty6 | .8514017 .0407545 20.89 0.000 .7110721

pty7 | .9420751 .0480753 19.60 0.000 .5457559

\_cons | .0183116 .0766335 0.24 0.811 .

------------------------------------------------------------------------------

The coefficients for the dummy variables are very close to the mean Obama vote, which is what we used in creating the effect proportional scores:

table ptyYou, c(m ptyDem m ptyEp m Obama freq)

------------------------------------------------------------------

ptyYou | mean(ptyDem) mean(ptyEp) mean(Obama) Freq.

----------+-------------------------------------------------------

1 | 0 0 .0211268 152

2 | 17 9 .1134752 343

3 | 33 13 .1446541 204

4 | 50 57 .5689655 415

5 | 67 91 .8894472 256

6 | 83 91 .8912387 442

7 | 100 100 .9798658 168

8 | .5862069 121

------------------------------------------------------------------

\*and compare this with our effect proportional scaling

regress Obama ptyEp famInc female status chGoNow2Q edNow2Q, b

predict wEp

Source | SS df MS Number of obs = 1012

-------------+------------------------------ F( 6, 1005) = 234.47

Model | 146.786333 6 24.4643888 Prob > F = 0.0000

Residual | 104.8609 1005 .104339204 R-squared = 0.5833

-------------+------------------------------ Adj R-squared = 0.5808

Total | 251.647233 1011 .248909232 Root MSE = .32302

------------------------------------------------------------------------------

Obama | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

ptyEp | .0094183 .0002703 34.84 0.000 .74183

famInc | -.0000692 .0001845 -0.37 0.708 -.0080216

female | -.0157304 .0204044 -0.77 0.441 -.0157638

status | -.0000449 .0004863 -0.09 0.926 -.0021981

chGoNow2Q | -.0185406 .0054255 -3.42 0.001 -.0726477

edNow2Q | .0040271 .0051978 0.77 0.439 .0186346

\_cons | .0110412 .0693061 0.16 0.873 .

------------------------------------------------------------------------------

\*The predicted values are essentially identical

. corr wDummys wEp Obama

(obs=1012)

| wDummys wEp Obama

-------------+---------------------------

wDummys | 1.0000

wEp | 0.9986 1.0000

Obama | 0.7648 0.7637 1.0000

#### \*Probit based scoring

\*make a version of party with no missing data or minor parties

gen ptyYou2 = ptyYou if ptyYou != -1 & ptyYou != 8

(315 missing values generated)

. tab ptyYou2

ptyYou2 | Freq. Percent Cum.

------------+-----------------------------------

1 | 152 7.68 7.68

2 | 343 17.32 25.00

3 | 204 10.30 35.30

4 | 415 20.96 56.26

5 | 256 12.93 69.19

6 | 442 22.32 91.52

7 | 168 8.48 100.00

------------+-----------------------------------

Total | 1,980 100.00

\*Now run an ordinal probit. It assumes the categories are in order from low to high, or vice versa, but makes no assumption about the intervals between them. In particular, no equal interval assumption.

oprobit ptyYou2 Obama famInc female status chGoNow2Q edNow2Q

Iteration 0: log likelihood = -1914.3736

Iteration 1: log likelihood = -1587.2114

Iteration 2: log likelihood = -1581.0807

Iteration 3: log likelihood = -1581.052

Iteration 4: log likelihood = -1581.052

Ordered probit regression Number of obs = 1012

LR chi2(6) = 666.64

Prob > chi2 = 0.0000

Log likelihood = -1581.052 Pseudo R2 = 0.1741

------------------------------------------------------------------------------

ptyYou2 | Coef. Std. Err. z P>|z| [95% Conf. Interval]

-------------+----------------------------------------------------------------

Obama | 1.957447 .0836218 23.41 0.000 1.793551 2.121342

famInc | -.0009279 .0006053 -1.53 0.125 -.0021143 .0002585

female | .1129205 .0659989 1.71 0.087 -.0164349 .242276

status | .0008971 .0015855 0.57 0.572 -.0022104 .0040046

chGoNow2Q | -.0500002 .0176868 -2.83 0.005 -.0846657 -.0153348

edNow2Q | -.0079875 .0168609 -0.47 0.636 -.0410342 .0250593

-------------+----------------------------------------------------------------

/cut1 | -1.027648 .2274682 -1.473477 -.5818184

/cut2 | .0196208 .2241549 -.4197148 .4589564

/cut3 | .4906342 .2249439 .0497521 .9315162

/cut4 | 1.156692 .2279769 .709866 1.603519

/cut5 | 1.711913 .2299921 1.261137 2.162689

/cut6 | 2.793545 .2357246 2.331533 3.255557

------------------------------------------------------------------------------

predict hat\_pbt, xb

. table ptyYou2, c(m hat\_pbt freq)

----------------------------------------

ptyYou2 | mean(hat\_pbt) Freq.

----------+-----------------------------

1 | -.1713569 152

2 | .0695712 343

3 | .0430627 204

4 | .9542851 415

5 | 1.657552 256

6 | 1.643283 442

7 | 1.822726 168

----------------------------------------

\*Make a 1st version of probit scoring (right intervals but not yet 0 to 100)

recode ptyYou2 (1=-.17)(2=.069)(3=.043)(4=.954)(5=1.657)(6=1.64)(7=1.82)(else= .), gen( rawPbt)

. tab rawPbt

RECODE of |

ptyYou2 | Freq. Percent Cum.

------------+-----------------------------------

-.17 | 152 7.68 7.68

.043 | 204 10.30 17.98

.069 | 343 17.32 35.30

.954 | 415 20.96 56.26

1.64 | 442 22.32 78.59

1.657 | 256 12.93 91.52

1.82 | 168 8.48 100.00

------------+-----------------------------------

Total | 1,980 100.00

##### \*now make a version with same intervals, but in range 0 to 100 in the usual way

gen adjPbt = (rawPbt + .17) / ( 1.82 + .17)

\*don't let the minus signs confuse you: remember that subtracting -.17 is same as adding .17 ie - -.17 is +.17

replace adjPbt = adjPbt \* 100

replace adjPbt = round( adjPbt, 1)

. tab adjPbt

adjPbt | Freq. Percent Cum.

------------+-----------------------------------

0 | 152 7.68 7.68

11 | 204 10.30 17.98

12 | 343 17.32 35.30

56 | 415 20.96 56.26

91 | 442 22.32 78.59

92 | 256 12.93 91.52

100 | 168 8.48 100.00

------------+-----------------------------------

Total | 1,980 100.00

\*How does this compare with the other scorings:

. table ptyYou, c(m adjPbt m ptyDem m ptyEp m Obama freq)

--------------------------------------------------------------------------------

ptyYou | mean(adjPbt) mean(ptyDem) mean(ptyEp) mean(Obama) Freq.

----------+---------------------------------------------------------------------

-1 | .6875 194

1 | 0 0 0 .0211268 152

2 | 12 17 9 .1134752 343

3 | 11 33 13 .1446541 204

4 | 56 50 57 .5689655 415

5 | 92 67 91 .8894472 256

6 | 91 83 91 .8912387 442

7 | 100 100 100 .9798658 168

8 | .5862069 121

--------------------------------------------------------------------------------

. corr adjPbt ptyDem ptyEp Obama , m

(obs=1494)

Variable | Mean Std. Dev. Min Max

-------------+----------------------------------------------------

adjPbt | 54.52075 38.92229 0 100

ptyDem | 51.77175 31.63156 0 100

ptyEp | 54.18942 39.20928 0 100

Obama | .54083 .498497 0 1

| adjPbt ptyDem ptyEp Obama

-------------+------------------------------------

adjPbt | 1.0000

ptyDem | 0.9599 1.0000

ptyEp | 0.9993 0.9637 1.0000

Obama | 0.7493 0.7228 0.7497 1.0000

. \*So we are pretty much the same as effect proportional scoring.

## \*Week 12: Direct and indirect effects, path models

### \*Data for this: Another special file from the World Inequality Study

\*\*this is where the file is on Kelley's computer; yours will differ

\*\*cd "C:\\_ACTIVE\28\_Teaching\2012\_APST\_470670\)Stata\_PQ\Data\WIS\_extract\_XTREG"

use WIS\_extract2\_stata10.dta, replace

Data are from WIS, extracted at "\*Data for ME's regression class ^^2012\_03\_28" in file "C:\\_ACTIVE\00\_Projects\_Major\2006\_WelfState\_WLFP\_Semyonov\SPSS\Semyonov\_SPSS\_21b\_Ozy.doc"

at location ^^2012\_03\_28 .

#### \*Description of the data

VARIABLES:

>> ed is education in years

>> pnteducq is the average of mother's and father's education, in years.

>> faocc4t is father’s occupational status, in Kelley’s worldwide scores which range from a low of zero (farm laborers) to a high of 100 (higher professionals)

>> occs4q is respondent's occupational status, again in Kelley’s worldwide scores which range from a low of zero (farm laborers) to a high of 100 (higher professionals)

>> lnFamInc is family income, natural log thereof. (Income analyses often work better when the DV is logged rather than in dollars. We can then get back to the original units by exp( predicted values). The original units here are what we call "minimum incomes" -- the ratio of respondent's family income to the mean income of semi-skilled male workers in the same country (ie relative rather than absolute income) . Thus, for example, an American with twice the income of a semi-skilled worker in the USA would be treated as having two "minimum incomes" (as we call them), as would someone in Brazil who earns twice the income of a semi-skilled Brazilian worker. This despite the fact that the American wage scale is much higher than Brazil's, so the American could buy much more than the Brazilian on the world market.

>> For simplicity, we have tossed out all missing data.

>> And we use 500 cases per nation just to avoid worrying about disproportionate influence of countries with larger samples.

Some description:

table countryx, c(m gdp\_ppUSA m pnteducq m faocc4t m occs4q freq)

-------------------------------------------------------------------------------------------

country (un code) | mean(gdp\_~A) mean(pnte~q) mean(fao~4t) mean(occs4q) Freq.

-------------------------+-----------------------------------------------------------------

36. oz: australia | 0.770 8.289 31.052 42.654 500

40. aut: austria | 0.800 6.768 26.648 35.610 500

76. bzl: brazil | 0.240 3.196 23.314 36.262 500

100. blg: bulgaria | 0.200 6.862 29.708 40.834 500

124. can: canada | 0.830 12.381 46.750 54.812 500

152. chl: chile | 0.270 6.246 29.044 36.401 481

156. chi: china 1999 | 0.095 2.030 11.130 21.730 500

196. cyp: cyprus | 0.600 8.044 33.030 49.248 500

203. cze: czech rpb | 0.440 9.421 34.552 42.942 500

208. den: denmark | 0.840 6.571 30.604 38.604 500

246. fin: finland | 0.680 5.890 25.908 32.442 500

250. fra: france | 0.750 8.369 43.582 62.187 500

278. g\_e: germany-east | 0.700 7.699 34.813 47.766 166

280. g\_w: germany-west | 0.820 7.881 33.964 43.468 500

348. hun: hungary | 0.340 8.584 27.246 41.834 500

356. ind: india | 0.070 3.372 27.138 28.900 500

376. isr: israel | 0.620 7.796 47.396 47.124 500

392. jpn: japan | 0.840 7.876 23.150 32.108 500

428. lva: latvia | 0.180 10.421 38.570 48.403 500

528. nl: netherlands | 0.750 7.642 37.076 49.070 500

554. nzl: new zealand | 0.630 10.421 44.472 51.105 500

566. nigeria | 0.030 0.412 14.428 18.180 500

578. nor: norway | 0.900 6.671 31.412 39.792 500

608. phl: philippines | 0.130 7.556 24.378 42.022 500

616. pol: poland | 0.240 7.679 26.608 41.360 500

620. prt portugal | 0.490 4.246 30.452 39.989 500

643. rus: russia | 0.260 8.526 39.022 56.466 500

703. svk: slovakia | 0.310 9.220 36.612 47.224 500

705. svn: slovenia | 0.470 8.277 33.695 44.858 423

724. esp: spain | 0.550 6.350 30.753 39.177 469

752. swe: sweden | 0.730 6.661 34.444 43.532 500

756. swz: switzerland | 0.920 8.321 29.738 44.190 500

840. usa: united states | 1.000 9.058 32.754 43.925 500

-------------------------------------------------------------------------------------------

### \*1st question: what influences occupational status? AKA social mobility.

#### \*The dependent variable

tab occs4q, m

Occ status (with extra |

categories) | Freq. Percent Cum.

---------------------------------+-----------------------------------

0. Farm worker | 277 1.73 1.73

10. Farmer | 1,907 11.89 13.62

14. Unskilled worker | 571 3.56 17.18

18. Unskilled service | 851 5.31 22.48

24. Semi-skilled worker | 2,703 16.85 39.34

30. Large farmer (India version) | 19 0.12 39.45

32. Routine sales | 812 5.06 44.52

33. Higher service | 814 5.08 49.59

37. Skilled worker | 1,403 8.75 58.34

38. Routine clerical | 777 4.84 63.18

38.100 | 59 0.37 63.55

51. Higher sales | 886 5.52 69.08

60. Higher clerical | 1,220 7.61 76.68

70. Technical | 1,539 9.60 86.28

75. Administrative | 791 4.93 91.21

100. Professional | 1,410 8.79 100.00

---------------------------------+-----------------------------------

Total | 16,039 100.00

#### \*The independent variable we are particularly interested in is father's occupational status

##### \*Some description (always a good idea if the area is new to you!)

table faocc4t, c(m occs4q freq)

----------------------------------------------------

Fathers occ status, |

recoded to 14 main |

categories | mean(occs4q) Freq.

------------------------+---------------------------

0. Farm worker | 26.922 1,410

10. Farmer | 28.143 3,196

14. Unskilled worker | 37.578 846

18. Unskilled service | 43.412 228

24. Semi-skilled worker | 41.021 3,259

32. Routine sales | 41.713 376

33. Higher service | 45.442 437

37. Skilled worker | 44.564 2,100

38. Routine clerical | 48.733 658

51. Higher sales | 50.087 863

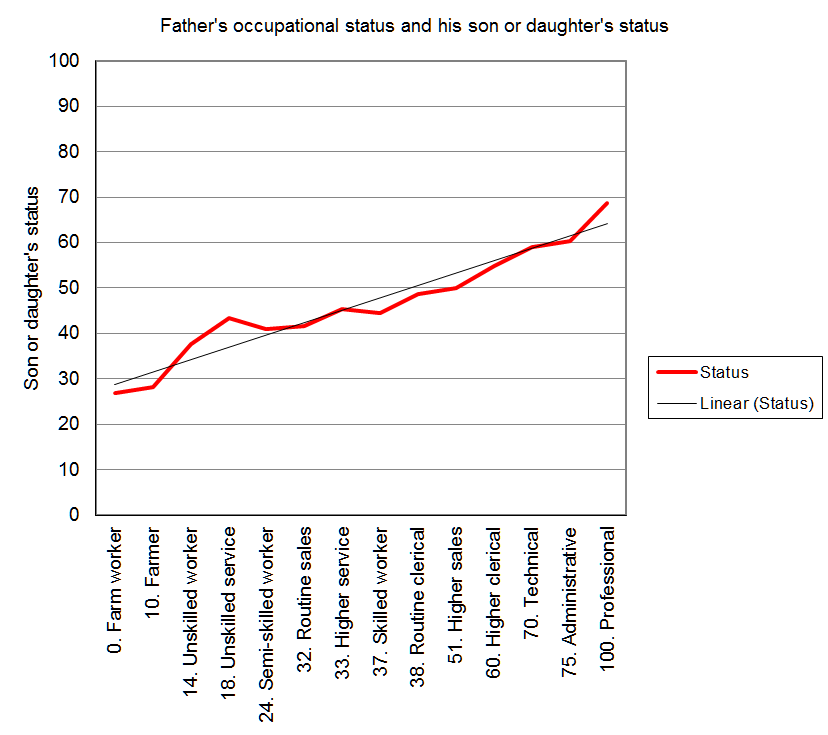
60. Higher clerical | 54.846 598

70. Technical | 58.950 636

75. Administrative | 60.418 676

100. Professional | 68.734 756

----------------------------------------------------



Lets assume the relationship is linear -- that's true, close enough. And it makes what is to come much simpler!

#### \*First estimate

regress occs4q faocc4t ,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 1, 16037) = 3267.14

Model | 1985427.22 1 1985427.22 Prob > F = 0.0000

Residual | 9745612.55 16037 607.695489 R-squared = 0.1692

-------------+------------------------------ Adj R-squared = 0.1692

Total | 11731039.8 16038 731.452785 Root MSE = 24.651

------------------------------------------------------------------------------

occs4q | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .4447849 .0077815 57.16 0.000 .4113947

\_cons | 27.79085 .3132955 88.70 0.000 .

------------------------------------------------------------------------------

QUESTIONS, BY WAY OF REVIEW

>> What status do we estimate for children from farm worker's families?

>> What status do we estimate for children from professional families?

>> How close are those to what we actually saw (in the means)?

>> What status do we estimate for children of skilled worker families?

#### \*A better estimate

QUESTION: Is this a good estimate of the effect of father's occupational status?

PART OF THE QUESTION: Is father's occupational status correlated with other things that are relevant to the issue, and might confound our very simple analysis?

-- how about age and gender (the "usual suspects")

##### \*control age and sex

Not much correlation, but a little. So we will control them

corr faocc4t agem maleq occs4q

(obs=16039)

| faocc4t agem maleq occs4q

-------------+------------------------------------

faocc4t | 1.0000

agem | -0.0805 1.0000

maleq | -0.0361 0.0640 1.0000

occs4q | 0.4114 -0.0227 -0.0954 1.0000

regress occs4q faocc4t agem maleq,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 3, 16035) = 1141.53

Model | 2064486.93 3 688162.31 Prob > F = 0.0000

Residual | 9666552.84 16035 602.840838 R-squared = 0.1760

-------------+------------------------------ Adj R-squared = 0.1758

Total | 11731039.8 16038 731.452785 Root MSE = 24.553

------------------------------------------------------------------------------

occs4q | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .4429471 .0077794 56.94 0.000 .4096949

agem | .0346712 .016152 2.15 0.032 .015467

maleq | -4.563328 .4017244 -11.36 0.000 -.081637

\_cons | 29.31967 .7662147 38.27 0.000 .

------------------------------------------------------------------------------

QUESTION: Do age and sex control make much difference to our estimate of the effect of father's occupational status?

##### \*control parents' education

corr faocc4t agem maleq pnteducq occs4q

Parent's education is more of a danger: big correlation with father's status AND with child's status.

| faocc4t agem maleq pnteducq occs4q

-------------+---------------------------------------------

faocc4t | 1.0000

agem | -0.0805 1.0000

maleq | -0.0361 0.0640 1.0000

pnteducq | 0.5492 -0.1660 -0.0546 1.0000

occs4q | 0.4114 -0.0227 -0.0954 0.4127 1.0000

regress occs4q faocc4t pnteducq agem maleq,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 4, 16034) = 1173.94

Model | 2657352.48 4 664338.12 Prob > F = 0.0000

Residual | 9073687.29 16034 565.902912 R-squared = 0.2265

-------------+------------------------------ Adj R-squared = 0.2263

Total | 11731039.8 16038 731.452785 Root MSE = 23.789

------------------------------------------------------------------------------

occs4q | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .2845262 .008987 31.66 0.000 .2631667

pnteducq | 1.854651 .0573 32.37 0.000 .2720888

agem | .1084001 .0158142 6.85 0.000 .0483578

maleq | -4.146164 .3894357 -10.65 0.000 -.0741741

\_cons | 17.71237 .8244481 21.48 0.000 .

------------------------------------------------------------------------------

QUESTION: Does the parents education control make much difference to our estimate of the effect of father's occupational status?

QUESTION: In this model what is our predicted status for a child from a farm laborer's family -- say a girl child, aged 30?

-- how about a girl age 30 from a professional home?

\*More generally

eval4 linear faocc4t "0 10 14 18 24 32 33 37 38 51 60 70 75 100" maleq "0 1"

\*---------- maleq = 0

faocc4t Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 35.44 .5033 34.46 36.43 0 0 1

10 38.29 .459 37.39 39.19 2.845 2.845 1.08

14 39.43 .4451 38.55 40.3 1.138 3.983 1.112

18 40.56 .4337 39.71 41.41 1.138 5.121 1.144

24 42.27 .4216 41.45 43.1 1.707 6.829 1.193

32 44.55 .4159 43.73 45.36 2.276 9.105 1.257

33 44.83 .416 44.02 45.65 .2845 9.389 1.265

37 45.97 .4184 45.15 46.79 1.138 10.53 1.297

38 46.25 .4195 45.43 47.08 .2845 10.81 1.305

51 49.95 .4497 49.07 50.84 3.699 14.51 1.409

60 52.51 .486 51.56 53.47 2.561 17.07 1.482

70 55.36 .5377 54.31 56.41 2.845 19.92 1.562

75 56.78 .5671 55.67 57.89 1.423 21.34 1.602

100 63.9 .7383 62.45 65.34 7.113 28.45 1.803

\*---------- maleq = 1

faocc4t Predicted s.e. ci Low ci High Now-prev Now-begin Now/begin

0 31.3 .4595 30.4 32.2 0 0 1

10 34.14 .4111 33.34 34.95 2.845 2.845 1.091

14 35.28 .3958 34.5 36.06 1.138 3.983 1.127

18 36.42 .3832 35.67 37.17 1.138 5.121 1.164

24 38.13 .3701 37.4 38.85 1.707 6.829 1.218

32 40.4 .3645 39.69 41.12 2.276 9.105 1.291

33 40.69 .3648 39.97 41.4 .2845 9.389 1.3

37 41.82 .3681 41.1 42.55 1.138 10.53 1.336

38 42.11 .3695 41.38 42.83 .2845 10.81 1.345

51 45.81 .4052 45.01 46.6 3.699 14.51 1.464

60 48.37 .4461 47.49 49.24 2.561 17.07 1.545

70 51.21 .5029 50.23 52.2 2.845 19.92 1.636

75 52.64 .5346 51.59 53.68 1.423 21.34 1.682

100 59.75 .7148 58.35 61.15 7.113 28.45 1.909

\*---------------------------------------------------------------------------------\*

By way of reminder re the meaning of the occupation codes:

0. Farm worker

10. Farmer

14. Unskilled worker

18. Unskilled service

24. Semi-skilled worker

32. Routine sales

33. Higher service

37. Skilled worker

38. Routine clerical

51. Higher sales

60. Higher clerical

70. Technical

75. Administrative

100. Professional

QUESTION: Is this new and smaller estimate of effect of father's status a more reasonable/logical/appropriate estimate?

#### \*Yet a further control: education

corr ed faocc4t agem maleq pnteducq occs4q

Education looks like something that ought to be controlled: highly correlated with father's status and with child's status:

| ed faocc4t agem maleq pnteducq occs4q

-------------+------------------------------------------------------

ed | 1.0000

faocc4t | 0.4278 1.0000

agem | -0.1340 -0.0805 1.0000

maleq | -0.0576 -0.0361 0.0640 1.0000

pnteducq | 0.6368 0.5492 -0.1660 -0.0546 1.0000

occs4q | 0.5850 0.4114 -0.0227 -0.0954 0.4127 1.0000

\*Lets add education to the controls in our analysis of father's status effects:

regress occs4q faocc4t ed pnteducq agem maleq,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 5, 16033) = 1981.13

Model | 4479945.04 5 895989.008 Prob > F = 0.0000

Residual | 7251094.73 16033 452.260633 R-squared = 0.3819

-------------+------------------------------ Adj R-squared = 0.3817

Total | 11731039.8 16038 731.452785 Root MSE = 21.266

------------------------------------------------------------------------------

occs4q | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .2220736 .0080941 27.44 0.000 .2054024

ed | 3.204102 .0504726 63.48 0.000 .515582

pnteducq | -.1460084 .060143 -2.43 0.015 -.0214203

agem | .1421255 .0141474 10.05 0.000 .0634029

maleq | -3.551408 .3482703 -10.20 0.000 -.063534

\_cons | -.1265746 .7887856 -0.16 0.873 .

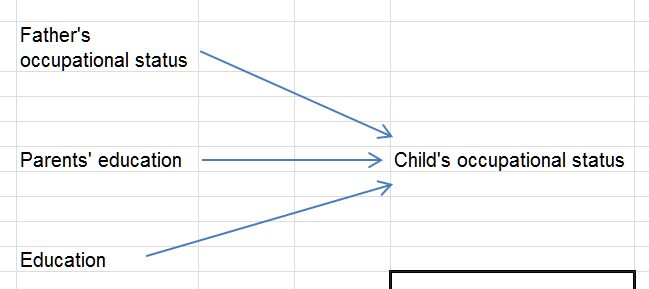
------------------------------------------------------------------------------

\*Question: Does education matter?

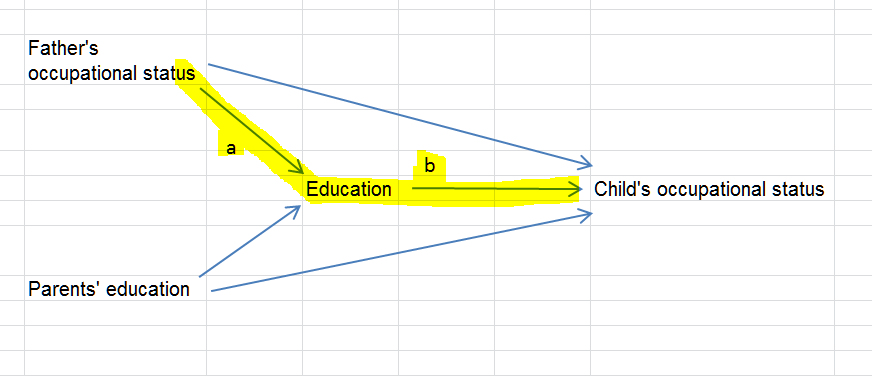
\*Question: does controlling education change the estimate of the effect of father's status?

\*Question (and this is the tricky one): is this control appropriate?

\*The answer mainly has to do with causal order. Is this the causal order?



\*Or is it instead this:



\*If it is the second diagram -- ie education is an INTERVENING variable, then part of the way that father's status influences son's status is INDIRECTLY via education (the path highlighted in yellow). So that should be counted as part of the effect of father's status.

#### \*Total effect of father's status (assuming that education is an INTERVENING variable, ie partly a consequence of father's status))

\*total effect of father's status

regress occs4q faocc4t pnteducq agem maleq,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 4, 16034) = 1173.94

Model | 2657352.48 4 664338.12 Prob > F = 0.0000

Residual | 9073687.29 16034 565.902912 R-squared = 0.2265

-------------+------------------------------ Adj R-squared = 0.2263

Total | 11731039.8 16038 731.452785 Root MSE = 23.789

------------------------------------------------------------------------------

occs4q | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .2845262 .008987 31.66 0.000 .26.31667

pnteducq | 1.854651 .0573 32.37 0.000 .2720888

agem | .1084001 .0158142 6.85 0.000 .0483578

maleq | -4.146164 .3894357 -10.65 0.000 -.0741741

\_cons | 17.71237 .8244481 21.48 0.000 .

------------------------------------------------------------------------------

\*direct effect of father's status

regress occs4q faocc4t ed pnteducq agem maleq,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 5, 16033) = 1981.13

Model | 4479945.04 5 895989.008 Prob > F = 0.0000

Residual | 7251094.73 16033 452.260633 R-squared = 0.3819

-------------+------------------------------ Adj R-squared = 0.3817

Total | 11731039.8 16038 731.452785 Root MSE = 21.266

------------------------------------------------------------------------------

occs4q | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .2220736 .0080941 27.44 0.000 .2054024

ed | 3.204102 .0504726 63.48 0.000 .515582

pnteducq | -.1460084 .060143 -2.43 0.015 -.0214203

agem | .1421255 .0141474 10.05 0.000 .0634029

maleq | -3.551408 .3482703 -10.20 0.000 -.063534

\_cons | -.1265746 .7887856 -0.16 0.873 .

------------------------------------------------------------------------------

##### \*Indirect effect of father's status via child's education

\* Get the indirect effect by subtraction: .284 - .222 = .062

\* or for standardized effects: .26 - .20 = .06

\*Compare: Total effect of father's status (standardized version): .26

\* Indirect effect of father's status (standardized version): .06

\* So percent indirect is ( .06 / .26 = 23% )

\* Thus we have found about a fifth of the father's status effect works through education (an important finding but that still leaves a lot more to be figured out)

\*The usual citation:   
Alwin, Duane F., and Robert M. Hauser. 1975. "The Decomposition of Effects in Path Analysis." American Sociological Review 40:37-47.

##### \*Indirect effect of parents' education via child's education

\*Now the same calculations for parents' education (lets again do the standardized version). Same regressions as above for father's status.

\* Get the indirect effect by subtraction: .27 - -.02 = .27 + .02 = .29

\*Compare: Total effect of father's status (standardized version): .27

\* Indirect effect of father's status (standardized version): .29

\* So everything is indirect (assuming away that annoying -.02 which is presumably nothing real). Thus we have explained ALL of the effect of growing up in a well educated home: its all a matter of schooling.

#### \*Path diagrams

\*Path diagrams often clarify things. By convention they go left (causally prior) to right (consequences); arrows are straight if possible (to reduce clutter); and standardized direct effects are shown.   
(We skip age and male just for simplicity)

regress ed faocc4t pnteducq ,beta

regress occs4q ed faocc4t pnteducq ,beta

regress ed faocc4t pnteducq ,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 2, 16036) = 5669.67

Model | 125819.405 2 62909.7027 Prob > F = 0.0000

Residual | 177932.651 16036 11.0958251 R-squared = 0.4142

-------------+------------------------------ Adj R-squared = 0.4141

Total | 303752.057 16038 18.9395222 Root MSE = 3.331

------------------------------------------------------------------------------

ed | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

faocc4t | .0194451 .0012583 15.45 0.000 .1117707

pnteducq | .631114 .0079329 79.56 0.000 .5753937

\_cons | 4.983379 .0552448 90.21 0.000 .

------------------------------------------------------------------------------

.

. regress occs4q ed faocc4t pnteducq ,beta

Source | SS df MS Number of obs = 16039

-------------+------------------------------ F( 3, 16035) = 3198.78

Model | 4392088.48 3 1464029.49 Prob > F = 0.0000

Residual | 7338951.29 16035 457.683274 R-squared = 0.3744

-------------+------------------------------ Adj R-squared = 0.3743

Total | 11731039.8 16038 731.452785 Root MSE = 21.394

------------------------------------------------------------------------------

occs4q | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

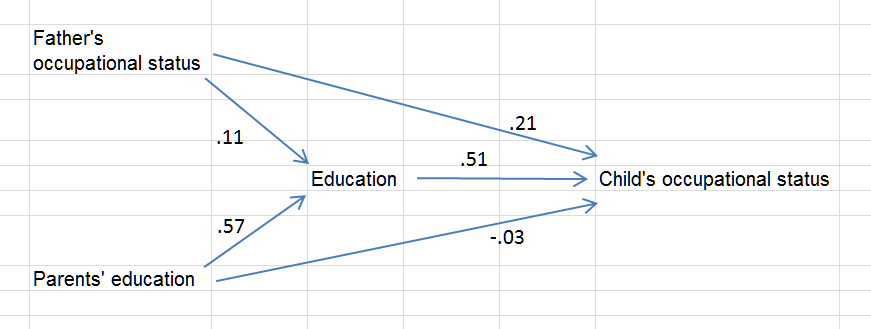
ed | 3.199204 .0507171 63.08 0.000 .5147939

faocc4t | .2238159 .0081412 27.49 0.000 .207014

pnteducq | -.196714 .0601693 -3.27 0.001 -.0288592

\_cons | 3.694396 .4356238 8.48 0.000 .

------------------------------------------------------------------------------



### \*Now do a path diagram and the appropriate regressions for a model including ln family income

\*First sort out the causal order, and draw the diagram

\*Then do the regressions

\*Then put the direct effects in the path diagram

\*Then look at how much of the effect of parent's education and father's occupation is indirect through child's education and occupation.

## \*Week 13: Missing data

\*Missing data are the bane of the analyzing classes. In a "normal" soiological analysis of attitude data perhaps a quarter of your time, and even more of your errors, come from missing data.

\*Stata 12 has some good new routines that you will eventually want to learn, but since we don't yet have Stata 12 at UNR, that will have to wait. Their underlying logic is regression imputation, which we deal with here.

\*For studies with little missing data simple regression imputation is probably all you need.

\*We will start with easy cases and work up to harder ones.

#### \*Data:

\*Use our usual data set: IsssUSA\_AnalysisVars\_1.dta

\*Run everything in Section #2 and Section #3 above, in the usual way. Then we will begin here

### \*Missing data in multiple-item scales

\*Lets start with missing data mainly in multiple-item scales:

\* -- for example, 5 questions measuring attitudes toward a single item;

\* -- or 6 alternative measures of vegitation density

\* -- or 4 measures of a sheep's weight taken at different times, with different scales

\*The key is that all these measure the same thing, so we want a measure that averages them for our analysis. And if one measurement is missing, the others provide a fine estimate of what the answer should have been.

\*Example: Questions on who do you trust for moral advice (we have looked at them before)

sum mrlPopeQ mrlABCQ mrlFundQ mrlMDQ mrlNIHQ mrlEthCQ mrlYou1stQ mrlYouQ

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

mrlPopeQ | 2196 31.73953 30.76663 0 100

mrlABCQ | 2194 33.80811 29.68837 0 100

mrlFundQ | 2180 35.12615 30.87444 0 100

mrlMDQ | 2201 63.15311 27.19222 0 100

mrlNIHQ | 2193 62.1979 28.24478 0 100

-------------+--------------------------------------------------------

mrlEthCQ | 2201 49.25034 30.52579 0 100

mrlYou1stQ | 2195 67.67654 26.77517 0 100

mrlYouQ | 2194 72.76664 25.77178 0 100

corr mrlPopeQ mrlABCQ mrlFundQ mrlMDQ mrlNIHQ mrlEthCQ mrlYou1stQ mrlYouQ

There are 3 separate concepts here. We will look at moral authority related to religion -- the Pope, the Archbishop of Canterbury, and fundamentalist leaders:

| mrlPopeQ mrlABCQ mrlFundQ mrlMDQ mrlNIHQ mrlEthCQ mrlYo~tQ mrlYouQ

-------------+------------------------------------------------------------------------

mrlPopeQ | 1.0000

mrlABCQ | 0.8192 1.0000

mrlFundQ | 0.7679 0.8836 1.0000

mrlMDQ | 0.1384 0.1884 0.1225 1.0000

mrlNIHQ | 0.1220 0.1851 0.1466 0.6620 1.0000

mrlEthCQ | 0.2822 0.3203 0.2706 0.5476 0.5239 1.0000

mrlYou1stQ | 0.1491 0.1735 0.1773 0.3940 0.3361 0.2405 1.0000

mrlYouQ | 0.0844 0.1040 0.1278 0.3507 0.3517 0.1941 0.7221 1.0000

\*These three questions do measure one concept:

factor mrlPopeQ mrlABCQ mrlFundQ mrlMDQ mrlNIHQ mrlEthCQ mrlYou1stQ mrlYouQ

Factor analysis/correlation Number of obs = 2135

Method: principal factors Retained factors = 3

Rotation: (unrotated) Number of params = 21

--------------------------------------------------------------------------

Factor | Eigenvalue Difference Proportion Cumulative

-------------+------------------------------------------------------------

Factor1 | 3.01094 1.31448 0.6255 0.6255

Factor2 | 1.69647 1.03499 0.3524 0.9779

Factor3 | 0.66148 0.68745 0.1374 1.1153

Factor4 | -0.02597 0.05307 -0.0054 1.1099

Factor5 | -0.07904 0.02888 -0.0164 1.0935

Factor6 | -0.10792 0.03539 -0.0224 1.0711

Factor7 | -0.14331 0.05541 -0.0298 1.0413

Factor8 | -0.19872 . -0.0413 1.0000

--------------------------------------------------------------------------

LR test: independent vs. saturated: chi2(28) = 1.0e+04 Prob>chi2 = 0.0000

rotate, varimax

Factor analysis/correlation Number of obs = 2135

Method: principal factors Retained factors = 3

Rotation: orthogonal varimax (Kaiser off) Number of params = 21

--------------------------------------------------------------------------

Factor | Variance Difference Proportion Cumulative

-------------+------------------------------------------------------------

Factor1 | 2.48672 0.92418 0.5166 0.5166

Factor2 | 1.56253 0.24289 0.3246 0.8412

Factor3 | 1.31964 . 0.2741 1.1153

--------------------------------------------------------------------------

LR test: independent vs. saturated: chi2(28) = 1.0e+04 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

-----------------------------------------------------------

Variable | Factor1 Factor2 Factor3 | Uniqueness

-------------+------------------------------+--------------

mrlPopeQ | 0.8408 0.0769 0.0387 | 0.2856

mrlABCQ | 0.9323 0.1255 0.0484 | 0.1128

mrlFundQ | 0.9029 0.0508 0.0857 | 0.1749

mrlMDQ | 0.0844 0.7366 0.2696 | 0.3776

mrlNIHQ | 0.0880 0.7169 0.2431 | 0.4192

mrlEthCQ | 0.2545 0.6279 0.0944 | 0.5321

mrlYou1stQ | 0.1145 0.2198 0.7613 | 0.3590

mrlYouQ | 0.0510 0.1981 0.7669 | 0.3701

-----------------------------------------------------------

#### \*look at some cases that have missing data

sum Religion\_\_\_\_\_ mrlPopeQ mrlABCQ mrlFundQ

There are 2295 cases in all. Only about 100 cases with missing data (unusually low for US surveys)

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

Religion\_\_~\_ | 2295 20 0 20 20

mrlPopeQ | 2196 31.73953 30.76663 0 100

mrlABCQ | 2194 33.80811 29.68837 0 100

mrlFundQ | 2180 35.12615 30.87444 0 100

\*Here we list only cases with missing data.

\*There are two distinct kinds: cases that are missing on all three items (and, in this example, on sex and education too). These are people who skipped a whole section of the questionnaire. These are hard to fix!

\*The other kind are folk who answered some of the "moral authority" questions but not all three. These are easy to fix

list maleQ edNowQ mrlPopeQ mrlABCQ mrlFundQ if mrlPopeQ== . | mrlABCQ== . | mrlFundQ== . , sep(0)

+------------------------------------------------+

| maleQ edNowQ mrlPopeQ mrlABCQ mrlFundQ |

|------------------------------------------------|

5. | . . . . . |

6. | . . . . . |

12. | . . . . . |

13. | . . . . . |

14. | . . . . . |

16. | . . . . . |

20. | . . . . . |

24. | . . . . . |

26. | . . . . . |

28. | . . . . . |

...etc...

167. | . . . . . |

178. | 1 16.5 . . . |

284. | 1 16.5 50 75 . |

345. | 0 16.5 0 0 . |

480. | 0 16.5 0 0 . |

487. | 1 16.5 . . . |

494. | 1 13 . 0 0 |

589. | 0 12 . . . |

597. | 0 18 . 0 0 |

604. | 1 13 0 0 . |

637. | 1 13 . 100 100 |

656. | 0 16.5 . . . |

759. | 0 18 . . . |

865. | 0 16.5 25 25 . |

898. | 1 13 . . . |

918. | 0 12 . 100 100 |

994. | 1 16.5 50 . 50 |

1033. | 0 14 . . . |

1120. | 0 14 50 50 . |

1144. | 0 13 25 25 . |

1150. | 0 13 25 75 . |

1217. | 0 16.5 0 0 . |

1242. | 1 12 . . . |

1269. | 0 13 0 0 . |

1274. | 1 13 25 25 . |

1281. | 0 16.5 0 . 75 |

1296. | 0 13 50 50 . |

1317. | 1 16.5 0 0 . |

1321. | 1 18 0 0 . |

1368. | 1 13 . . . |

1377. | 1 12 . 50 50 |

1385. | 1 13 . . . |

1390. | 0 18 0 0 . |

1435. | 0 13 50 . 75 |

1449. | 0 13 . 50 25 |

1506. | 0 14 50 . 50 |

1528. | 0 18 0 . 0 |

1565. | 0 13 . . . |

1645. | 0 13 0 0 . |

1826. | 1 12 50 50 . |

1862. | 0 14 0 0 . |

1891. | 0 18 0 0 . |

1899. | 0 12 . . . |

1904. | 0 18 0 50 . |

1971. | . 12 0 0 . |

2006. | 0 13 . . . |

2033. | 0 13 0 . 75 |

2157. | 0 10 . . . |

2226. | 1 11 25 . 25 |

2284. | . 12 . . . |

+------------------------------------------------+

\*correlations & regressions are normally only for cases that answered all questions. So here we would have for a typical analysis (with only 2091 cases):

. corr cHumn9 maleQ edNowQ mrlPopeQ mrlABCQ mrlFundQ

(obs=2091)

| cHumn9 maleQ edNowQ mrlPopeQ mrlABCQ mrlFundQ

-------------+------------------------------------------------------

cHumn9 | 1.0000

maleQ | 0.1843 1.0000

edNowQ | -0.0458 0.0300 1.0000

mrlPopeQ | 0.2229 0.0663 -0.0364 1.0000

mrlABCQ | 0.2231 0.0512 -0.0293 0.8222 1.0000

mrlFundQ | 0.1915 0.0547 -0.0608 0.7757 0.8842 1.0000

regress cHumn9 maleQ edNowQ mrlPopeQ mrlABCQ mrlFundQ, beta

Source | SS df MS Number of obs = 2091

-------------+------------------------------ F( 5, 2085) = 39.44

Model | 146661.116 5 29332.2232 Prob > F = 0.0000

Residual | 1550730.98 2085 743.755866 R-squared = 0.0864

-------------+------------------------------ Adj R-squared = 0.0842

Total | 1697392.1 2090 812.149329 Root MSE = 27.272

------------------------------------------------------------------------------

cHumn9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

maleQ | 9.834163 1.196426 8.22 0.000 .1725511

edNowQ | -.571297 .262947 -2.17 0.030 -.045654

mrlPopeQ | .1082996 .0345968 3.13 0.002 .1172061

mrlABCQ | .1690391 .0485752 3.48 0.001 .1761712

mrlFundQ | -.0623897 .0423037 -1.47 0.140 -.067417

\_cons | 19.95576 3.870153 5.16 0.000 .

------------------------------------------------------------------------------

#### \*How we normally make a scale

egen author9 = rowmean(mrlPopeQ mrlABCQ mrlFundQ )

\*Note the scale has less missing that any of its component items -- because it takes the average of whatever answers each respondent made: all three, or two out of three, or even one out of the three:

sum author9 mrlPopeQ mrlABCQ mrlFundQ

sum author9 mrlPopeQ mrlABCQ mrlFundQ

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

author9 | 2202 33.50931 28.60084 0 100

mrlPopeQ | 2196 31.73953 30.76663 0 100

mrlABCQ | 2194 33.80811 29.68837 0 100

mrlFundQ | 2180 35.12615 30.87444 0 100

list maleQ edNowQ author9 mrlPopeQ mrlABCQ mrlFundQ if mrlPopeQ== . | mrlABCQ== . | mrlFundQ== . , sep(0)

+----------------------------------------------------------+

| maleQ edNowQ author9 mrlPopeQ mrlABCQ mrlFundQ |

|----------------------------------------------------------|

5. | . . . . . . |

6. | . . . . . . |

12. | . . . . . . |

13. | . . . . . . |

14. | . . . . . . |

16. | . . . . . . |

20. | . . . . . . |

24. | . . . . . . |

26. | . . . . . . |

28. | . . . . . . |

...etc...

167. | . . . . . . |

178. | 1 16.5 . . . . |

284. | 1 16.5 62.5 50 75 . |

345. | 0 16.5 0 0 0 . |

480. | 0 16.5 0 0 0 . |

487. | 1 16.5 . . . . |

494. | 1 13 0 . 0 0 |

589. | 0 12 . . . . |

597. | 0 18 0 . 0 0 |

604. | 1 13 0 0 0 . |

637. | 1 13 100 . 100 100 |

656. | 0 16.5 . . . . |

759. | 0 18 . . . . |

865. | 0 16.5 25 25 25 . |

898. | 1 13 . . . . |

918. | 0 12 100 . 100 100 |

994. | 1 16.5 50 50 . 50 |

1033. | 0 14 . . . . |

1120. | 0 14 50 50 50 . |

1144. | 0 13 25 25 25 . |

1150. | 0 13 50 25 75 . |

1217. | 0 16.5 0 0 0 . |

1242. | 1 12 . . . . |

1269. | 0 13 0 0 0 . |

1274. | 1 13 25 25 25 . |

1281. | 0 16.5 37.5 0 . 75 |

1296. | 0 13 50 50 50 . |

1317. | 1 16.5 0 0 0 . |

1321. | 1 18 0 0 0 . |

1368. | 1 13 . . . . |

1377. | 1 12 50 . 50 50 |

1385. | 1 13 . . . . |

1390. | 0 18 0 0 0 . |

1435. | 0 13 62.5 50 . 75 |

1449. | 0 13 37.5 . 50 25 |

1506. | 0 14 50 50 . 50 |

1528. | 0 18 0 0 . 0 |

1565. | 0 13 . . . . |

1645. | 0 13 0 0 0 . |

1826. | 1 12 50 50 50 . |

1862. | 0 14 0 0 0 . |

1891. | 0 18 0 0 0 . |

1899. | 0 12 . . . . |

1904. | 0 18 25 0 50 . |

1971. | . 12 0 0 0 . |

2006. | 0 13 . . . . |

2033. | 0 13 37.5 0 . 75 |

2157. | 0 10 . . . . |

2226. | 1 11 25 25 . 25 |

2284. | . 12 . . . . |

+----------------------------------------------------------+

\*And now the regression is a lot better: better measurement (with the scale rather than separate items) and a few more cases (because the scale in effect imputes the missing data):

regress cHumn9 maleQ edNowQ author9, beta

Source | SS df MS Number of obs = 2124

-------------+------------------------------ F( 3, 2120) = 63.34

Model | 141327.88 3 47109.2932 Prob > F = 0.0000

Residual | 1576655.58 2120 743.705464 R-squared = 0.0823

-------------+------------------------------ Adj R-squared = 0.0810

Total | 1717983.46 2123 809.224429 Root MSE = 27.271

------------------------------------------------------------------------------

cHumn9 | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

maleQ | 10.0039 1.186634 8.43 0.000 .1758274

edNowQ | -.517668 .260304 -1.99 0.047 -.0414504

author9 | .2092862 .0207324 10.09 0.000 .2107539

\_cons | 19.02836 3.838599 4.96 0.000 .

------------------------------------------------------------------------------

### \*Regression imputation of missing data

\*The multiple item scales are an easy case, because how to imput the missing data is clear: use answers to the other items in the scale.

\*But in other cases we have a more difficult hunt for other information to use in the prediction, and our normal hunting procedure is OLS regression (or logistic or probit regression sometimes).   
  
--we could do this for the data that is still missing in our author9 scale (those would be the folk who did not answer any of the 3 questions)  
  
--or we could do it for a measure that was not part of a scale, like age, education, occupational status, and such.

#### \*Example #1: Education

tab edNowQ, m

RECODE of |

EdNow | Freq. Percent Cum.

------------+-----------------------------------

10 | 42 1.83 1.83

11 | 31 1.35 3.18

12 | 444 19.35 22.53

13 | 729 31.76 54.29

14 | 255 11.11 65.40

16.5 | 453 19.74 85.14

18 | 166 7.23 92.37

22 | 35 1.53 93.90

. | 140 6.10 100.00

------------+-----------------------------------

Total | 2,295 100.00

\*We need to find something that predicts education, to use in imputing. But the thing itself should not have much missing data. Lets look around:

sum edNowQ cThrp9 female mrlRel9 ageM

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

edNowQ | 2155 14.09211 2.2767 10 22

cThrp9 | 2294 64.95205 29.77205 0 100

female | 2148 .5121043 .4999699 0 1

mrlRel9 | 2202 33.50931 28.60084 0 100

ageM | 2295 46.46536 15.44872 17 87

corr edNowQ cThrp9 female mrlRel9 ageM

(obs=2126)

| edNowQ cThrp9 female mrlRel9 ageM

-------------+---------------------------------------------

edNowQ | 1.0000

cThrp9 | -0.0027 1.0000

female | -0.0264 -0.0820 1.0000

mrlRel9 | -0.0532 -0.0599 -0.0626 1.0000

ageM | 0.1343 -0.0497 -0.0132 -0.1100 1.0000

regress edNowQ cThrp9 female mrlRel9 ageM, beta

Only age is going to help, and even it not very much! ( variable ageM has a little missing data imputed itself, which is OK if we are happy with how we imputed it. In fact is was a crude imputation, so its not ideal but better than nothing).

Source | SS df MS Number of obs = 2126

-------------+------------------------------ F( 4, 2121) = 10.97

Model | 223.77446 4 55.943615 Prob > F = 0.0000

Residual | 10813.1794 2121 5.09815155 R-squared = 0.0203

-------------+------------------------------ Adj R-squared = 0.0184

Total | 11036.9539 2125 5.19386066 Root MSE = 2.2579

------------------------------------------------------------------------------

edNowQ | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

cThrp9 | -.0000723 .0016842 -0.04 0.966 -.0009297

female | -.1243827 .0985535 -1.26 0.207 -.0272883

mrlRel9 | -.0032447 .0017287 -1.88 0.061 -.0407751

ageM | .0185928 .0031126 5.97 0.000 .1294158

\_cons | 13.4134 .2189084 61.27 0.000 .

------------------------------------------------------------------------------

\*throw everything but age out, and try again:

regress edNowQ ageM, beta

Source | SS df MS Number of obs = 2155

-------------+------------------------------ F( 1, 2153) = 39.71

Model | 202.211674 1 202.211674 Prob > F = 0.0000

Residual | 10962.7542 2153 5.09185054 R-squared = 0.0181

-------------+------------------------------ Adj R-squared = 0.0177

Total | 11164.9659 2154 5.18336392 Root MSE = 2.2565

------------------------------------------------------------------------------

edNowQ | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

ageM | .0192729 .0030583 6.30 0.000 .1345781

\_cons | 13.19677 .1501621 87.88 0.000 .

------------------------------------------------------------------------------

\*This is what the regression implies about ed (ignoring the random element in green)

gen edEst = .01927 \* ageM + 13.196

\*round, so its easier for us to see

replace edEst = round( edEst, 0.5)

(2295 real changes made)

tab edEst

edEst | Freq. Percent Cum.

------------+-----------------------------------

13.5 | 355 15.47 15.47

14 | 1,247 54.34 69.80

14.5 | 676 29.46 99.26

15 | 17 0.74 100.00

------------+-----------------------------------

Total | 2,295 100.00

\*Now make a new version of education, using these imputations (for now, ignore the random element)

gen edNowM = edNowQ

label variable edNowM "Education now, with missing imputed from age"

tab edNowM, m

Education |

now, with |

missing |

imputed |

from age | Freq. Percent Cum.

------------+-----------------------------------

10 | 42 1.83 1.83

11 | 31 1.35 3.18

12 | 444 19.35 22.53

13 | 729 31.76 54.29

14 | 255 11.11 65.40

16.5 | 453 19.74 85.14

18 | 166 7.23 92.37

22 | 35 1.53 93.90

. | 140 6.10 100.00 -- these are the ones we want to fix

------------+-----------------------------------

Total | 2,295 100.00

replace edNowM = edEst if edNowM == .

(140 real changes made)

tab edNowM, m

Education |

now, with |

missing |

imputed |

from age | Freq. Percent Cum.

------------+-----------------------------------

10 | 42 1.83 1.83

11 | 31 1.35 3.18

12 | 444 19.35 22.53

13 | 729 31.76 54.29

13.5 | 2 0.09 54.38

14 | 391 17.04 71.42

14.5 | 1 0.04 71.46

15 | 1 0.04 71.50

16.5 | 453 19.74 91.24

18 | 166 7.23 98.47

22 | 35 1.53 100.00

------------+-----------------------------------

Total | 2,295 100.00

\*Since there is not much missing data on education, in practice we might leave things at this stage and use edNowM.

\*In the real world, we did not do that but used another question to impute -- it had to do with educational plans, and is a moste unusual question, so our way of using that is of no general interest. Besides its messy!

\*In general you should IN THEORY always add in a random element, reflecting uncertainty in your estimate. That matters if there is lots of missing, and is good practice anyway (although sometimes hard to explain to many a journal reviewer, or even a PhD supervisor).

\*So make a random variable, mean zero, std dev=1 in the usual way:

gen errEd = invnorm(uniform())

sum errEd

Variable | Obs Mean Std. Dev. Min Max

-------------+--------------------------------------------------------

errEd | 2295 .0364627 .9980252 -3.383373 3.194541

\*here is our regression again. We are after the "Root MSE" which is the proper amount of random error

Source | SS df MS Number of obs = 2155

-------------+------------------------------ F( 1, 2153) = 39.71

Model | 202.211674 1 202.211674 Prob > F = 0.0000

Residual | 10962.7542 2153 5.09185054 R-squared = 0.0181

-------------+------------------------------ Adj R-squared = 0.0177

Total | 11164.9659 2154 5.18336392 Root MSE = 2.2565

------------------------------------------------------------------------------

edNowQ | Coef. Std. Err. t P>|t| Beta

-------------+----------------------------------------------------------------

ageM | .0192729 .0030583 6.30 0.000 .1345781

\_cons | 13.19677 .1501621 87.88 0.000 .

------------------------------------------------------------------------------

\*Now add the proper amount of random error. We make another version of edNowM so everything is clear, and all in one place, just like before except for adding in the random bit in green:

gen edNowM2 = edNowQ

label variable edNowM2 "Education now, with missing imputed from age + random error"

replace edNowM2 = edEst + 2.2565 \* errEd if edNowM2 == .

(140 real changes made)

\*round, so its easier for us to see

replace edNowM2 = round(edNowM2, 0.5)

tab1 edNowQ edNowM2, m

RECODE of |

EdNow | Freq. Percent Cum.

------------+-----------------------------------

10 | 42 1.83 1.83

11 | 31 1.35 3.18

12 | 444 19.35 22.53

13 | 729 31.76 54.29

14 | 255 11.11 65.40

16.5 | 453 19.74 85.14

18 | 166 7.23 92.37

22 | 35 1.53 93.90

. | 140 6.10 100.00

------------+-----------------------------------

Total | 2,295 100.00

-> tabulation of edNowM2

error | Freq. Percent Cum.

------------+-----------------------------------

7.5 | 1 0.04 0.04

8 | 1 0.04 0.09

9 | 1 0.04 0.13

9.5 | 3 0.13 0.26

10 | 44 1.92 2.18

10.5 | 4 0.17 2.35

11 | 33 1.44 3.79

11.5 | 9 0.39 4.18

12 | 450 19.61 23.79

12.5 | 7 0.31 24.10

13 | 743 32.37 56.47

13.5 | 13 0.57 57.04

14 | 268 11.68 68.71

14.5 | 15 0.65 69.37

15 | 8 0.35 69.72

15.5 | 12 0.52 70.24

16 | 10 0.44 70.68

16.5 | 458 19.96 90.63

17 | 4 0.17 90.81

17.5 | 4 0.17 90.98

18 | 167 7.28 98.26

18.5 | 1 0.04 98.30

19 | 1 0.04 98.34

19.5 | 1 0.04 98.39

20.5 | 1 0.04 98.43

21 | 1 0.04 98.47

22 | 35 1.53 100.00

------------+-----------------------------------

Total | 2,295 100.00

##### \*Out of range values; fractional values:

\*Some of the predicted values in edNowM2 are "out of range" in the sense that they are below the minimum value for EdNow (which is 10 in these data). In other cases they might also be ones above the normal range (we have none in this example).

\*You might want to do something about that since they are confusing to many readers. For example if you imputed an education of -2.5 years, that would be strange (although statistically OK). But in the present case, the out of range codes are perfectly legitimate numbers for education, so you don't really need to do anything.

\*Some people might also worry about "fractional" years of education. Especially here where the fractional values are not from mainstream answers (except for our 16.5 for college graduation) but only from missing data imputation.

\*A tidy version withbout "out of range" and "fractional" values is statistically very very slightly worse than the version we have in edNowM2 but easier to think about. Especially if education is not a main interest in your analysis, it may be best to tidy all up:

gen edNowM3 = edNowM2

label variable edNowM3 "Education now, missing imputed from age + random error, tidy version"

\*we don't want to round away the "real" 16.5 for college grads

replace edNowM3 = round( edNowM3, 1) if edNowM3 != 16.5

(71 real changes made)

recode edNowM3 ( 0 / 10 = 10)

(edNowM3: 3 changes made)

tab edNowM3

edNowM3 | Freq. Percent Cum.

------------+-----------------------------------

10 | 50 2.18 2.18

11 | 37 1.61 3.79

12 | 459 20.00 23.79

13 | 750 32.68 56.47

14 | 281 12.24 68.71

15 | 23 1.00 69.72

16 | 22 0.96 70.68

16.5 | 458 19.96 90.63

17 | 4 0.17 90.81

18 | 171 7.45 98.26

19 | 2 0.09 98.34

20 | 1 0.04 98.39

21 | 2 0.09 98.47

22 | 35 1.53 100.00

------------+-----------------------------------

Total | 2,295 100.00