TEACHING STATA

Nyi Nyi Naing
Universiti Sultan Zainal Abidin (UniSZA)
Terengganu
Malaysia

01/02/2024

Top 10 Statistical Tools Used in Medical Research

BY RAMYA SRIRAM ON JULY 15, 2020

Table of Contents

Who are our trainees?

- Universities - Postgraduate candidates and lecturers especially for those trained for Biostatistics, Epidemiology, public health, clinical medicine and laboratory based medicine disciplines
- Health personnel under disciplines of Medicine, Dentistry, Pharmacy and Allied health in universities and Ministry of Health
- Research institutes under National Institute of Health, Ministry of Health
- Clinical Research Centres (CRCs) under Ministry of Health, Malaysia
- Pharmaceutical companies
- Non-governmental Organisations
- Biostatisticians, Epidemiologists, public health personnel in Southeast Asia

SOME HIGHLIGHTS OF TEACHING STATA IN MEDICAL AND HEALTH SCIENCES RESEARCH

HIGHLIGHTS OF TEACHING STATA

Data management
Types of study designs
Sample size determination
Examples of analysis

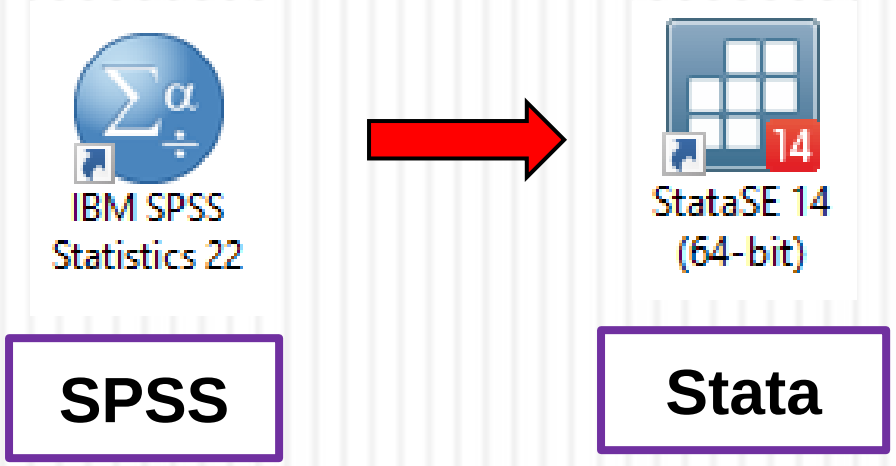
HIGHLIGHTS OF TEACHING STATA

Data management

Types of files

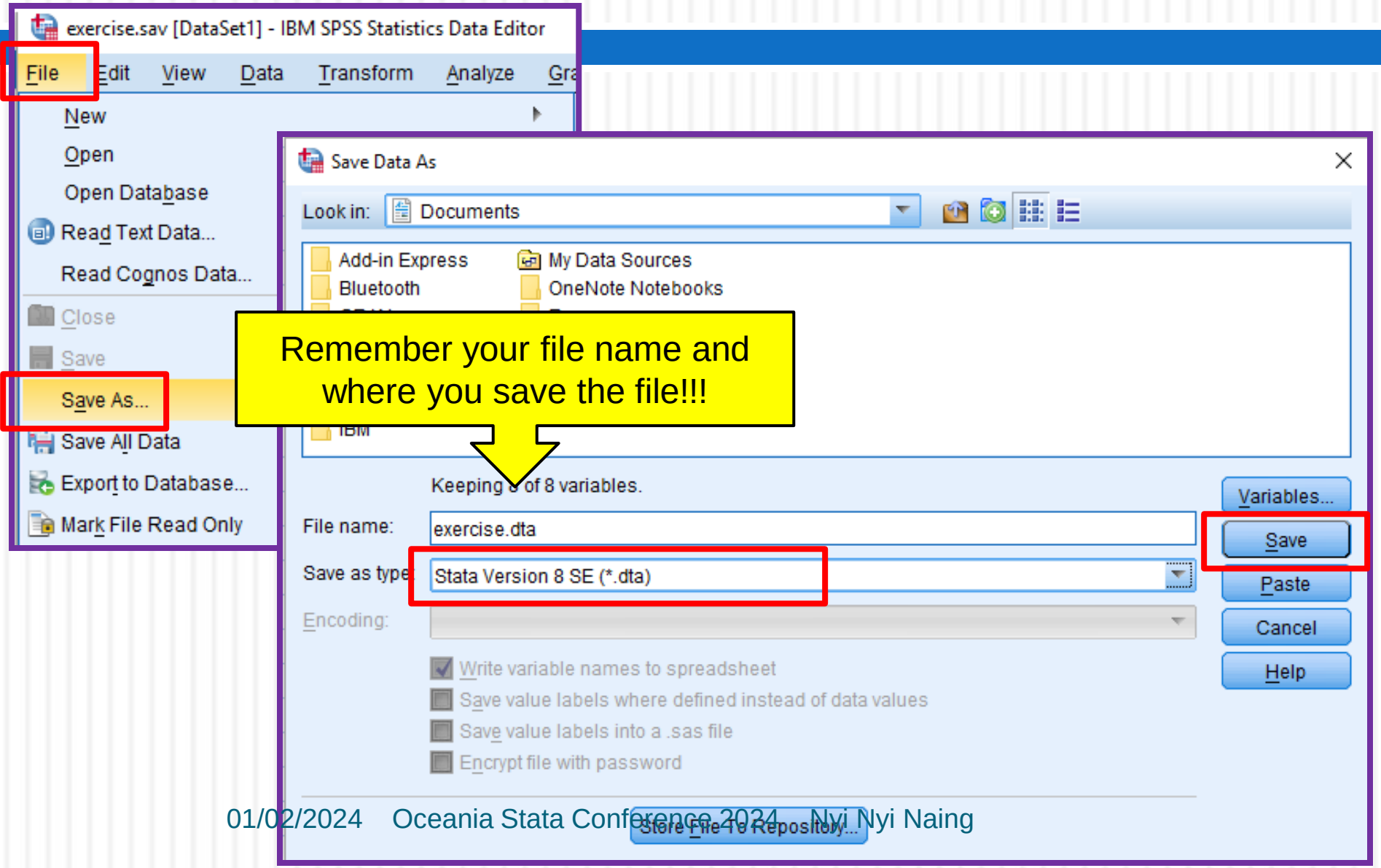
.dat	Tab-delimited files
.dta	Data files
.do	Command files
.ado	Programs files
.hlp	Help files
.log	Displayed in notepad (log file)
.smcl	Displayed in viewer window (formatted log file)
.gph	Graphic files

Import SPSS to Stata

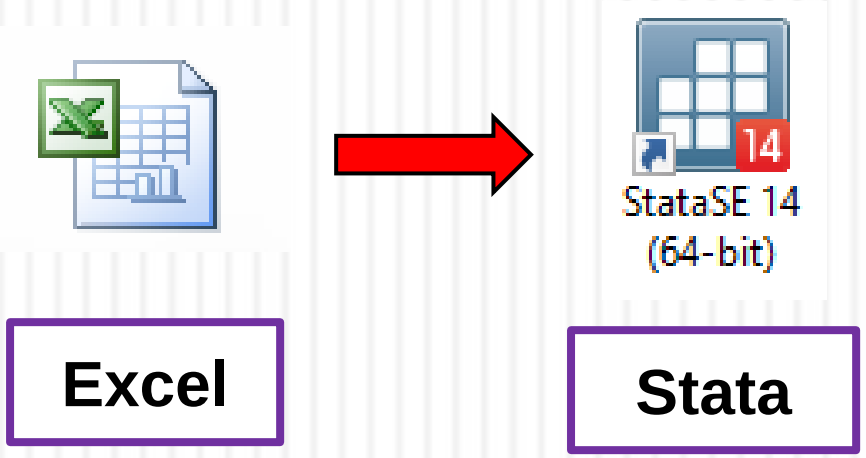


Open SPSS dataset (exercise.sav)

Import SPSS to Stata

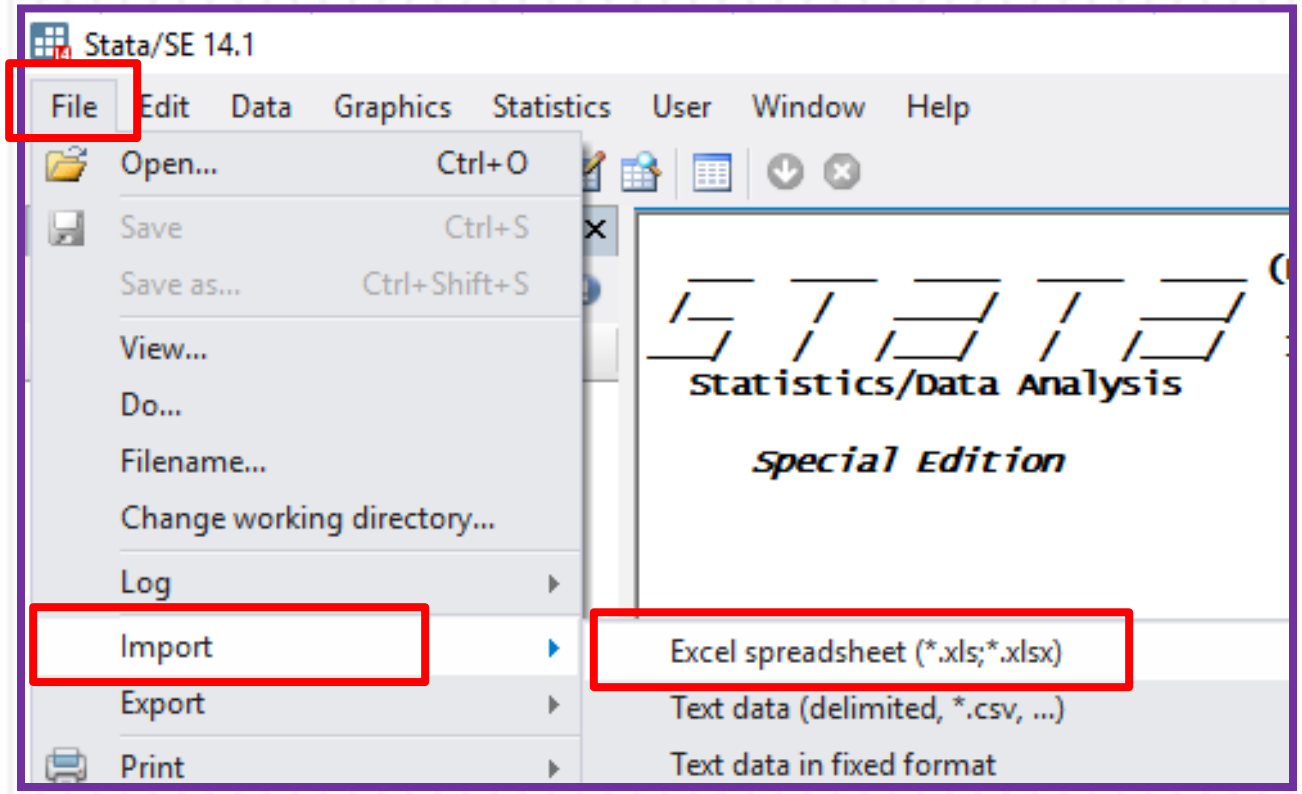


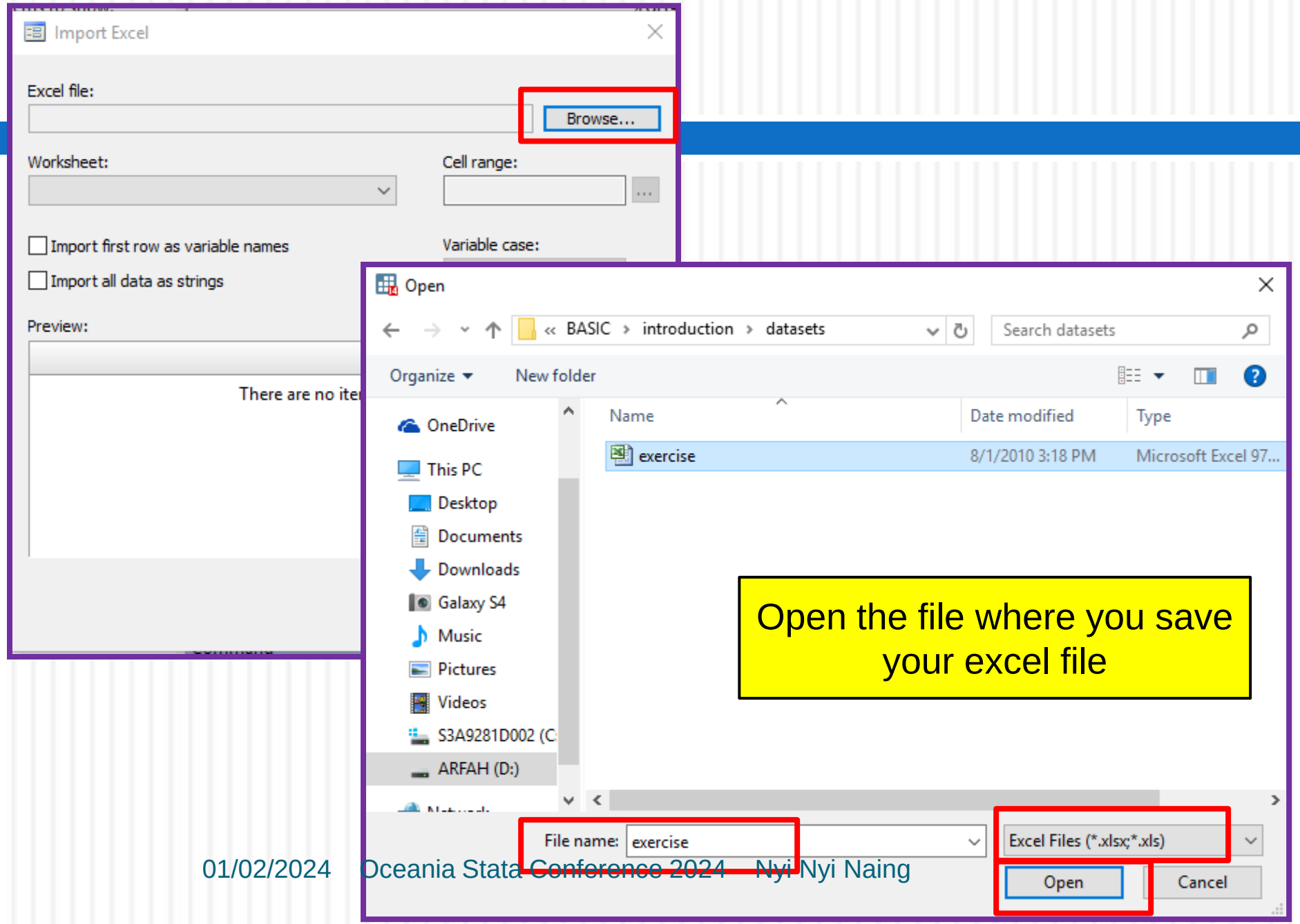
Import Excel to Stata

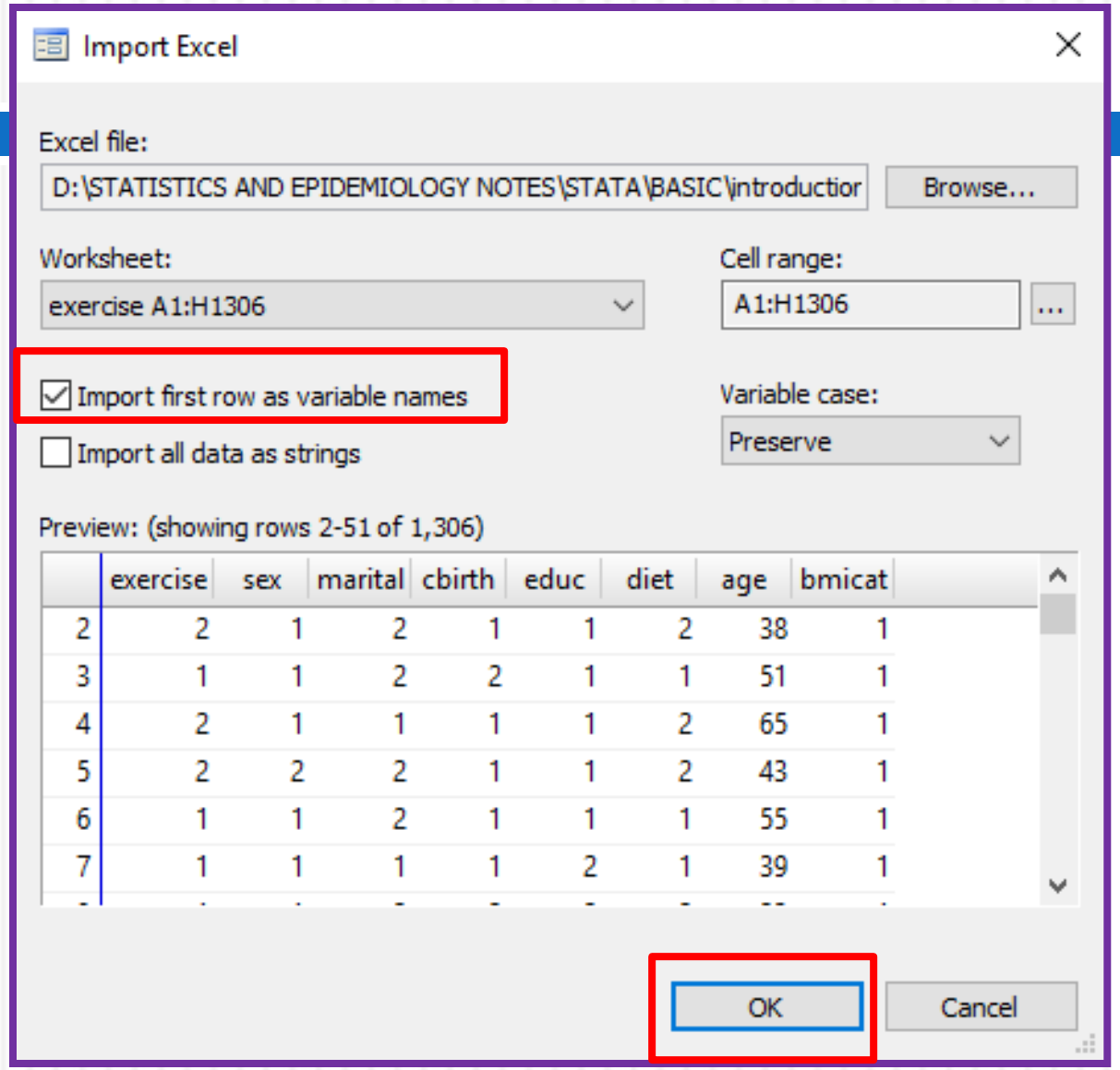


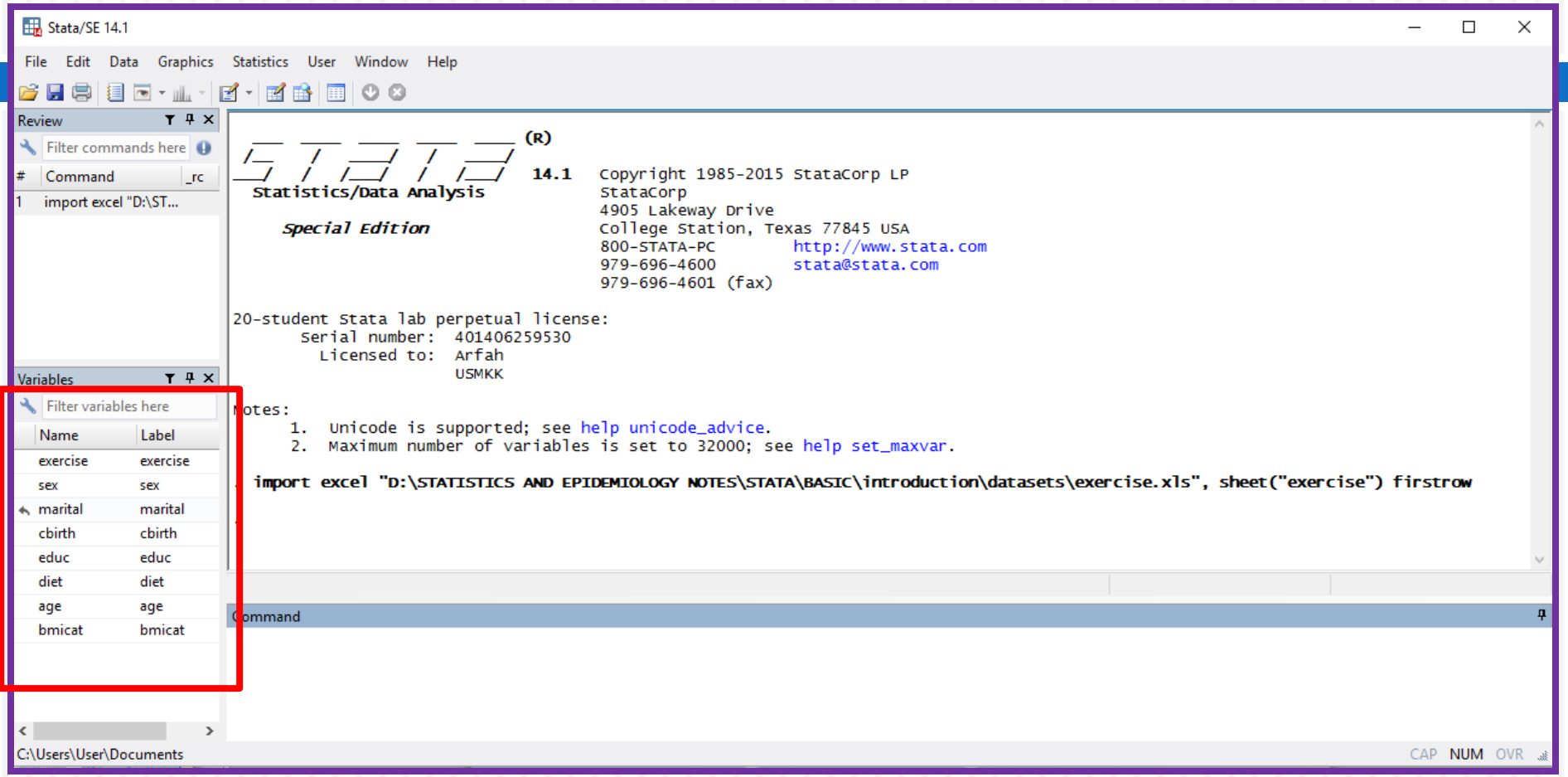
Import Excel to Stata

- Stata:
- File > Import > Excel spreadsheet (*.xls, *.xlsx)

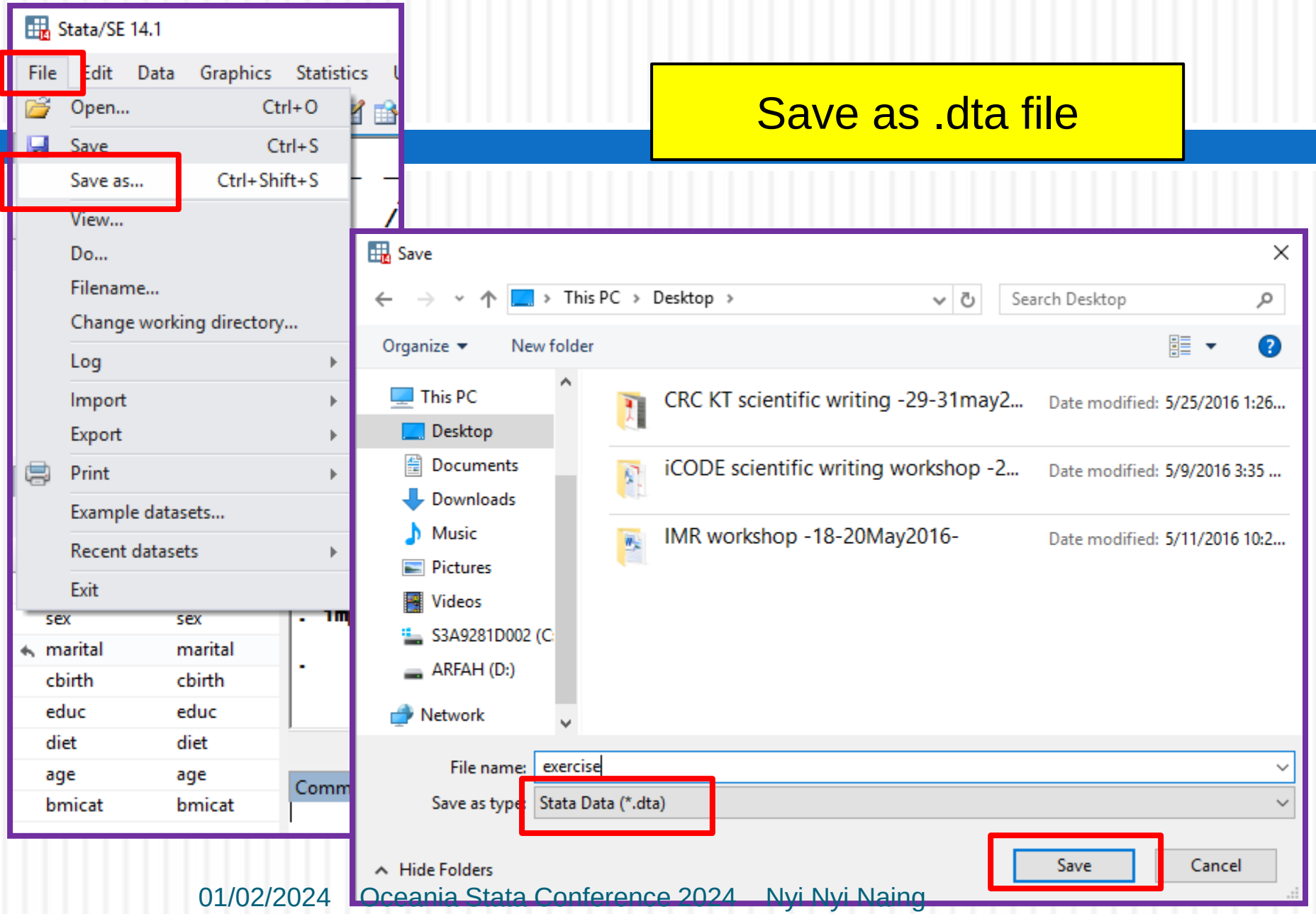



Import Excel
Excel file:
Browse...
Worksheet:
Cell range:
Import first row as variable names
Import all data as strings
Variable case:
Preview:
There are no ite
Open
« BASIC › introduction › datasets
Search datasets
Organize
New folder
OneDrive
This PC
Desktop
Documents
Downloads
Galaxy S4
Music
Pictures
Videos
S3A9281D002 (C
ARFAH (D:)
Name
Date modified
Type
exercise
8/1/2010 3:18 PM
Microsoft Excel 97...
Open the file where you save your excel file
File name: exercise
Excel Files (*.xlsx;*.xls)
Open
Cancel


Import Excel
Excel file:
D:\STATISTICS AND EPIDEMIOLOGY NOTES\STATA\BASIC\introductior
Browse...
Worksheet:
exercise A1:H1306
Cell range:
A1:H1306
...
Import first row as variable names
Import all data as strings
Variable case:
Preserve
Preview: (showing rows 2-51 of 1,306)
exercise sex marital cbirth educ diet age bmicat
2 2 1 2 1 1 2 38 1
3 1 1 2 2 1 1 51 1
4 2 1 1 1 1 2 65 1
5 2 2 2 1 1 2 43 1
6 1 1 2 1 1 1 55 1
7 1 1 1 1 2 1 39 1
OK
Cancel


Stata/SE 14.1
File Edit Data Graphics Statistics User Window Help
Review
Filter commands here
Command _rc
1 import excel "D:\ST...
Variables
Filter variables here
Name Label
exercise exercise
sex sex
marital marital
cbirth cbirth
educ educ
diet diet
age age
bmicat bmicat
(R)
14.1
Statistics/Data Analysis
Special Edition
Copyright 1985-2015 StataCorp LP
StataCorp
4905 Lakeway Drive
College Station, Texas 77845 USA
800-STATA-PC http://www.stata.com
979-696-4600 stata@stata.com
979-696-4601 (fax)
20-student Stata lab perpetual license:
Serial number: 401406259530
Licensed to: Arfah
USMKK
Notes:
1. Unicode is supported; see help unicode_advice.
2. Maximum number of variables is set to 32000; see help set_maxvar.
. import excel "D:\STATISTICS AND EPIDEMIOLOGY NOTES\STATA\BASIC\introduction\datasets\exercise.xls", sheet("exercise") firstrow
Command
C:\Users\User\Documents
CAP NUM OVR

Save as .dta file



input

- input exposure disease number
  (code: 0/1 unexposed/exposed,
  0/1 no disease/disease)
- 1 1 100
- 1 0 30
- 0 1 25
- 0 0 70
- end
- tab exposure disease [fweight=number],chi2
- cc exposure disease [fweight=number]

Immediate command-
tabi 100 30 \ 25 70, chi2
cci 100 30 25 70

```
. cc exposure disease [fweight=number]

                                                      Proportion
                |    Exposed   Unexposed |      Total       exposed
----------------+------------------------+------------------------
          Cases |        100          30 |        130        0.7692
       Controls |         25          70 |         95        0.2632
----------------+------------------------+------------------------
          Total |        125         100 |        225        0.5556
                |                        |
                |      Point estimate    |    [95% conf. interval]
----------------+------------------------+------------------------
     Odds ratio |           9.333333     |    4.851641    18.07077 (exact)
Attr. frac. ex. |           .8928571     |    .7938842     .944662 (exact)
Attr. frac. pop |           .6868132     |
----------------+------------------------+------------------------
                         chi2(1) =    56.93  Pr>chi2 = 0.0000
```

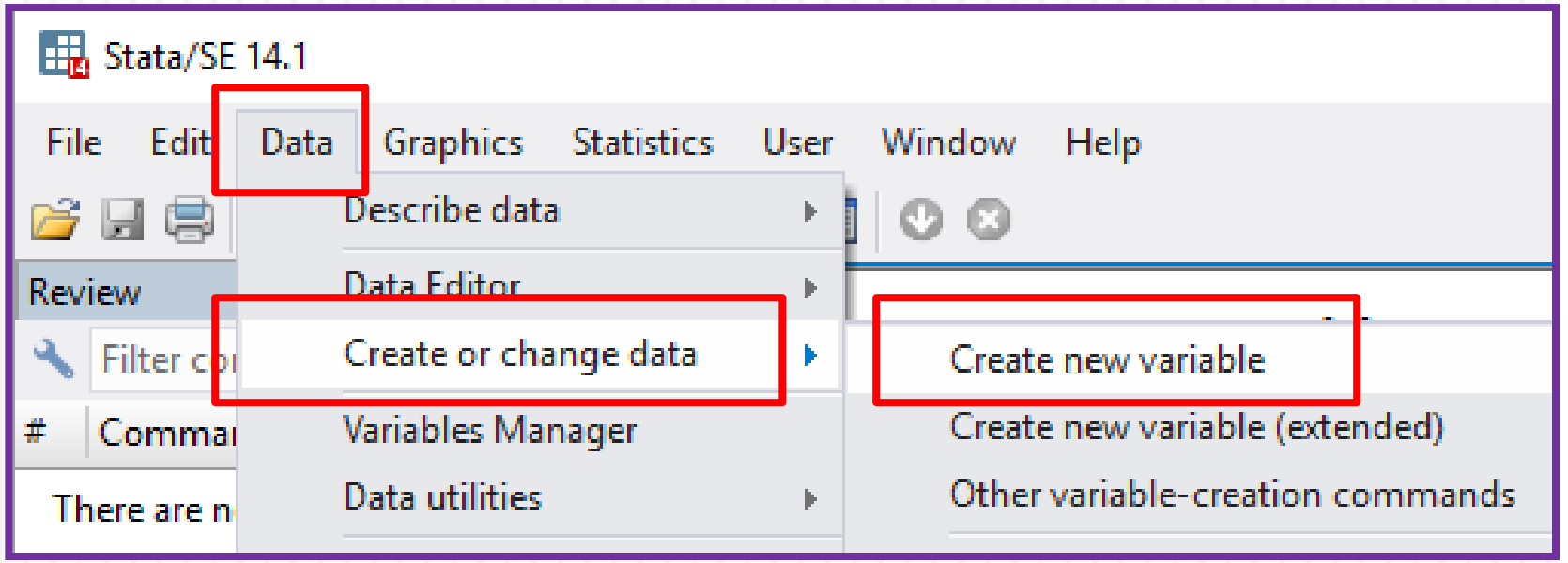
MANUAL DATA ENTRY IN STATA

Steps in Manual Data Entry

1. Create a new variable
2. Label a new variable’s name
3. Create a value label for a categorical variable (pseudo-numeric)
4. Data entry

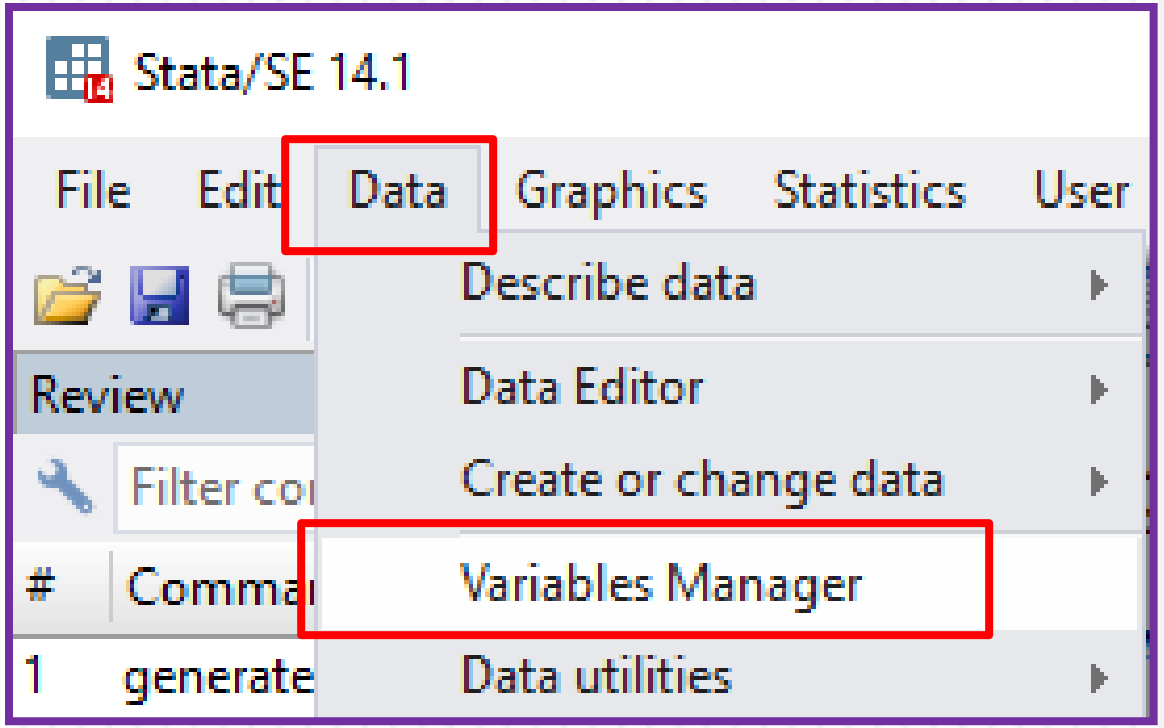
Step 1: Create a new variable

Data > Create or change data > Create new variable

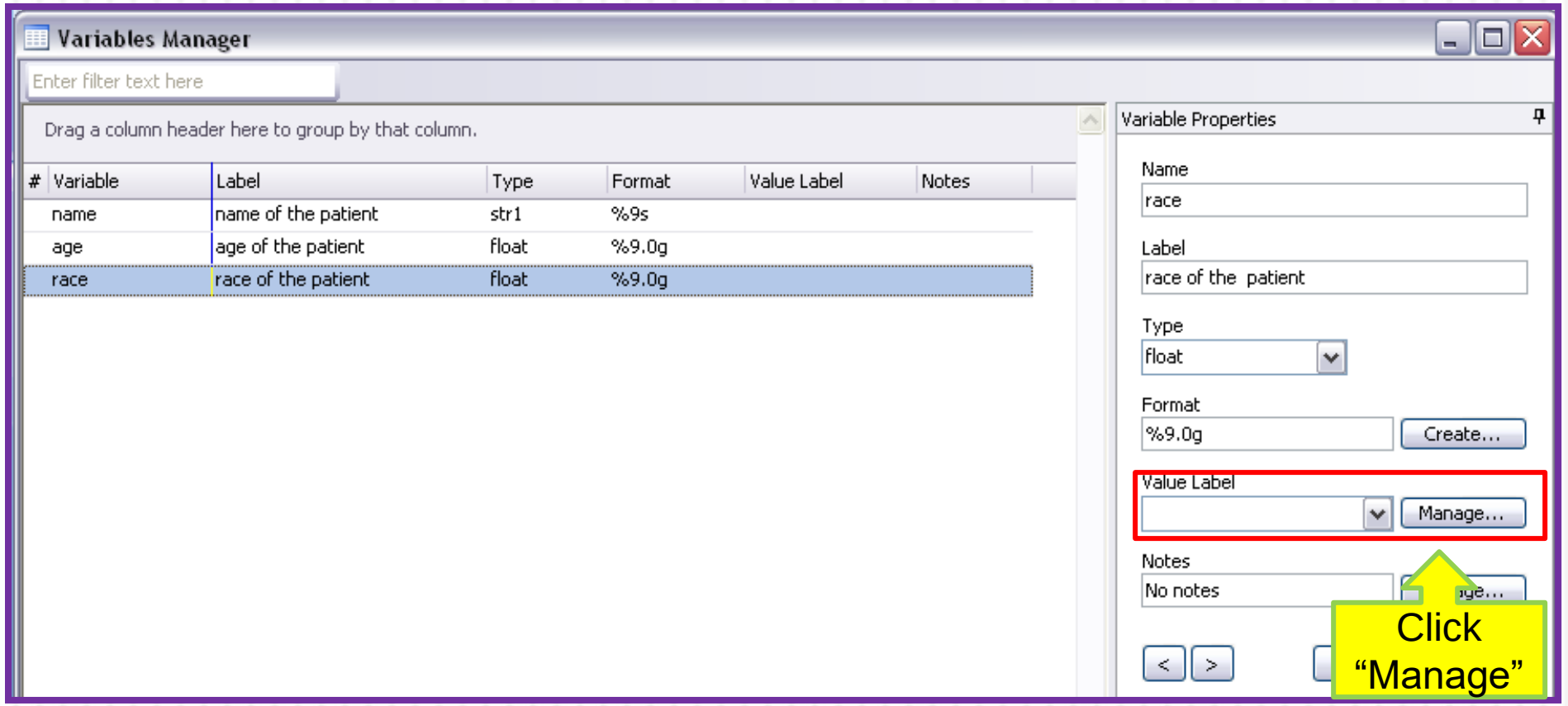


Step 2: Label a new variable's name

Data > Variables Manager **OR** Click the icon 

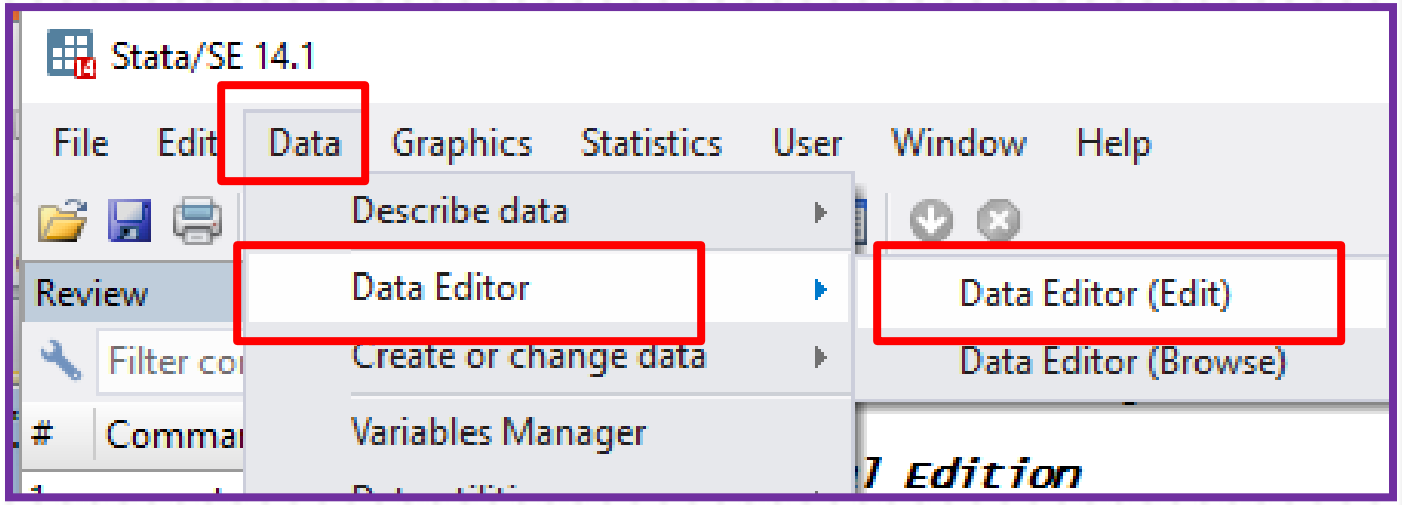


Step 3: Create values label

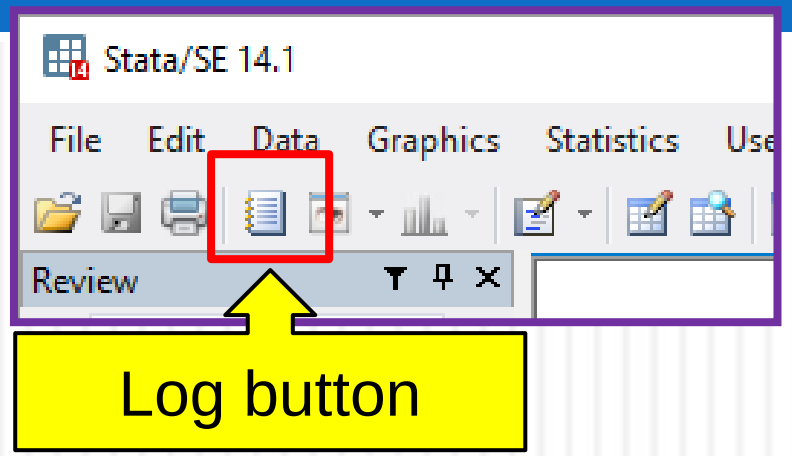


Step 4: Data entry

Data > Data Editor > Data Editor (Edit) **OR** Click the icon 



Saving results by using log



- A log file is a record of your Result window
- It records all commands and all textual output
- Stata can save the file in one of two different formats:
  - Stata Mark up and Control Language (SMCL) format
  - Plain log file

Saving results by using logs – two different formats

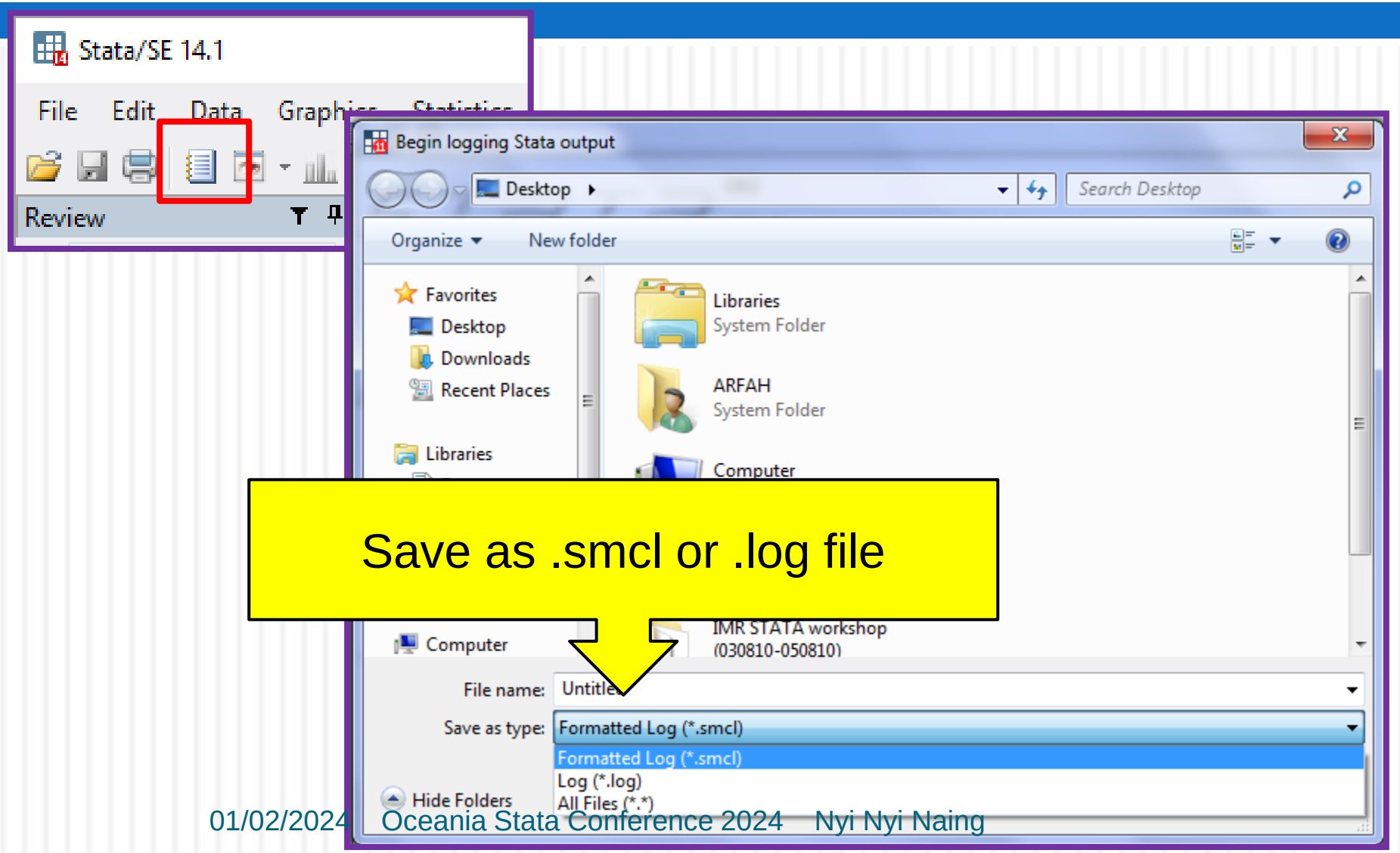
SMCL format

- Preserves all the formatting and links from Results window
- Can open these results in the Viewer
- Will behave as though they were in the Results window

Plain log file

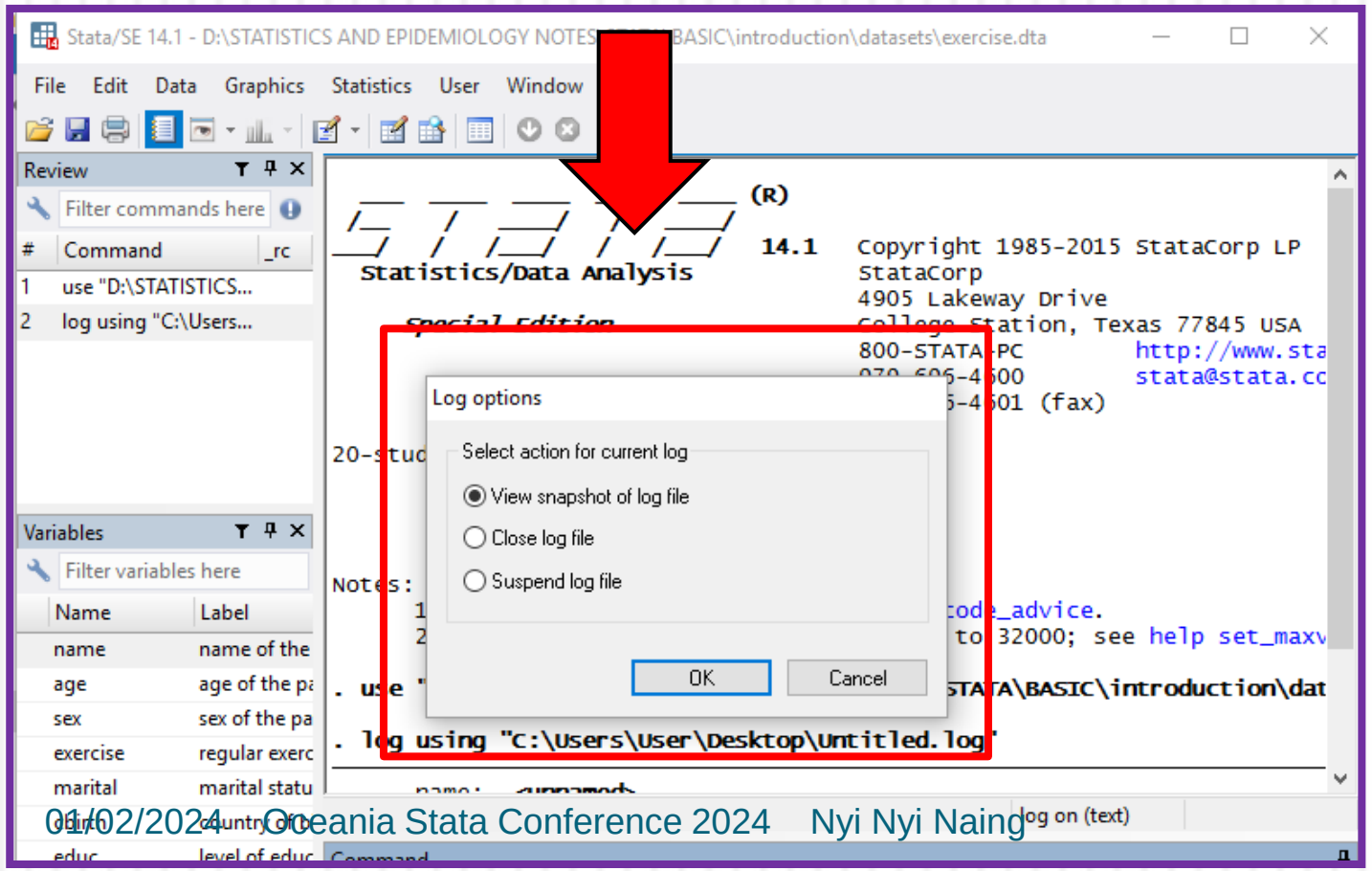
- Plain-text files without any formatting
- Can open by using notepad

Saving results by using logs



Saving results by using logs

- If you specify a file that already exists:



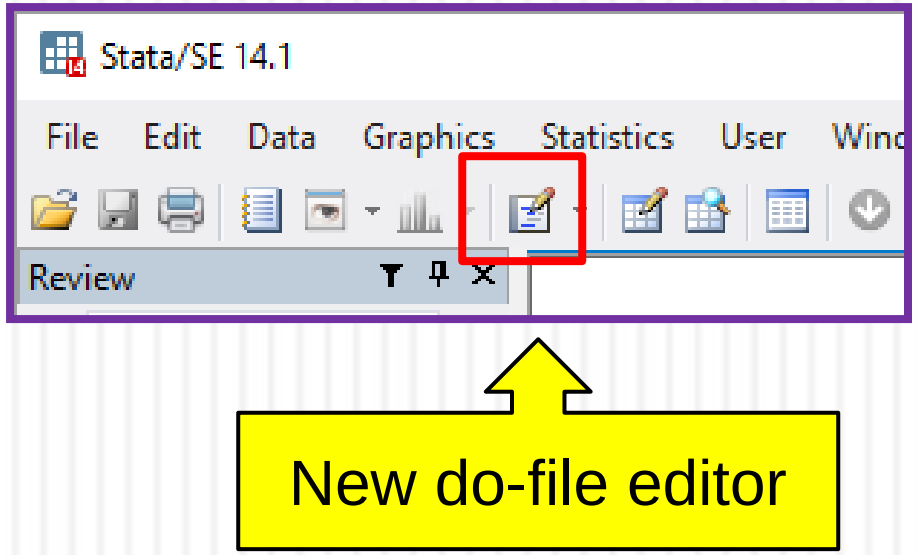
View logs (.log file)

smoking - Notepad

File Edit Format View Help

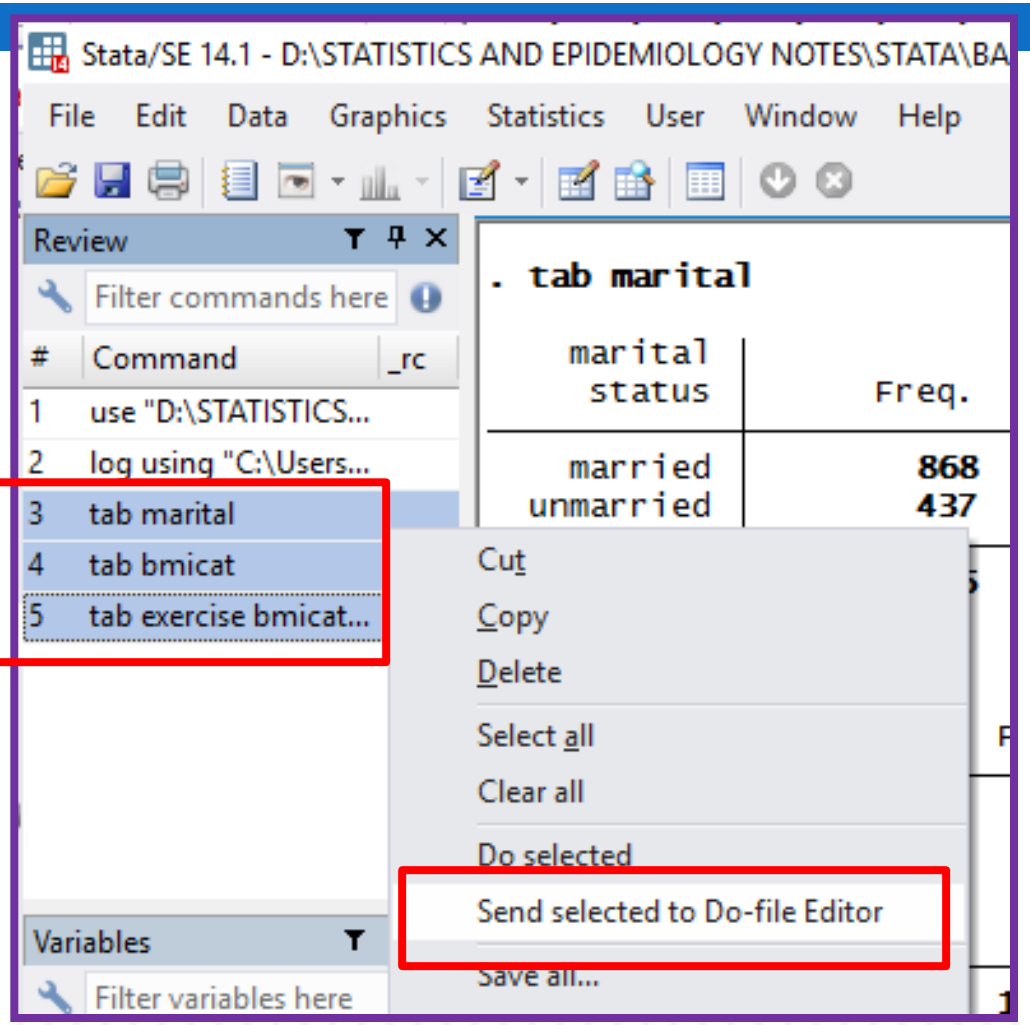
```
------------------------------------------------------------------------------
      name:  <unnamed>
       log:  C:\Users\user\Desktop\smoking.log
  log type:  text
 opened on:   1 Aug 2010, 23:59:27

. sum

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    exercise |      1305    1.270498    .4443874          1          2
         sex |      1305    1.504981    .5001669          1          2
     marital |      1305    1.334866    .4721245          1          2
     cbirth |      1305    1.303448    .4599234          1          2
       educ |      1305    1.594636    .4911506          1          2
-------------+--------------------------------------------------------
        diet |      1305    1.739464    .4390956          1          2
         age |      1305    45.14023    13.84321         20         70
      bmicat |      1305    2.422989    .8057074          1          4

. sum age, de

                            age
-------------------------------------------------------------
      Percentiles      Smallest
 1%           21             20
 5%           23             20
10%           26             20       Obs                1305
25%           34             20       Sum of Wgt.        1305

50%           44                      Mean           45.14023
                      Largest         Std. Dev.      13.84321
75%           57             70
90%           64             70       Variance       191.6345
95%           67             70       Skewness       .0142363
99%           69             70       Kurtosis       1.917969

. tab bmicat
```

Using the do-file editor

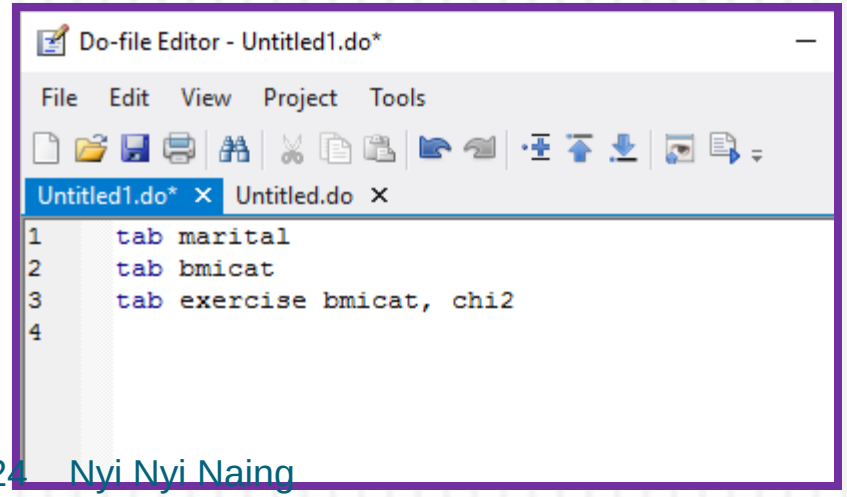


- Do-file → is a file containing a list of commands for Stata to run

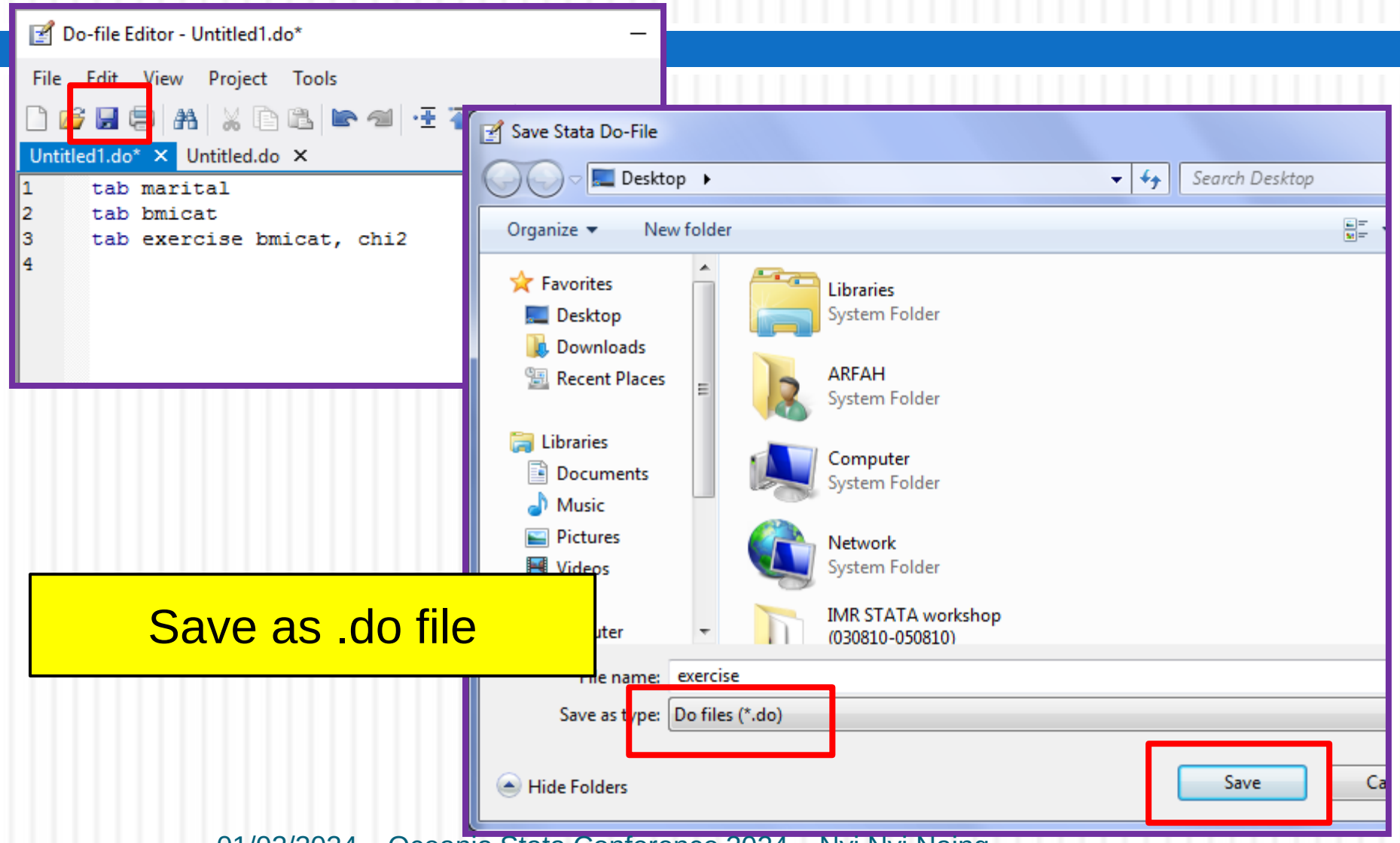
Using the do-file editor



Highlighted the selected command > Send to Do-file editor



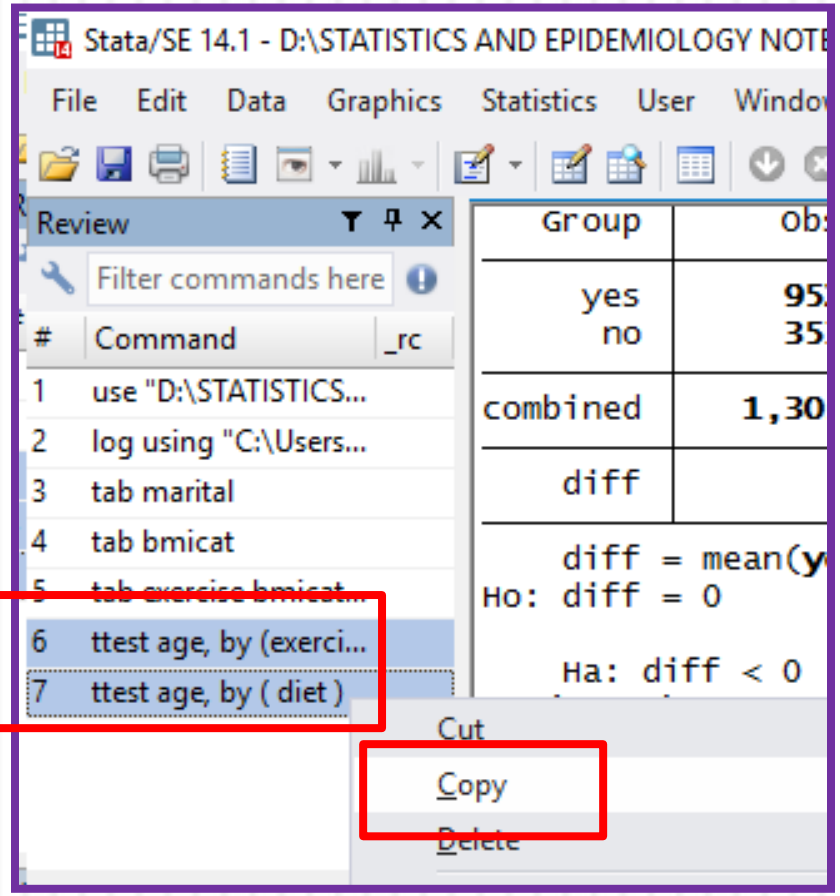
Save the do-file



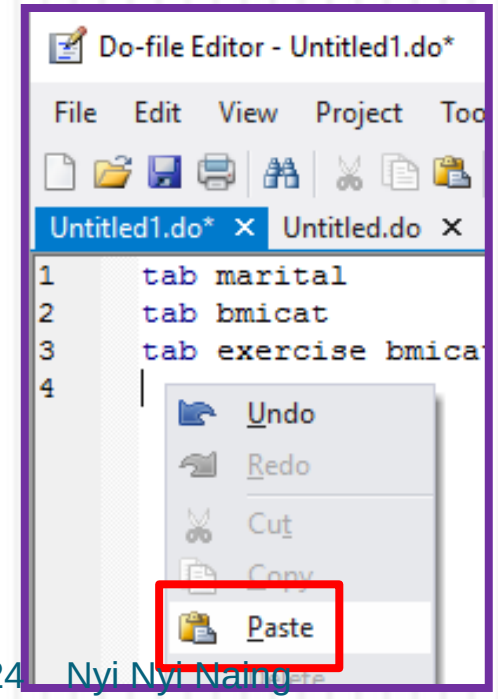
Save as .do file

Using the do-file editor

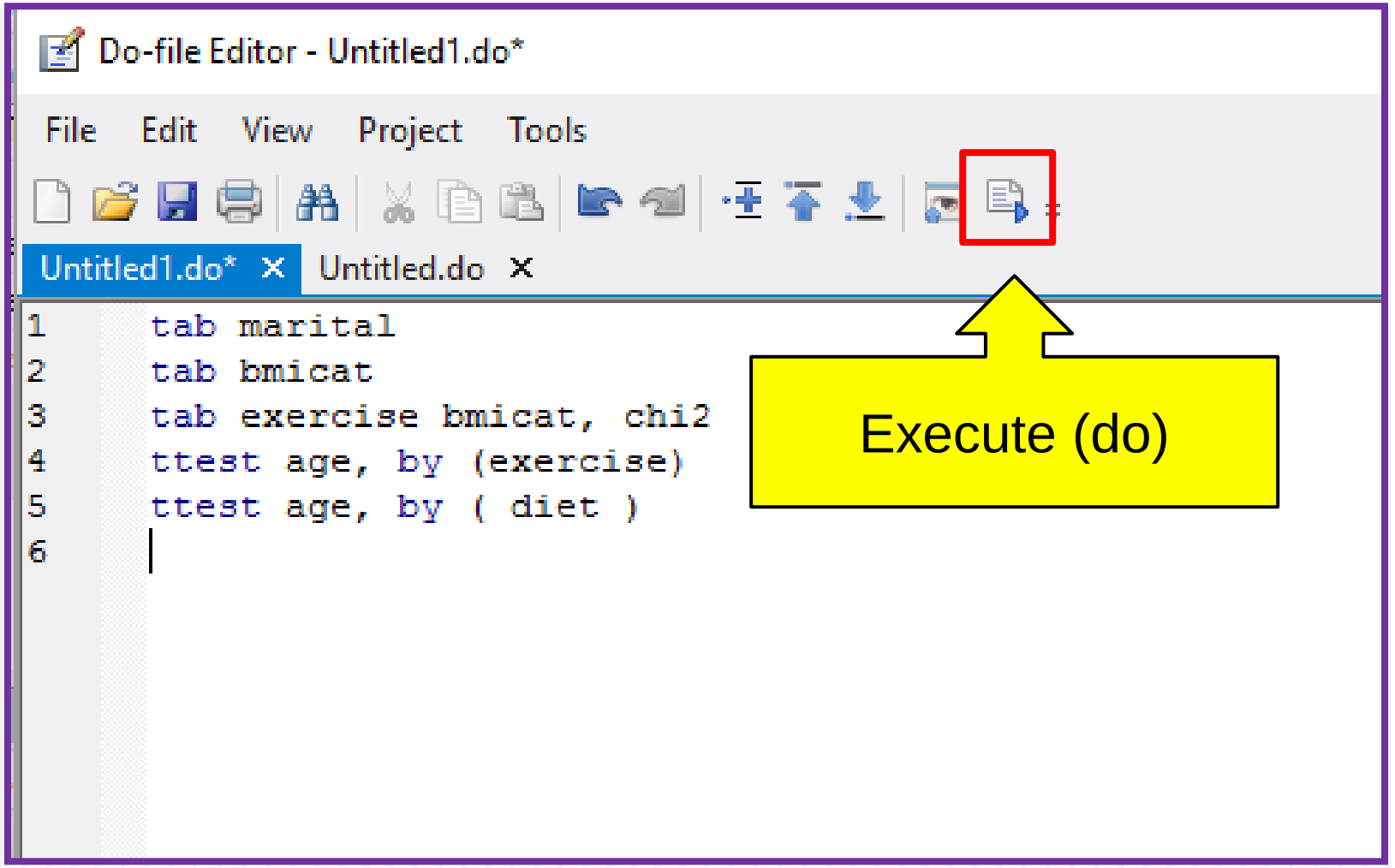
- If you want to add command in do-file that already exists:



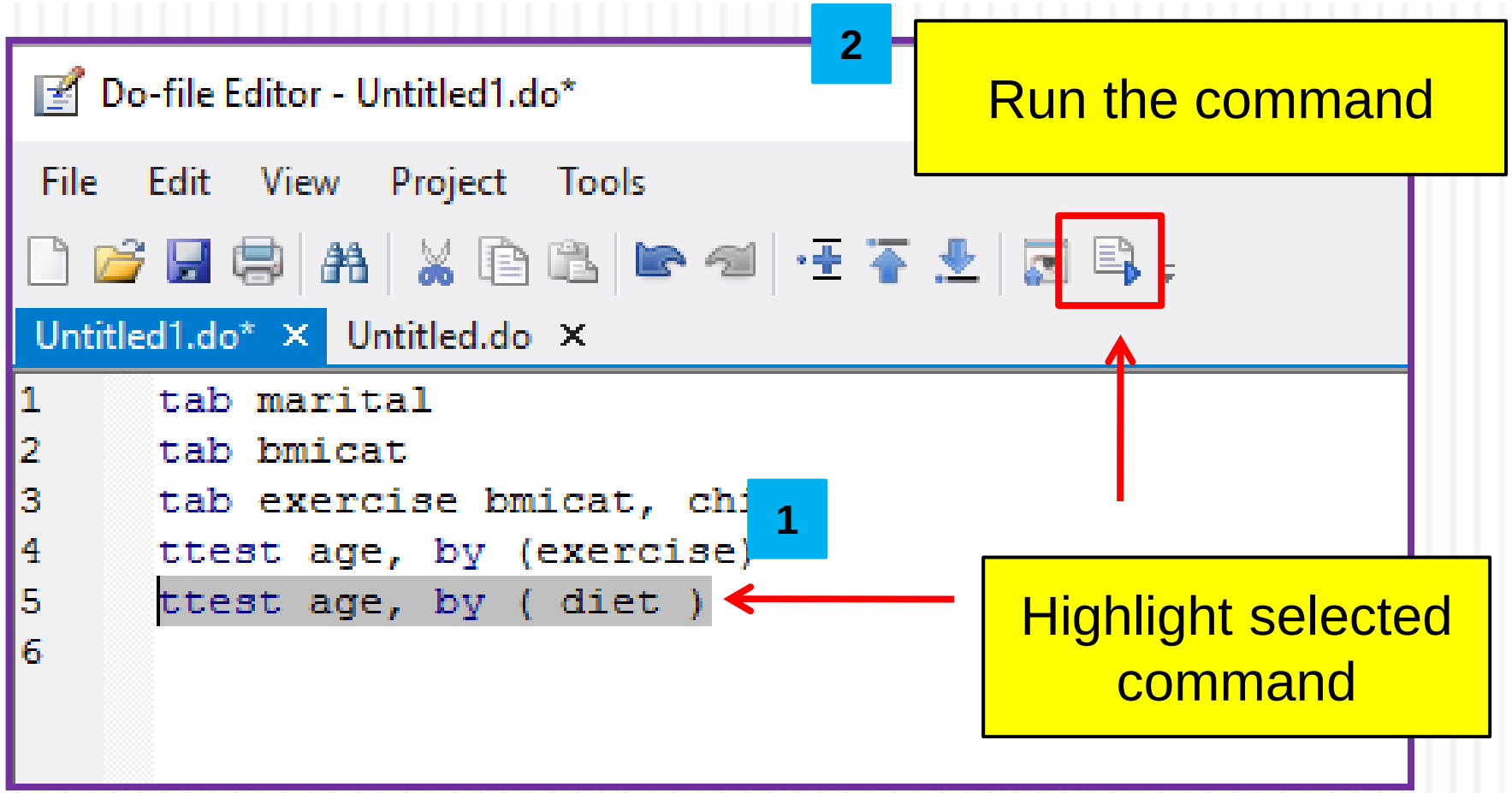
Copy the selected command and paste to the saved do-file



Run the command using the do-file editor



Run the command using the do-file editor



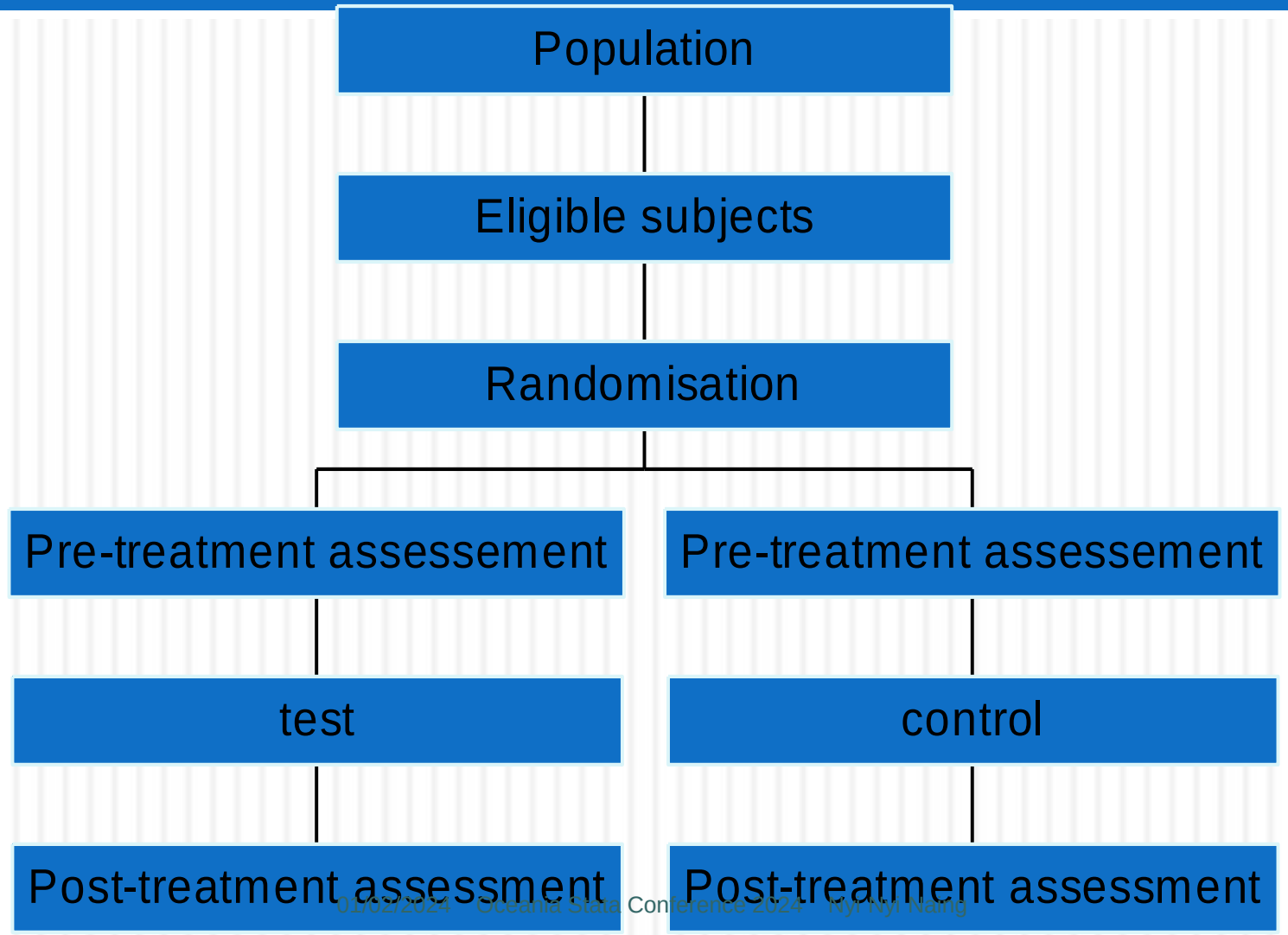
HIGHLIGHTS OF TEACHING STATA

Types of study designs

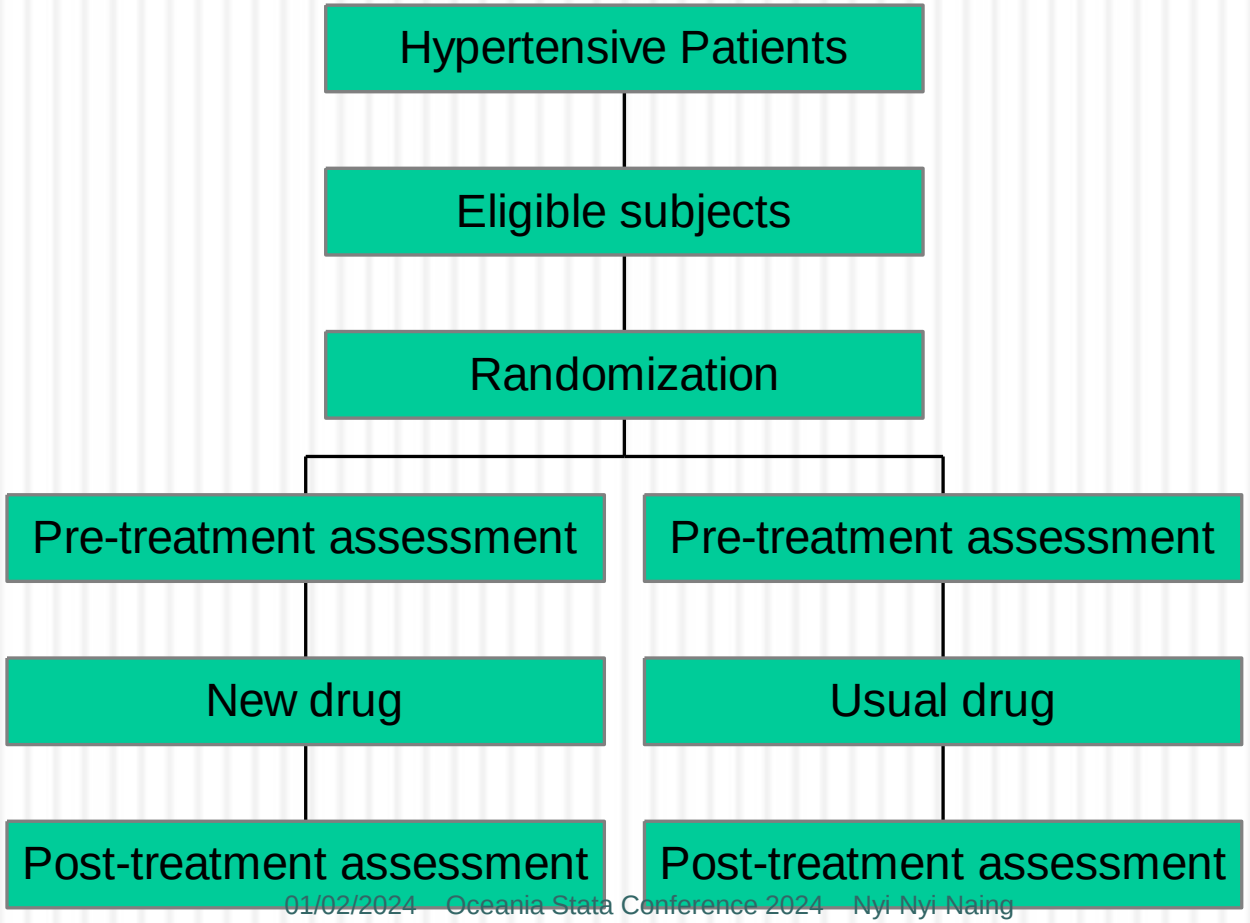
Overview of *study designs*

- **Systematic review and meta analysis**
- **Interventional studies**
  - clinical trials & community trials
- **Observational (Analytical) Studies**
  - Cohort studies
    - prospective, retrospective and historical
  - Case-control studies
    - matched, unmatched and nested
  - Cross-sectional studies
    - classical, comparative, diagnostic test, questionnaire
- **Observational (Descriptive) studies**
  - population : prevalence, incidence studies
  - individuals: case reports, case series

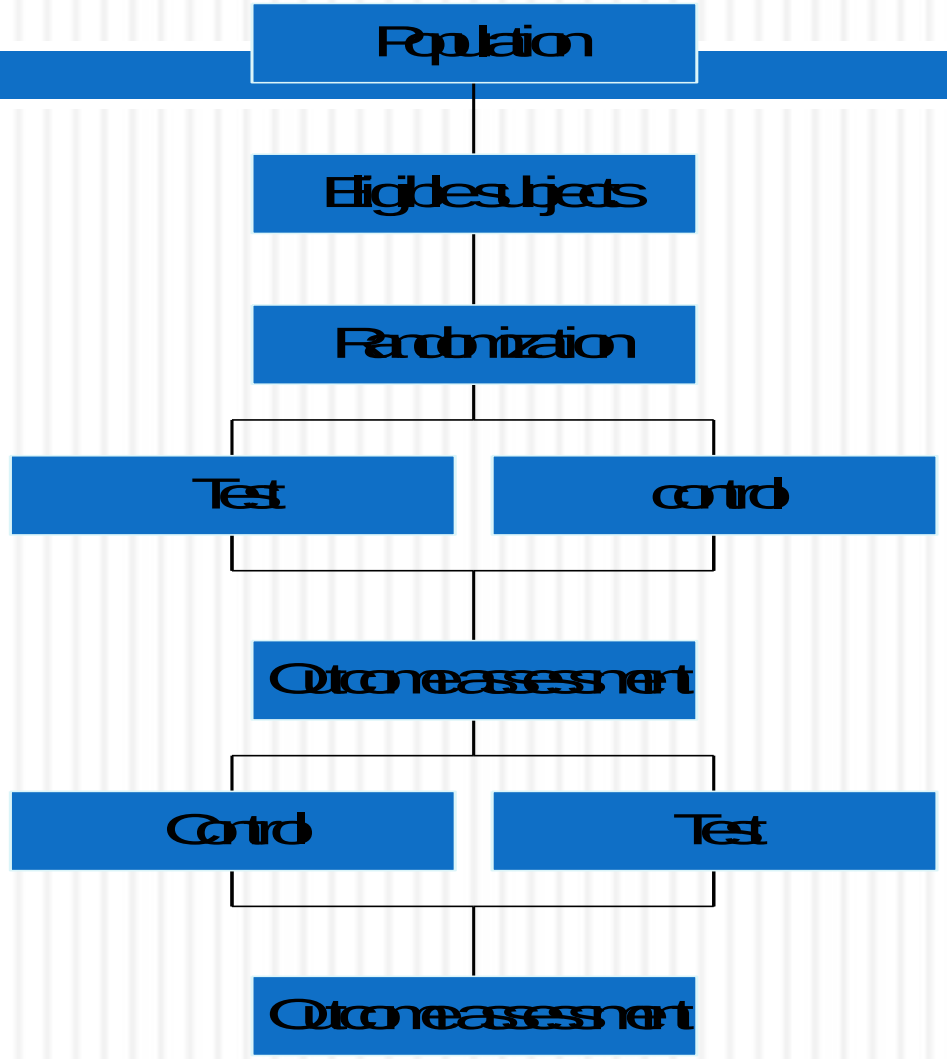
Parallel


Population
Eligible subjects
Randomisation
Pre-treatment assessement
Pre-treatment assessement
test
control
Post-treatment assessment
Post-treatment assessment

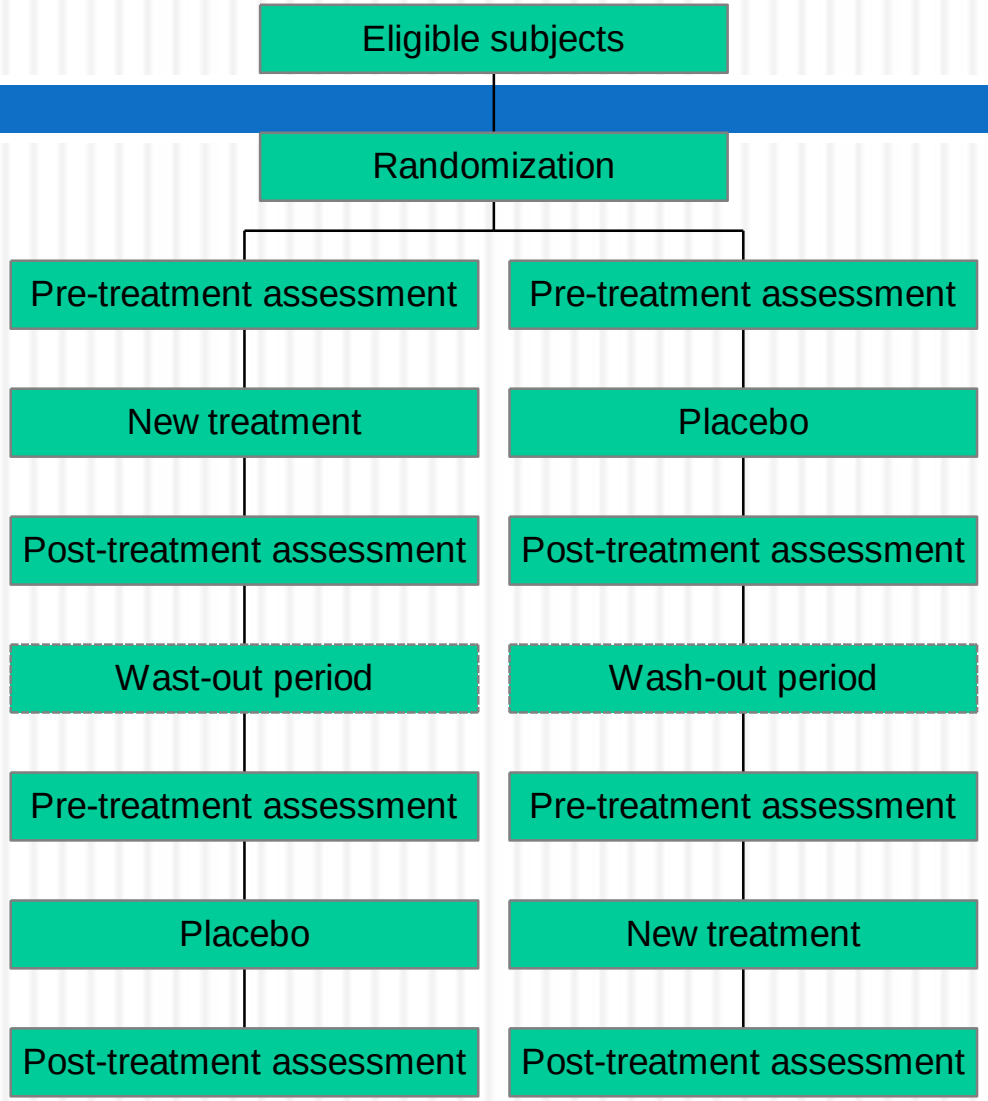
Example


Hypertensive Patients
Eligible subjects
Randomization
Pre-treatment assessment
Pre-treatment assessment
New drug
Usual drug
Post-treatment assessment
Post-treatment assessment

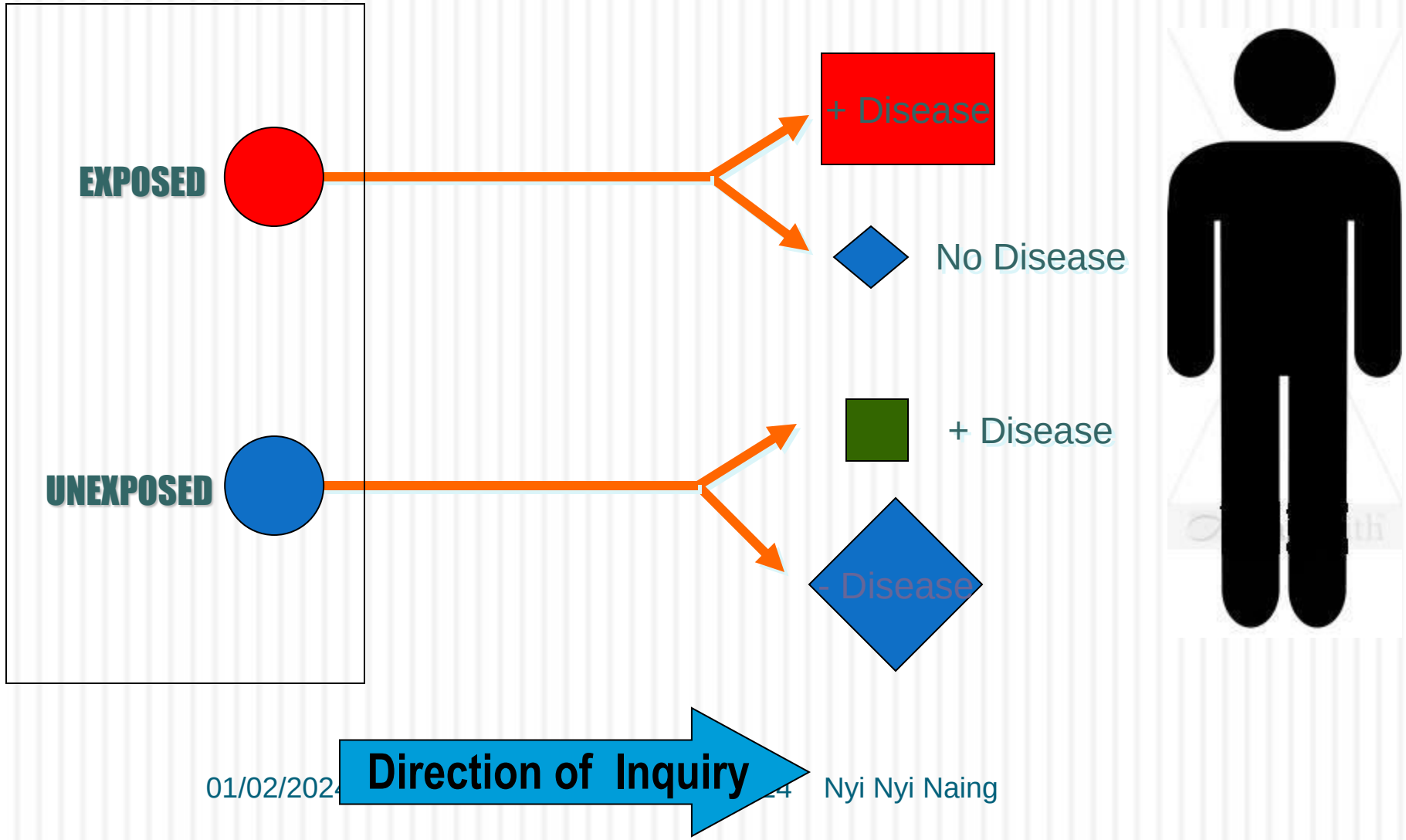
Crossover


Population
Eligible subjects
Randomization
Test
control
Outcome assessment
Control
Test
Outcome assessment

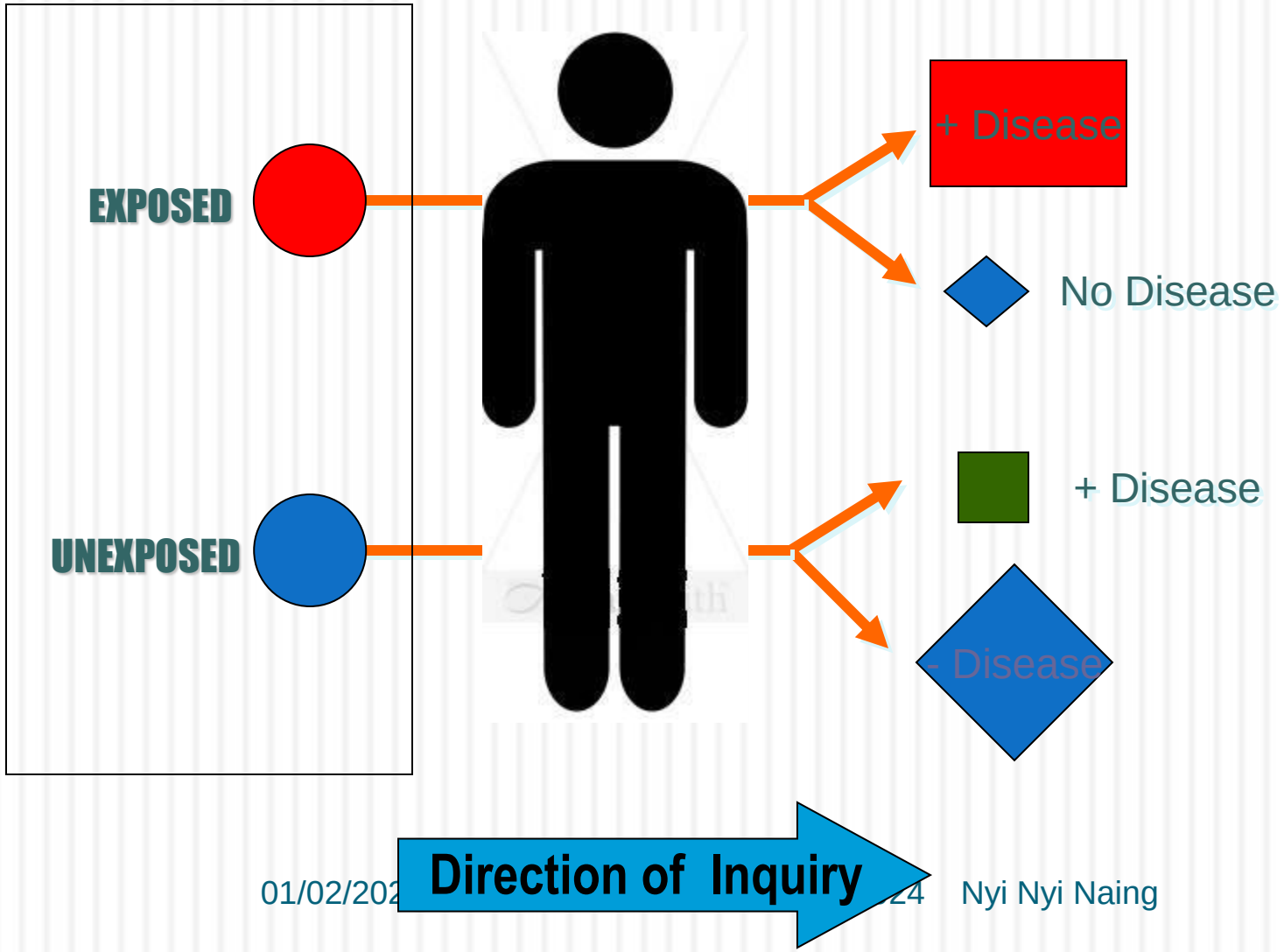
Cross-over


Eligible subjects
Randomization
Pre-treatment assessment
Pre-treatment assessment
New treatment
Placebo
Post-treatment assessment
Post-treatment assessment
Wast-out period
Wash-out period
Pre-treatment assessment
Pre-treatment assessment
Placebo
New treatment
Post-treatment assessment
Post-treatment assessment

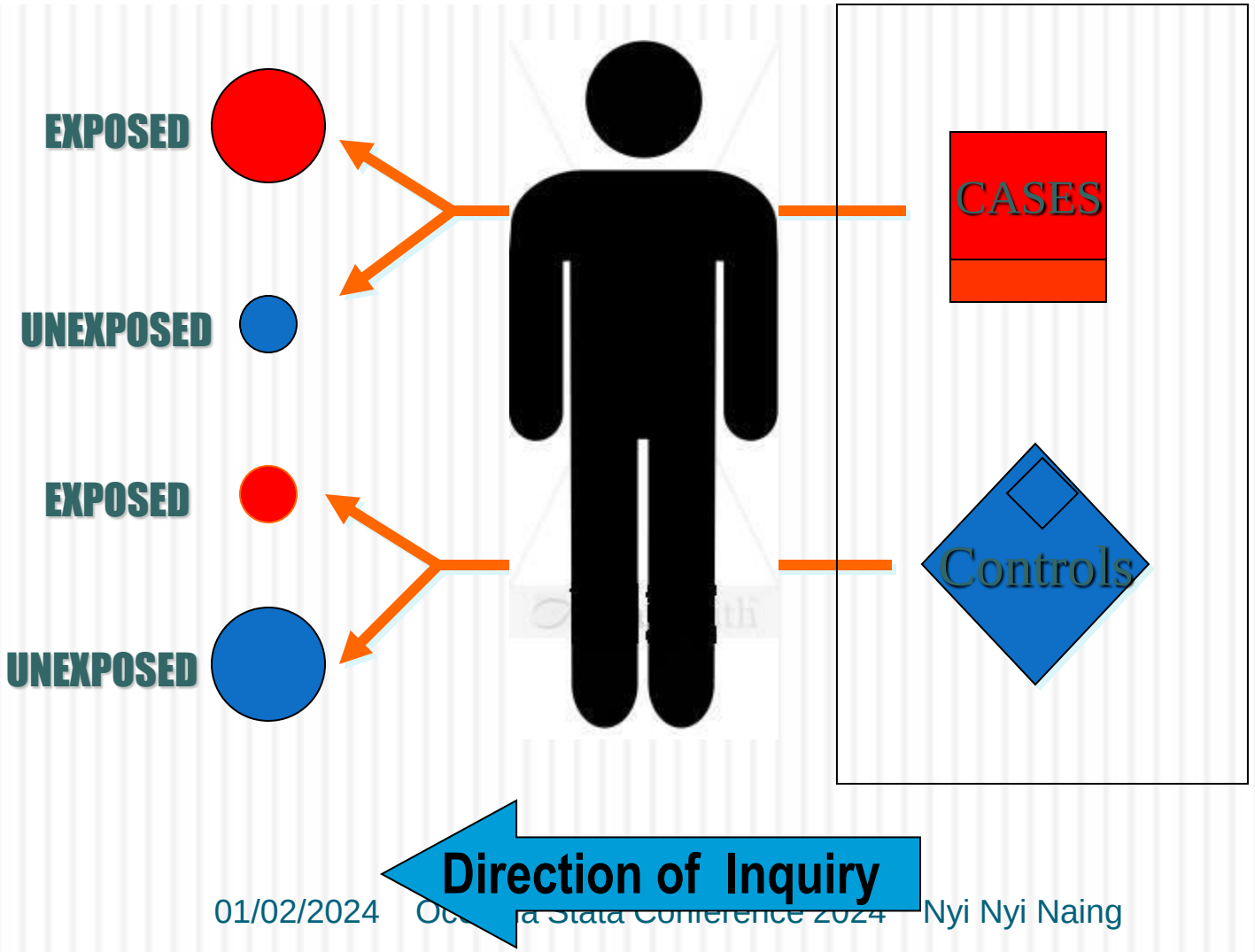
Design of a Prospective Cohort Study

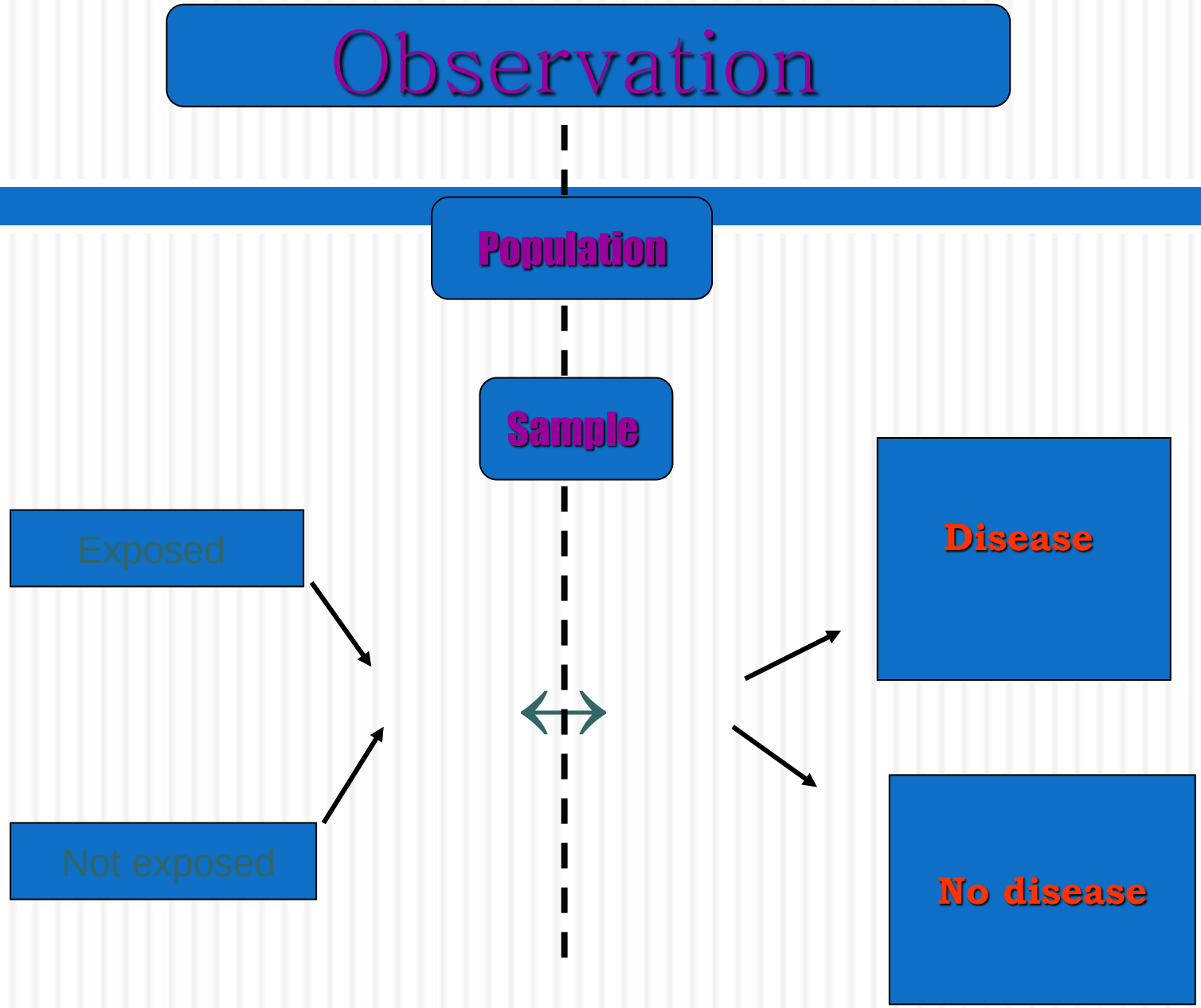


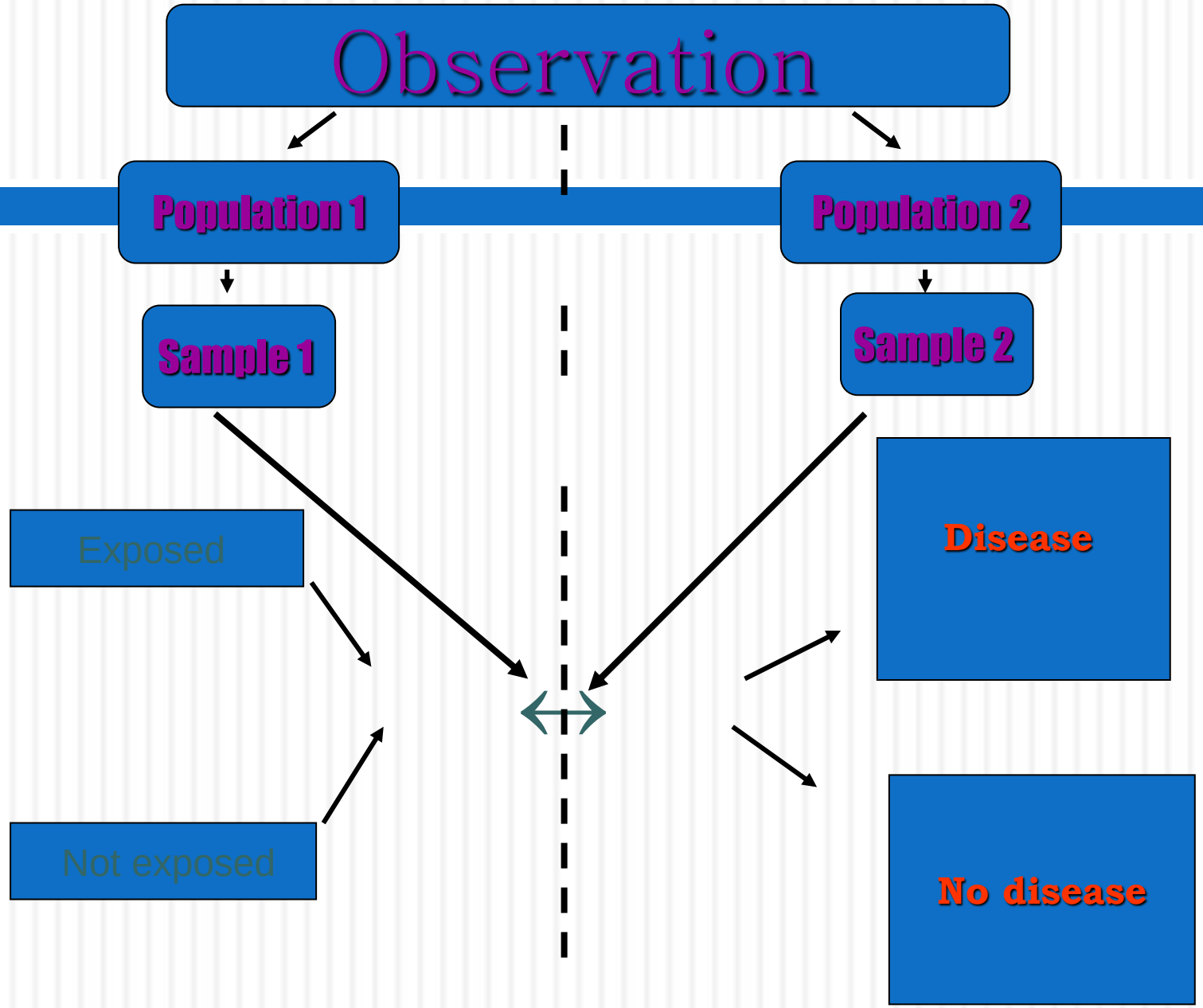
Design of a Retrospective Cohort Study



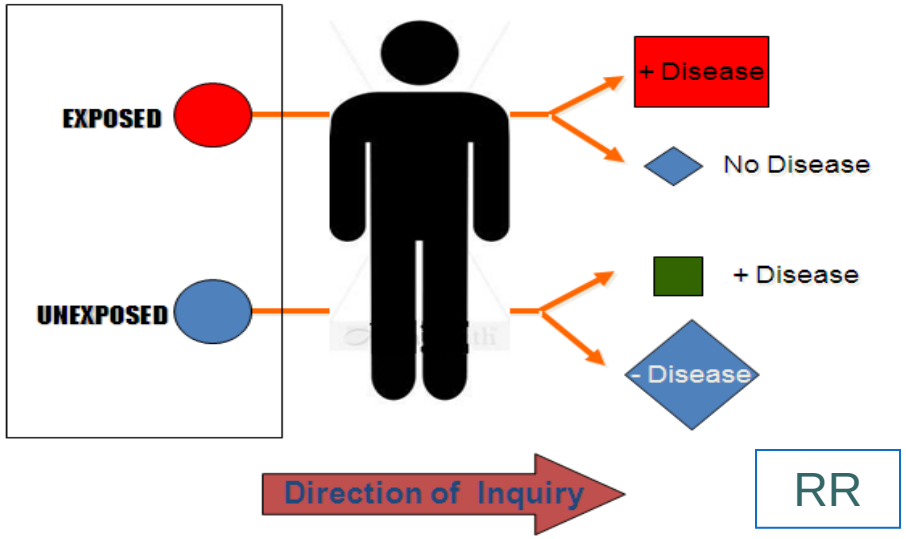
Design of a Case-Control study




Observation
Population
Sample
Exposed
Not exposed
Disease
No disease

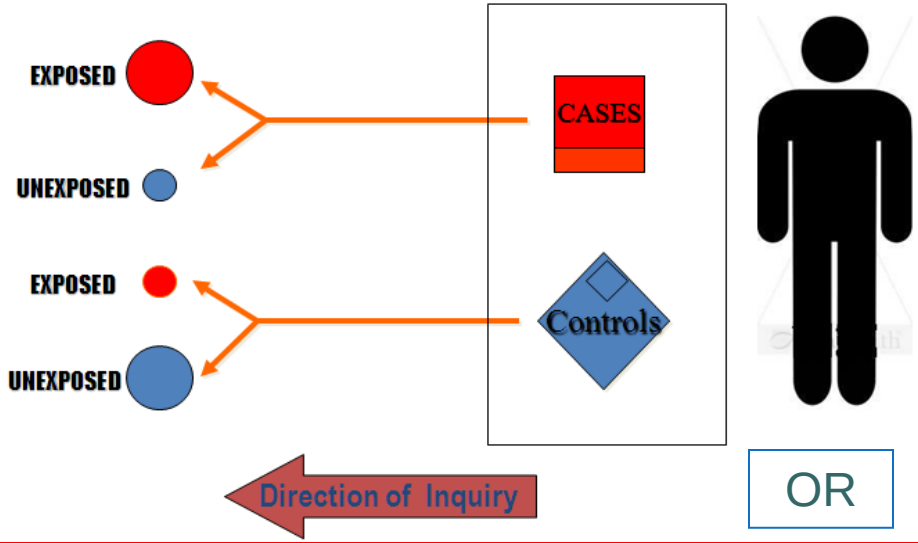

Observation
Population 1
Population 2
Sample 1
Sample 2
Exposed
Not exposed
Disease
No disease

Design of a Prospective Cohort Study



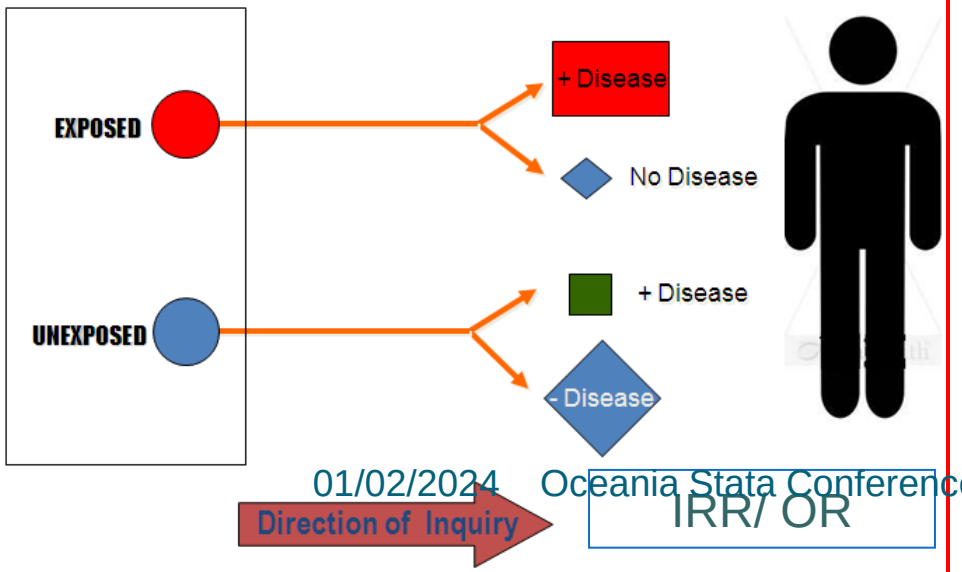
RR

Design of a Case-Control study



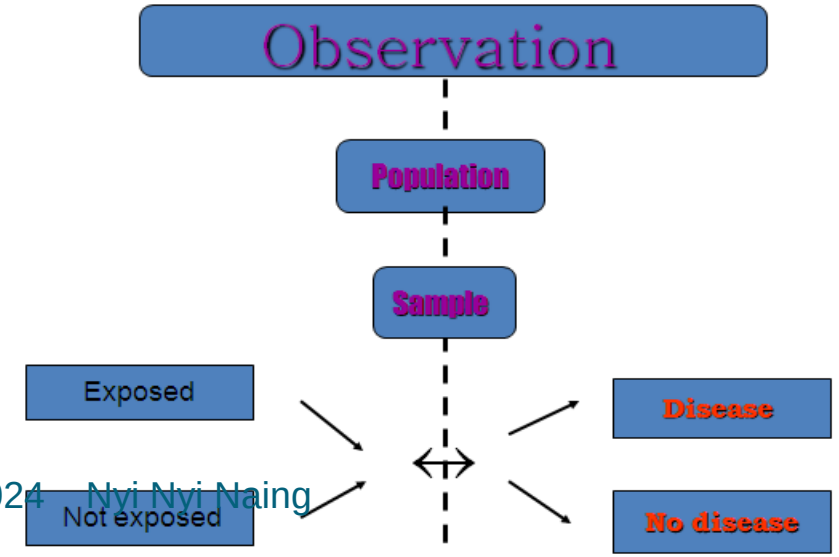
OR

Design of a Retrospective Cohort Study (RRR)



IRR/ OR

Design of a Cross-Sectional study



HIGHLIGHTS OF TEACHING STATA

Sample size determination

Estimated sample size for two samples with repeated measures
Assumptions:

alpha = 0.0500 (two-sided)
power = 0.9000
m1 = 25
m2 = 20
sd1 = 5
sd2 = 10
n2/n1 = 1.00
number of follow-up measurements = 3
correlation between follow-up measurements = 0.500
number of baseline measurements = 0

Method: POST
relative efficiency = 1.500
adjustment to sd = 0.816
adjusted sd1 = 4.082
adjusted sd2 = 8.165

Estimated required sample sizes:

n1 = 36
n2 = 36

HIGHLIGHTS OF TEACHING STATA

Examples of analyses

A general guide to select appropriate statistical methods in multivariable and multivariate analyses

Number of dependent variables	Independent variables	Dependent variable	Statistical test
One (Multivariable univariate)	Numerical and categorical	Numerical	- Multiple Linear Regression
	Numerical and categorical	Categorical (dichotomous)	- Multiple/ Binary Logistic Regression
	Numerical and categorical	Categorical (polytomous-nominal)	- Multinomial Logistic Regression
	Numerical and categorical	Categorical (ordinal)	- Ordinal Logistic Regression
	Numerical and categorical	Categorical (matched)	- Conditional Logistic Regression
	Numerical and categorical	Time to event	- Cox Proportional Hazards Regression
	Factors and covariates	Count/ Rate	- Loglinear Regression/ Poisson Regression
	One factor	Numerical	- [a] One-way ANOVA
	Two factors	Numerical	- Two-way ANOVA
	Three and more factors	Numerical	- Multi-factorial ANOVA
	Factors (with controlled covariates and factors)	Numerical	- ANCOVA
More than one (Multivariable multivariate)	Factors	Numerical	- MANOVA
	Factors (with controlled covariates and factor)	Numerical	- MANCOVA
	Factors	Numerical (repeated measures)	- Repeated Measures ANOVA
	Factors (with controlled covariates and factors)	Numerical (repeated measures)	- Repeated Measures ANCOVA
	Factors	Numerical (repeated measures)	- Repeated Measures MANOVA
	Factors (with controlled covariates and factors)	Numerical (repeated measures)	- Repeated Measures MANCOVA
	Factors and covariates	Categorical (repeated measures)	- Cross-sectional Time Series (Xt)

[a] One independent variable only

STEPS IN ADVANCED SURVIVAL ANALYSIS

1. Data exploration and cleaning
2. Univariable analysis (Simple Cox Regression)
3. Variables selection (Multiple Cox Regression) → Preliminary main effect model
4. Checking linearity of continuous variable
5. Checking multicollinearity and interactions → Preliminary final model
6. Checking the specification error of preliminary final model
7. Checking the assumption of the model

Graphs	Tests
Hazard function plot	Scaled Schoenfeld (Separate test)
Log Minus Log plot	Unscaled Schoenfeld (Global test)
Schoenfeld residual (partial residual)	C-statistics

8. Regression diagnostic
   - → Cox-Snell residual
   - → Deviance residual
   - → Martingale residual
   - → Influential analysis
9. Remedial measures → Final model

→ Change of regression coefficient

If change is < 20% → not influential

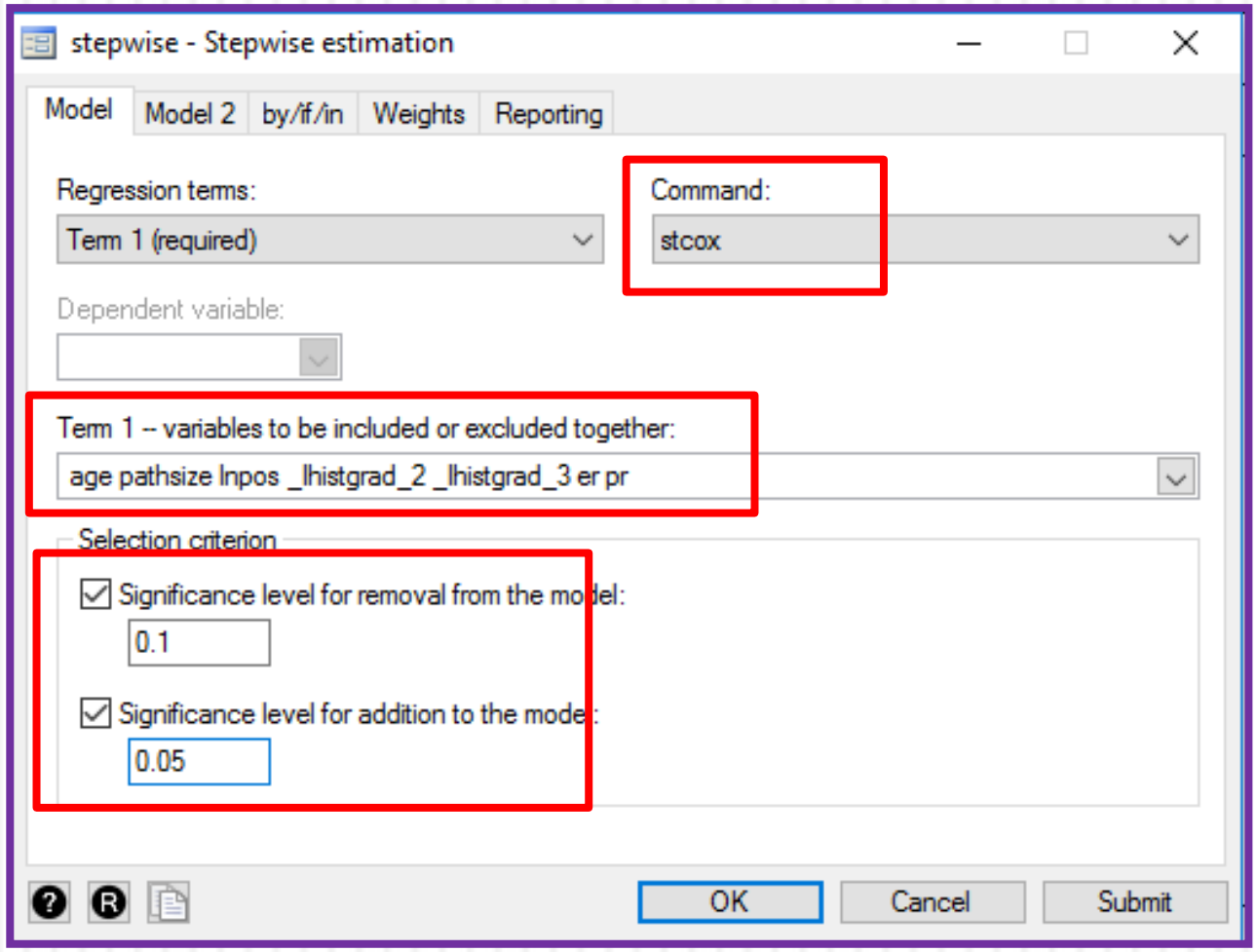
If change is ≥ 20% → influential

$$\frac{|\beta(\text{without outlier} - \beta(\text{with outlier})|}{\beta(\text{with outlier})} \times 100$$

10. Interpretation, conclusion and presentation

Step 3: variables selection (multiple cox regression)

Backward Stepwise Method

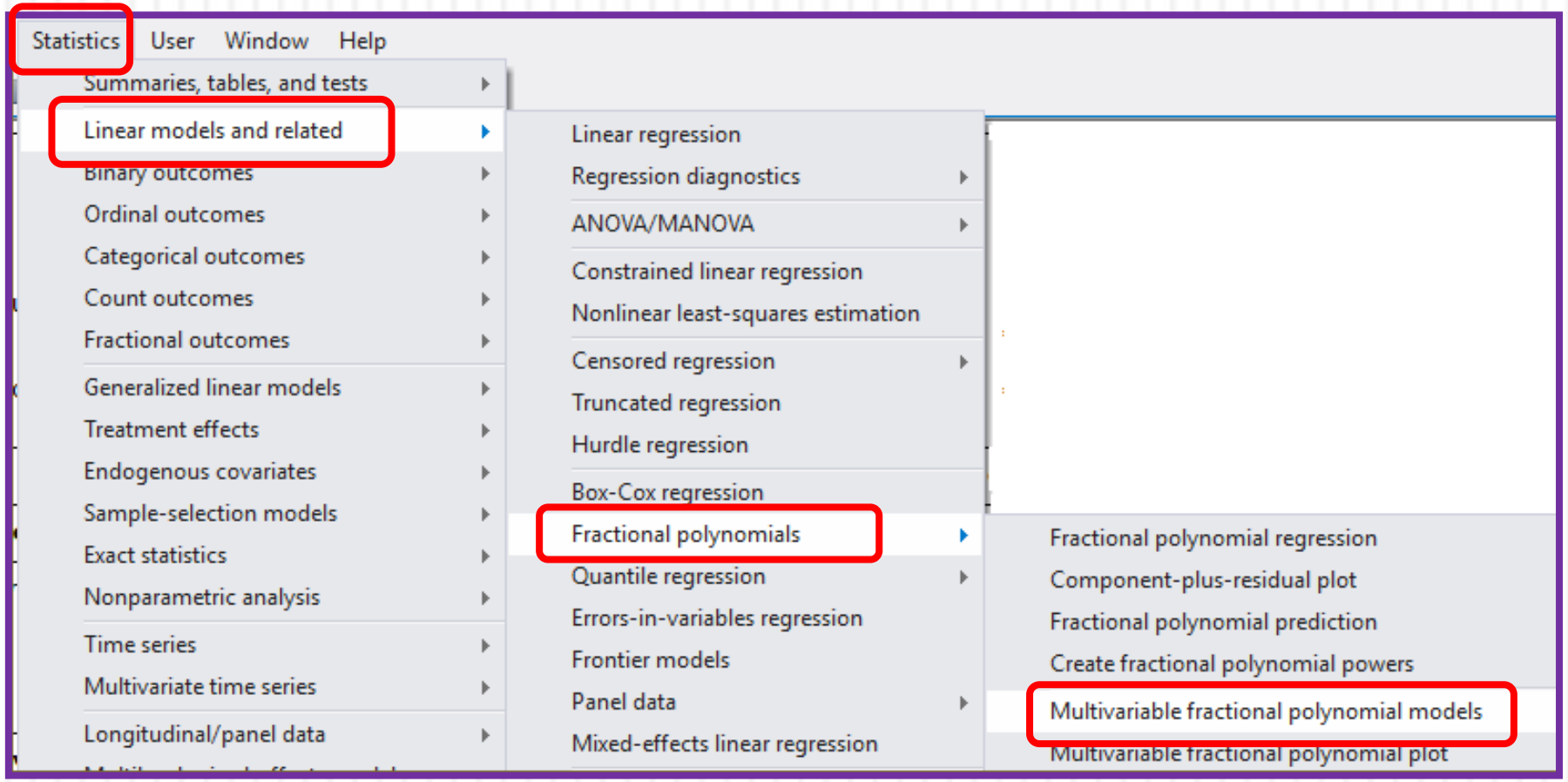


```
. stepwise, pr(0.1) pe(0.05): stcox age pathsize lnpos (_Ihistgrad_2 _Ihistgrad_3) er pr
                    begin with full model
p = 0.8567 >= 0.1000  removing er
p = 0.6686 >= 0.1000  removing _Ihistgrad_2 _Ihistgrad_3
p = 0.1121 >= 0.1000  removing age

Cox regression -- no ties

No. of subjects =          660                 Number of obs    =          660
No. of failures =           40
Time at risk    =  28866.76667
                                                LR chi2(3)       =        29.30
Log likelihood  =   -216.11104                  Prob > chi2      =       0.0000

------------------------------------------------------------------------------
          _t | Haz. Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          pr |   .5109172   .1715316    -2.00   0.045     .2645905    .9865674
    pathsize |   1.566191   .206621     3.40   0.001     1.209343    2.028338
       lnpos |   1.147063   .054402     2.89   0.004     1.045243    1.258802
------------------------------------------------------------------------------
```

Step 4: checking linearity of continuous variable

Multivariable Fractional Polynomial (mfp command)

- Statistics > Linear models and related > Fractional polynomials > Multivariable fractional polynomial models



```
. mfp, sequential : stcox pathsize lnpos pr

Deviance for model with all terms untransformed =   432.222, 660 observations

Variable     Model (vs.)    Deviance  Dev diff.    P      Powers    (vs.)
-------------------------------------------------------------------------
pathsize     FP2    FP1       428.352     1.536  0.464   -2 3       -1
             FP1    lin.      429.887     2.335  0.127   -1         1
             Final            432.222                    1

lnpos        FP2    FP1       430.826     1.092  0.579   -2 3       .5
             FP1    lin.      431.918     0.304  0.581   .5         1
             Final            432.222                    1

[pr included with 1 df in model]
```

- Based on the results, there is no significant difference in comparing linear model to the best two-term fractional polynomial model
- Thus, we can treat variables pathsize and lnpos as linear in the model

Step 7: checking the assumption of the model

Graphs	Tests
Hazard function plot - For categorical variable	Scaled Schoenfeld (Separate test)
Log Minus Log plot - For categorical variable	Unscaled Schoenfeld (Global test)
Schoenfeld residual (partial residual) - For both numerical and categorical variables	C-statistics

Partial likelihood estimate of regression coefficient

```
estat phtest, detail

    Test of proportional-hazards assumption

    Time:  Time
                 |      rho         chi2       df       Prob>chi2
    -------------+------------------------------------------------
    pathsize     |    0.03733       0.06        1          0.8097
    lnpos        |   -0.02143       0.02        1          0.8874
    pr           |    0.23658       2.22        1          0.1361
    -------------+------------------------------------------------
    global test  |                  2.24        3          0.5247
```

Separate test

Global test

- There was no evidence that the proportional hazards assumption has been violated

```
estat phtest, rank detail

    Test of proportional-hazards assumption

    Time:  Rank(t)
```

	rho	chi2	df	Prob>chi2
pathsize	0.05967	0.15	1	0.7003
lnpos	-0.02350	0.02	1	0.8766
pr	0.20575	1.68	1	0.1949
global test		1.71	3	0.6343

Separate test

Global test

- There was no evidence that the proportional hazards assumption has been violated

C-statistics

- Harrell's C or C-statistics – as the proportion of all pairs in which the predictions and outcomes were concordant
- 0.5 and below → no discrimination
- 0.7 – 0.8 → acceptable discrimination
- 0.8 – 0.9 → excellent discrimination
- 0.9 and above → outstanding discrimination

command:
- stcox *[independent]*
- estat concordance

```
estat concordance

      failure _d:  status
 analysis time _t:  time

Harrell's C concordance statistic

Number of subjects (N)                =       660
Number of comparison pairs (P)        =     14390
Number of orderings as expected (E) =     10394
Number of tied predictions (T)        =       128

         Harrell's C = (E + T/2) / P =    0.7268
                           Somers' D =    0.4535
```

- The model was able to discriminate between those who experienced the event versus those who did not by 72.7%
- It indicated the concordance between predicted and observed values of outcome was 72.7% in the preliminary final model

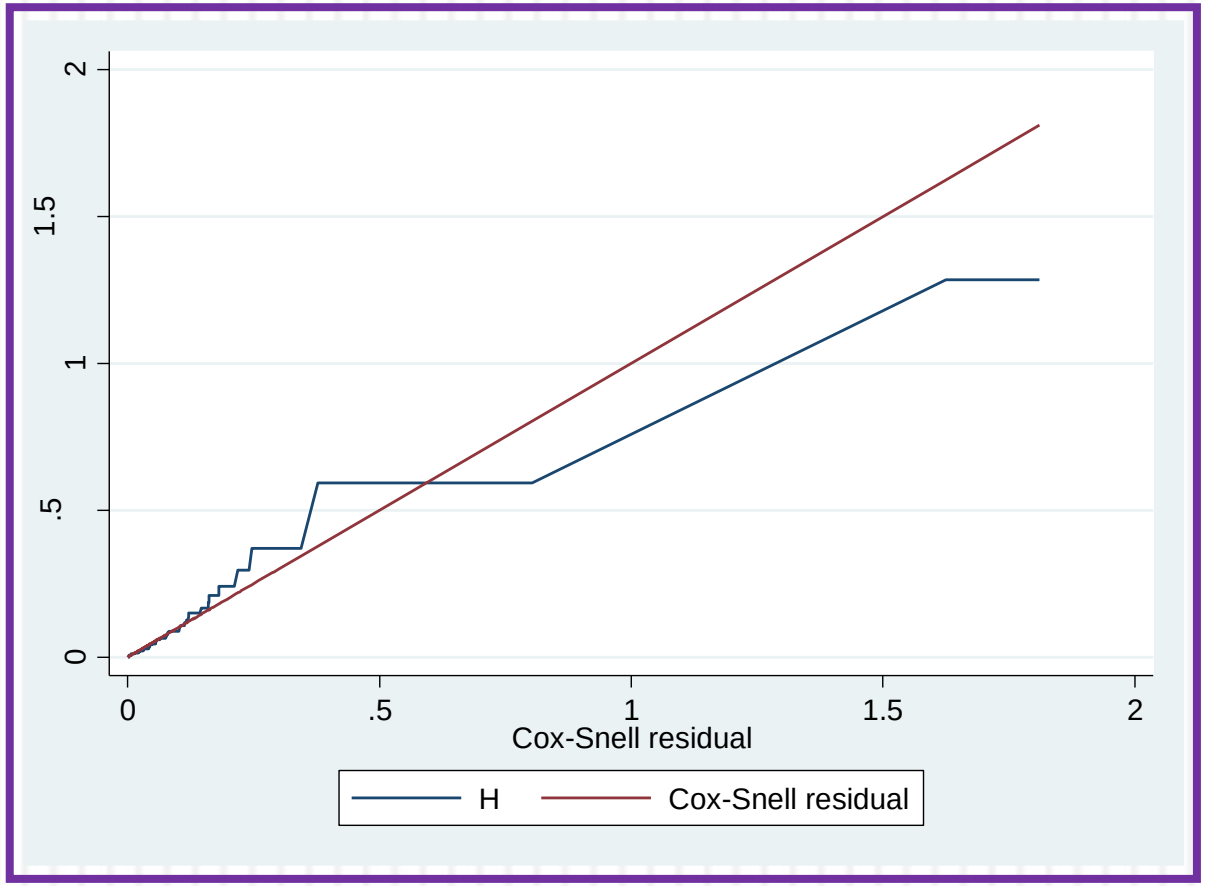
Regression diagnostic

- Cox-Snell residuals
- Martingale residuals
- Deviance residuals
- Influence analysis

Cox-Snell Residuals

- Assessing overall model fit
- Determining whether Cox regression model fits the data well
- If the Cox regression model fits the data, residuals should have a standard censored exponential distribution with hazard ratio 1
- Verification of the model's fit – calculate estimate of cumulative hazard function (using Cox-Snell residuals as the time variable and the data's original censoring variable)
- If the model fits the data, the plot of the cumulative hazard versus Cox-Snell residuals should be a straight line with slope 1

```
line H cs cs, sort clstyle(.)
```



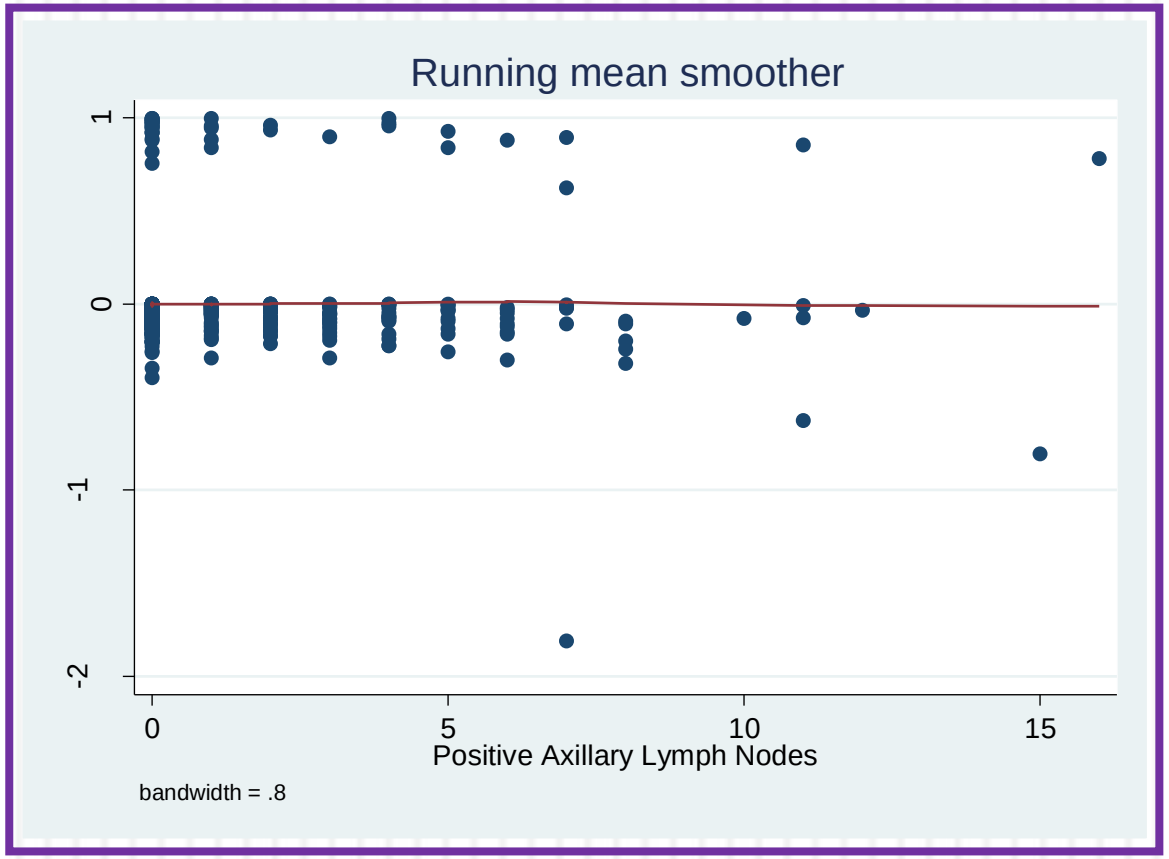
- Comparing the jagged line with reference line, it is interpreted that Cox regression model fits relatively well

Martingale Residuals

- Determining the functional form of covariates to be included in the model
- Assessing lack of fit
- Determining whether the transformation is necessary

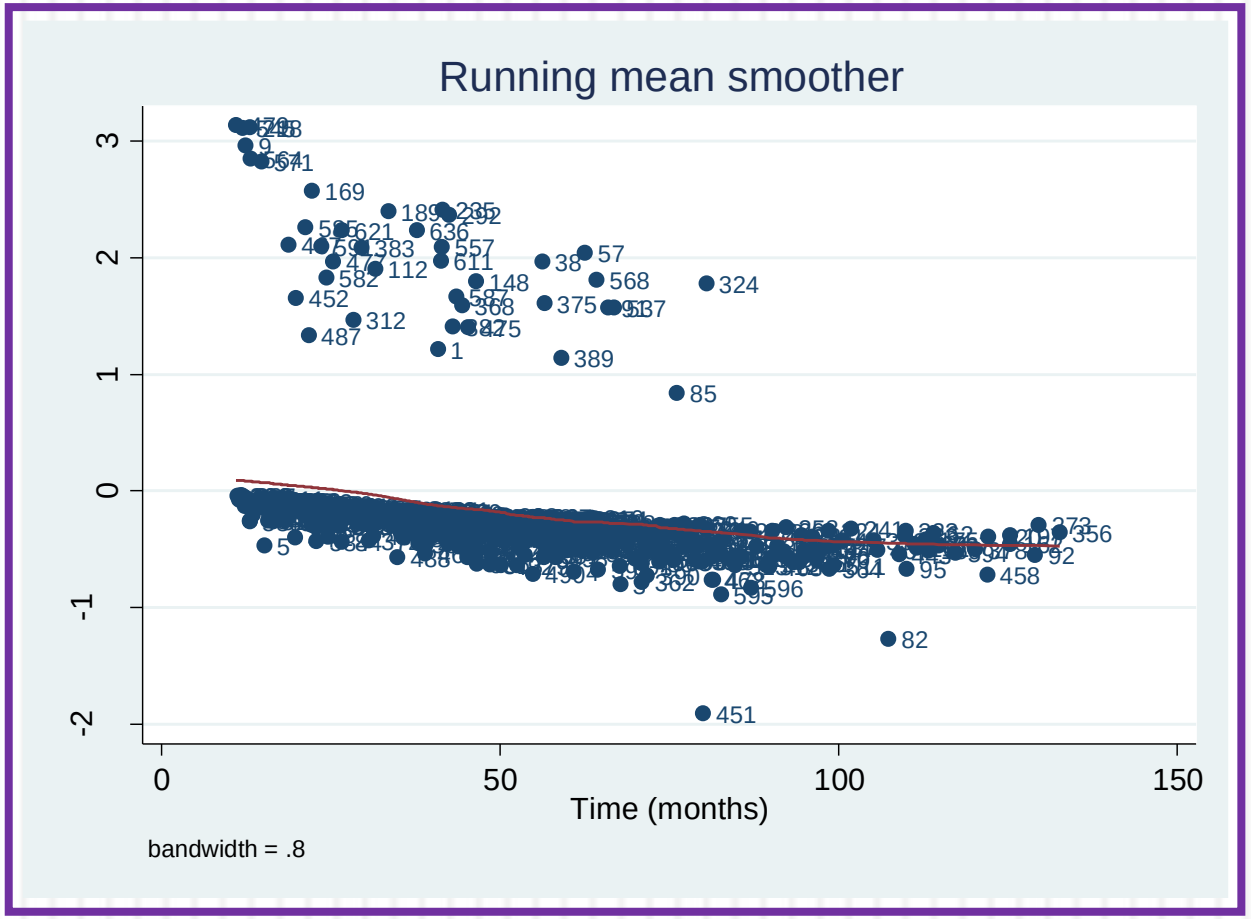
- The difference over time of the observed number of failures minus that predicted by the model

```
lowess mg lnpos , mean
```



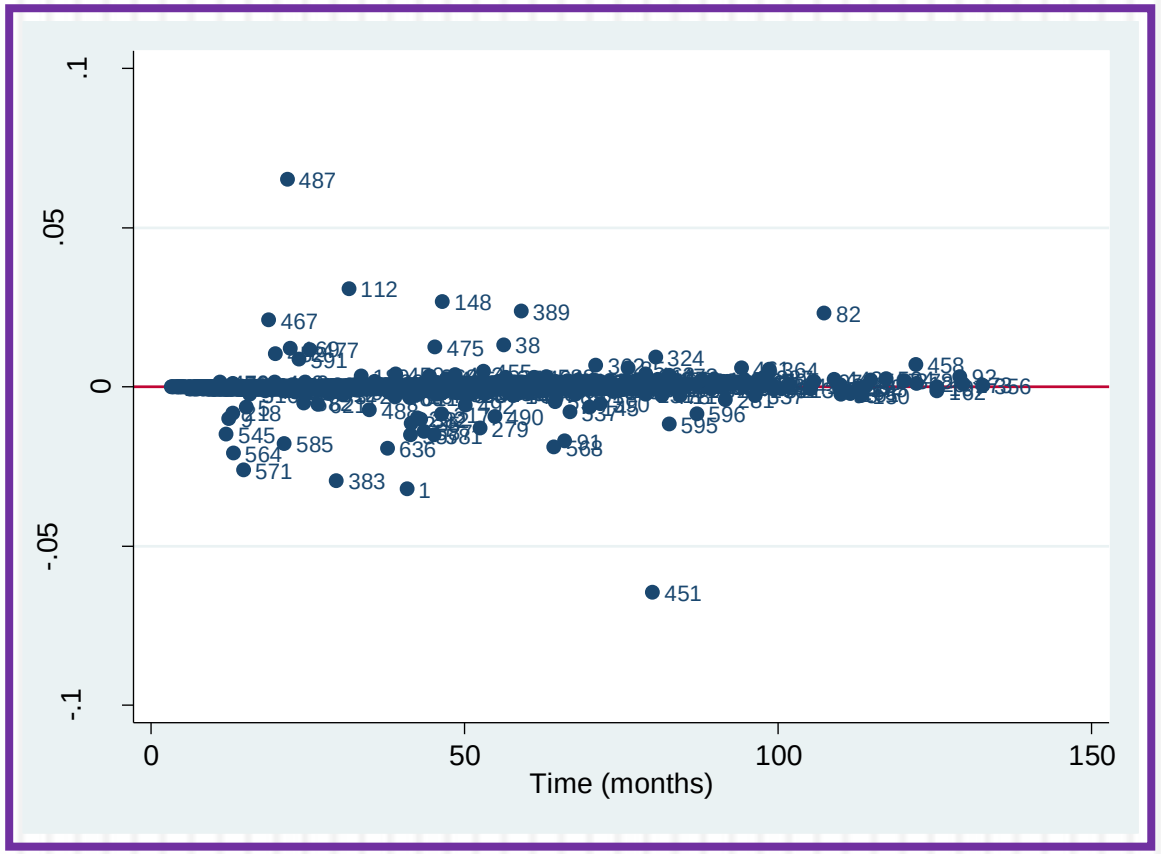
- Martingale residuals against positive axillary lymph nodes

```
lowess dev time, mean mlabel (n)
```



- Deviance residuals against time

```
predict df*, dfbeta

scatter df1 time, yline(0) mlabel (n)
```



- No observation outside the range of cut-off point ± 1

Step 9: remedial measures

Remedial measures

- Percent changes in regression coefficient ≥ 20%

$$\frac{|\beta(\text{without outlier} - \beta(\text{with outlier})|}{\beta(\text{with outlier})} \times 100$$

- Command:
- stcox *[independent]*
- stcox *[independent]* if n!=9

Table 1: Prognostic factors of breast cancer

Variables	Regression coefficient (b)	Adjusted Hazards Ratio (95% CI)	*P*-value
pathsize	0.45	1.57 (1.21, 2.03)	0.001
lnpos	0.14	1.15 (1.05,1.26)	0.004
pr			
Negative	0	1	
Positive	-0.67	0.51 (0.26, 0.99)	0.045

Backward stepwise Cox proportional hazards regression model applied.
Multicollinearity and interactions were unlikely.
The preliminary final model was properly specified.
Hazard function plot, Log-minus-log plot, Schoenfeld partial residuals plot, scaled and non-scaled Schoenfeld residuals test and C-statistics were applied to check the assumption of the model.
Regression diagnostics were performed by Cox-Snell residual, Martingale residual, deviance residual and influential analysis.
Influential outliers were identified by checking percent changes in regression coefficient set at 20%.

Steps in ordinal logistic regression

1) Data exploration and cleaning
2) Univariable analysis
3) Variables selection
4) Checking linearity of continuous variables
5) Checking multicollinearity and interactions
6) Checking the specification error of preliminary final model
7) Checking assumptions
8) Checking overall fit of the model
9) Regression diagnostic (outliers and influential)
10) Remedial measures
11) Data interpretation, conclusion and presentation

Step 4: Checking linearity of continuous variable

- Need to check the scale of continuous variable by using **three separate binary regressions of y=k vs. y=0**
- Linearity checking by using:
  - Multivariable fractional polynomial (**mfp** command)
  - **lintrend** command
  - **design variable**

- Second binary logit model


```
. mfp, sequential : ologit bwt4n lwt race smoke ui if bwt4n==0 | bwt4n==2

Deviance for model with all terms untransformed =   134.844, 105 observations

Variable     Model (vs.)    Deviance  Dev diff.    P      Powers    (vs.)
-------------------------------------------------------------------------
lwt          FP2    FP1     134.547     0.039    0.981   -2 3       3
             FP1    lin.    134.586     0.257    0.612   3          1
             Final          134.844                      1

[ui included with 1 df in model]

[smoke included with 1 df in model]

[race included with 1 df i

Fractional polynomial fitting algorit                    ed afte
```

Deviance

P-value

Dev. diff

Power

- Third binary logit model


```
. mfp, sequential : ologit bwt4n lwt race smoke ui if bwt4n==0 | bwt4n==3

Deviance for model with all terms untransformed =   107.468, 105 observations

[smoke included with 1 df in model]

[race included with 1 df in model]

[ui included with 1 df in model]

lwt           FP2     FP1      106.886     0.078   0.962   -2 3      -2
              FP1     lin.     106.964     0.504   0.478   -2        1
              Final            107.468                     1


Fractional polynomial fitting algorithm converged after 1 cycle.
```

Deviance

P-value

Dev. diff

Power

Step 6: Checking the specification error of preliminary final model

- Command “linktest” is used to detect a specification error
- The link test is based on the idea of if the model is properly specified, any additional independent variables that are statistically significant should not be found except by chance

- The link test uses the linear predicted value (_hat) and linear predicted value squared (_hatsq)
- _hat should be statistically significant as it was predicted value from the model
- _hatsq should not be statistically significant in order for the link function to be correctly specified

- Run the preliminary final model
- Run the linktest

command:
ologit *[dependent] [independent]*
linktest

```
. linktest

Iteration 0:   log likelihood = -259.65219
Iteration 1:   log likelihood = -239.88318
Iteration 2:   log likelihood =  -239.6969
Iteration 3:   log likelihood = -239.69642
Iteration 4:   log likelihood = -239.69642

Ordered logistic regression                     Number of obs     =        189
                                                LR chi2(2)        =      39.91
                                                Prob > chi2       =     0.0000
Log likelihood = -239.69642                     Pseudo R2         =     0.0769

------------------------------------------------------------------------------
       bwt4n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        _hat |   .9374316   .2069318     4.53   0.000     .5318528    1.343011
      _hatsq |   .0714174    .141142     0.51   0.613    -.2052159    .3480506
-------------+----------------------------------------------------------------
       /cut1 |  -.4615616    .185475                     -.8250859   -.0980373
       /cut2 |   .5066476   .1843405                      .1453469    .8679484
       /cut3 |   1.782627   .2279891                      1.335777    2.229478
------------------------------------------------------------------------------
```

- The model was properly specified

Proportional odds assumption

- **omodel** can be used to perform a test of the assumption of constancy of effects across categories assumed in models fitted by ologit
- omodel reports an approximate likelihood-ratio test of whether the coefficients are equal across categories (test of the proportional odds assumption)
- **omodel** command is not available in Stata by default, need to install by using command: findit omodel

command: omodel logit *[dependent] [independent]*

```
. xi:omodel logit bwt4n lwt i.race smoke ui
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

Iteration 0:   log likelihood = -259.65219
Iteration 1:   log likelihood = -240.01413
Iteration 2:   log likelihood = -239.82383
Iteration 3:   log likelihood = -239.82339

Ordered logit estimates                           Number of obs   =        189
                                                  LR chi2(5)      =      39.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -239.82339                       Pseudo R2       =     0.0764

------------------------------------------------------------------------------
       bwt4n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |   .0119279   .0045958     2.60   0.009     .0029204    .0209355
    _Irace_2 |  -1.515517   .4170844    -3.63   0.000    -2.332987   -.6980462
    _Irace_3 |  -.9356021   .3283015    -2.85   0.004    -1.579061    -.292143
       smoke |  -1.090056   .3078651    -3.54   0.000    -1.693461   -.4866515
          ui |  -.9647471   .3907561    -2.47   0.014    -1.730615   -.1988793
-------------+----------------------------------------------------------------
        _cut1 |  -.4963771   .6661124          (Ancillary parameters)
        _cut2 |   .4722462   .6689451
        _cut3 |   1.742281   .6832484
------------------------------------------------------------------------------

Approximate likelihood-ratio test of proportionality of odds
across response categories:
          chi2(10) =       8.11
        Prob > chi2 =     0.6178
```

Parallel regression assumption

- **brant** performs a Brant test of the parallel regression assumption after ologit
- It reports both the results of an omnibus test for the entire model and tests of the assumption for each of the independent variables in the model
- **brant** command is not available in Stata by default, need to install by using

command: findit brant

command: ologit *[dependent] [independent],* nolog or brant, detail


```
. brant, detail

Estimated coefficients from j-1 binary regressions

                 y>0          y>1          y>2
      lwt   .01197912   .01725155   .01024191
_Irace_2   -1.3103787   -1.5469911   -2.3300021
_Irace_3   -.95574965   -.49490369   -1.3439977
   smoke   -1.0291707   -.89569729   -1.4001067
      ui    -.7784749   -1.0179173   -1.3623835
   _cons    .38657762   -1.4190499   -1.2197784

Brant Test of Parallel Regression Assumption

    Variable |      chi2    p>chi2      df
-------------+--------------------------------
         All |      9.07     0.526      10
-------------+--------------------------------
         lwt |      2.00     0.367       2
      _Irace_2 |      1.42     0.492       2
      _Irace_3 |      5.58     0.061       2
       smoke |      1.70     0.427       2
          ui |      0.59     0.746       2
-------------+--------------------------------
```

Result for the entire model

Results for each of the independent variables in the model

The parallel regression assumption is met

- The tests indicate that the assumptions are not violated
- If the assumptions are violated, need to run model as a generalized ordered logistic model using **gologit2**
- **gologit2** is not available in Stata by default, need to install by using

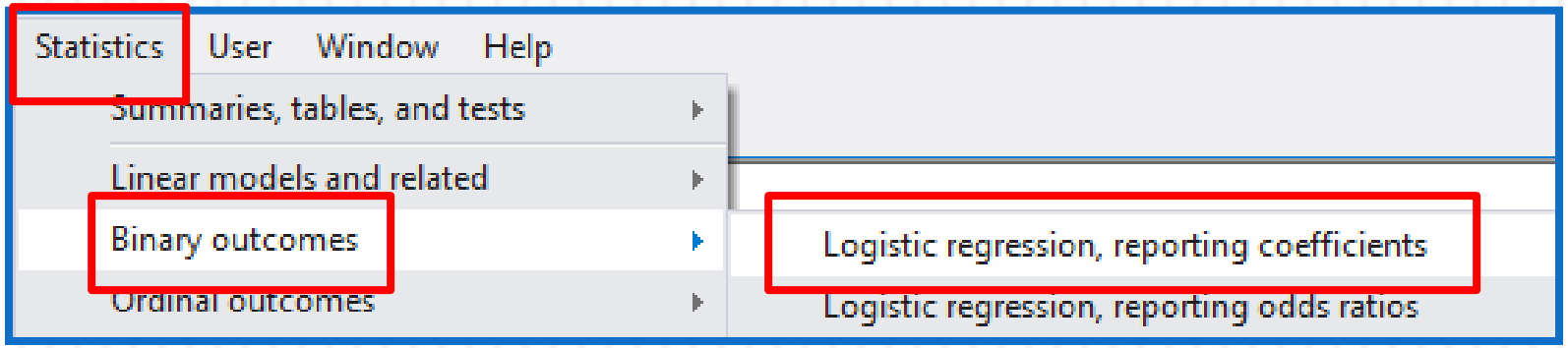
```
command: findit gologit2
```

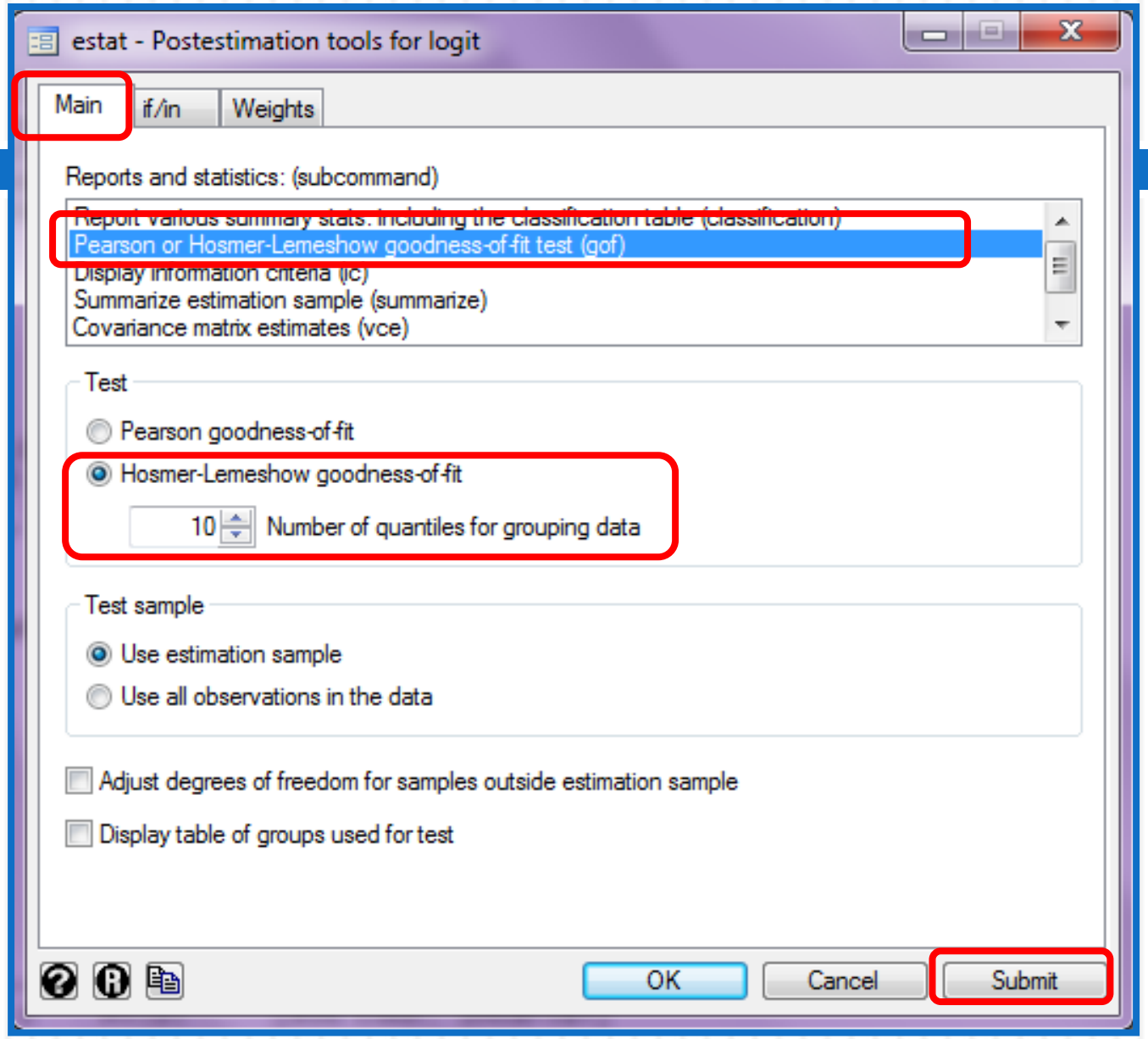
Step 8: Checking overall fit of the model

- Need to performed three separate full assessments of fit based on the binary logistic regressions of bwt4n=k vs. bwt4n=0
  - Hosmer-Lemeshow test
  - Pearson Chi-square Goodness of fit test computed by covariate pattern
  - Stukel's test (not available in Stata)
  - Classification table
  - Area under the ROC curve

Hosmer-Lemeshow test

- It is based on grouping cases into 10 groups
- It compares the observed probability with the expected probability within each group
- Check the ***P***-value. If it is >0.05, there is no significant difference between the observed probability and the expected probability. Thus, assumption is met
- Run the regression model separately for each binary logit model
  - Statistics > Binary outcomes > Logistic regression, reporting coefficients



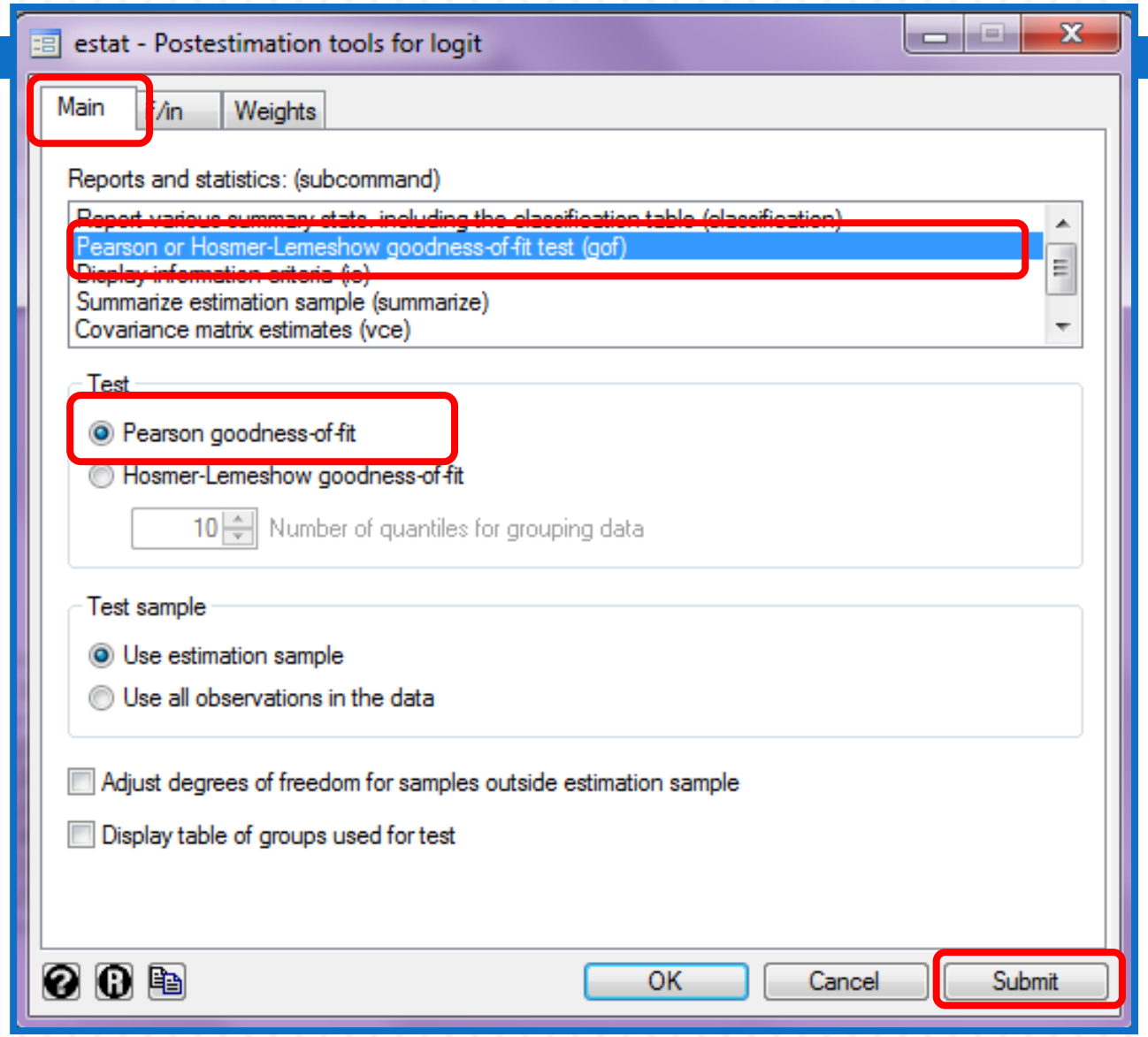

estat - Postestimation tools for logit
Main
if/in
Weights
Reports and statistics: (subcommand)
Report various summary stats. including the classification table (classification)
Pearson or Hosmer-Lemeshow goodness-of-fit test (gof)
Display information criteria (ic)
Summarize estimation sample (summarize)
Covariance matrix estimates (vce)
Test
Pearson goodness-of-fit
Hosmer-Lemeshow goodness-of-fit
10
Number of quantiles for grouping data
Test sample
Use estimation sample
Use all observations in the data
Adjust degrees of freedom for samples outside estimation sample
Display table of groups used for test
OK
Cancel
Submit

command: estat gof, group(10)

```
. estat gof, group(10)

Logistic model for bwt4n, goodness-of-fit test

  (Table collapsed on quantiles of estimated probabilities)

       number of observations =        97
             number of groups =        10
     Hosmer-Lemeshow chi2(8) =      3.59
                Prob > chi2 =    0.8921
```

- The p-value is >0.05, which is 0.892
- We can conclude that the model is fit

Pearson Chi-square Goodness of fit test



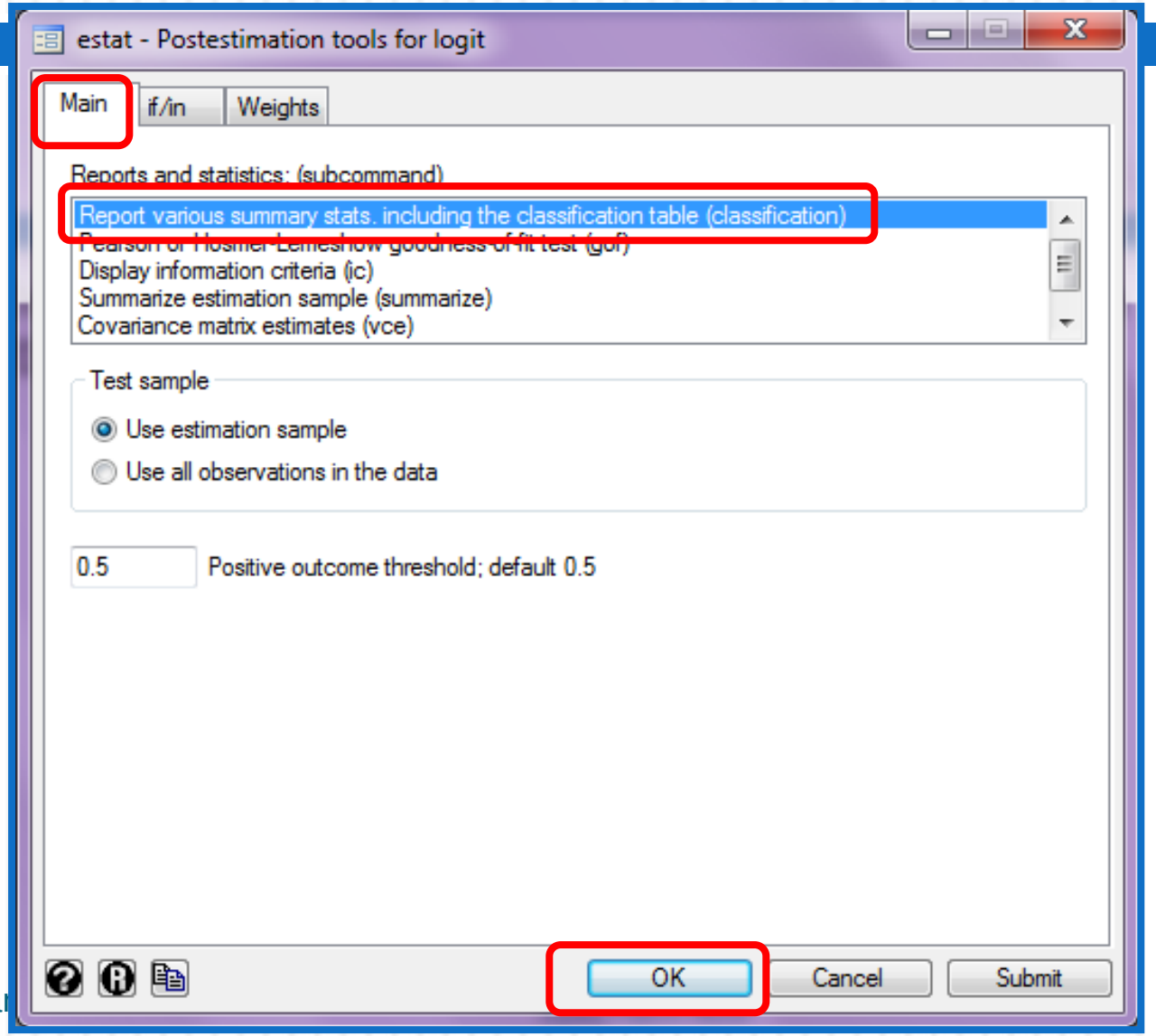
command: estat gof

```
. estat gof

Logistic model for bwt4n, goodness-of-fit test

       number of observations =        97
 number of covariate patterns =        82
            Pearson chi2(76) =     84.83
                  Prob > chi2 =    0.2285
```

- The p-value is >0.05, which is 0.229
- We can conclude that the model is fit

Classification table

- Overall correctly classified percentage is good if above 70%



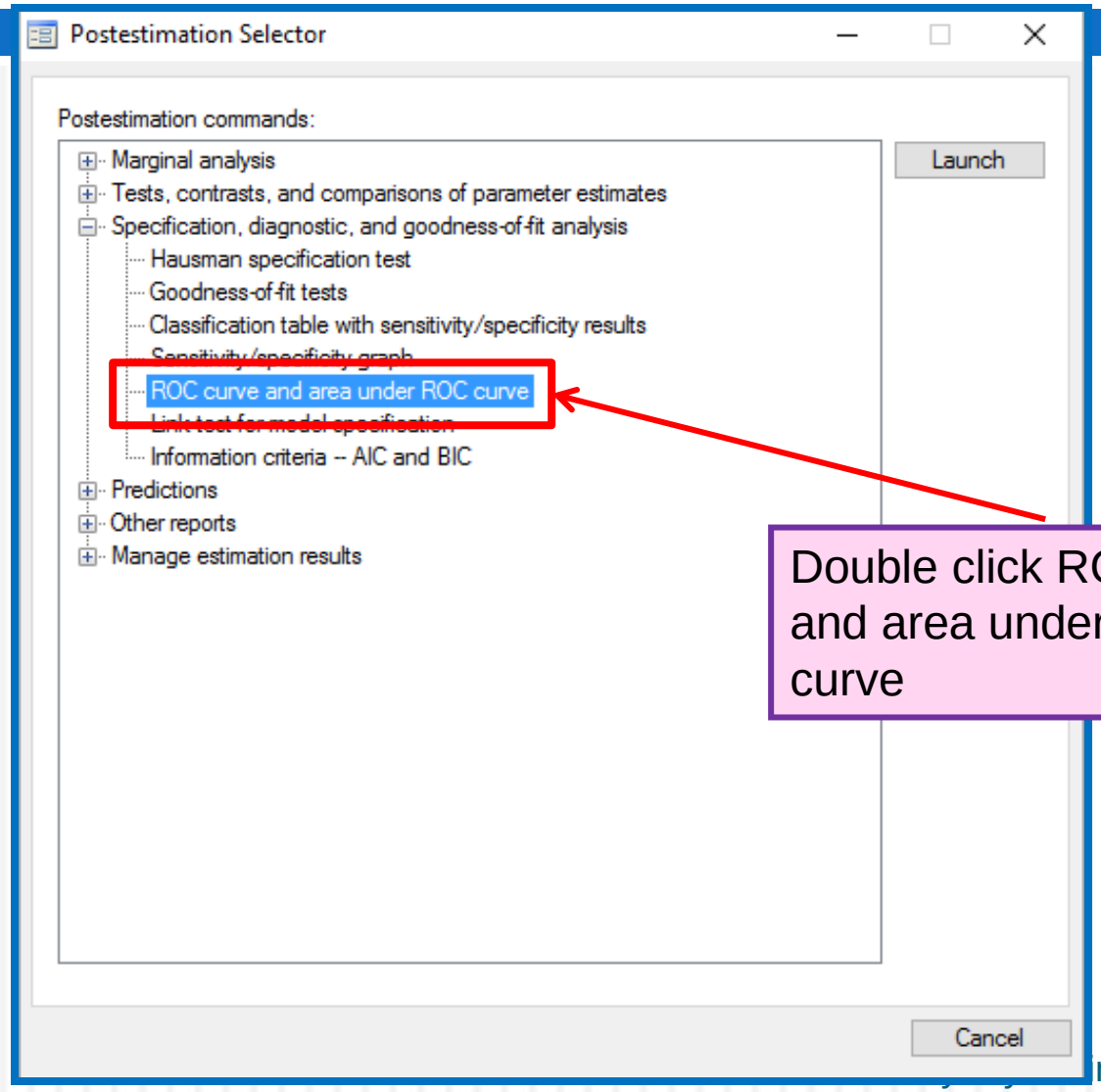
command: estat classification

```
. estat classification

Logistic model for bwt4n

                 ------- True -------
Classified |         D            ~D |      Total
-----------+--------------------------+-----------
         + |        10             7 |         17
         - |        28            52 |         80
-----------+--------------------------+-----------
     Total |        38            59 |         97

Classified + if predicted Pr(D) >= .5
True D defined as bwt4n != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   26.32%
Specificity                     Pr( -|~D)   88.14%
Positive predictive value       Pr( D| +)   58.82%
Negative predictive value       Pr(~D| -)   65.00%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)   11.86%
False - rate for true D         Pr( -| D)   73.68%
False + rate for classified +   Pr(~D| +)   41.18%
False - rate for classified -   Pr( D| -)   35.00%
--------------------------------------------------
Correctly classified                        63.92%
--------------------------------------------------
```

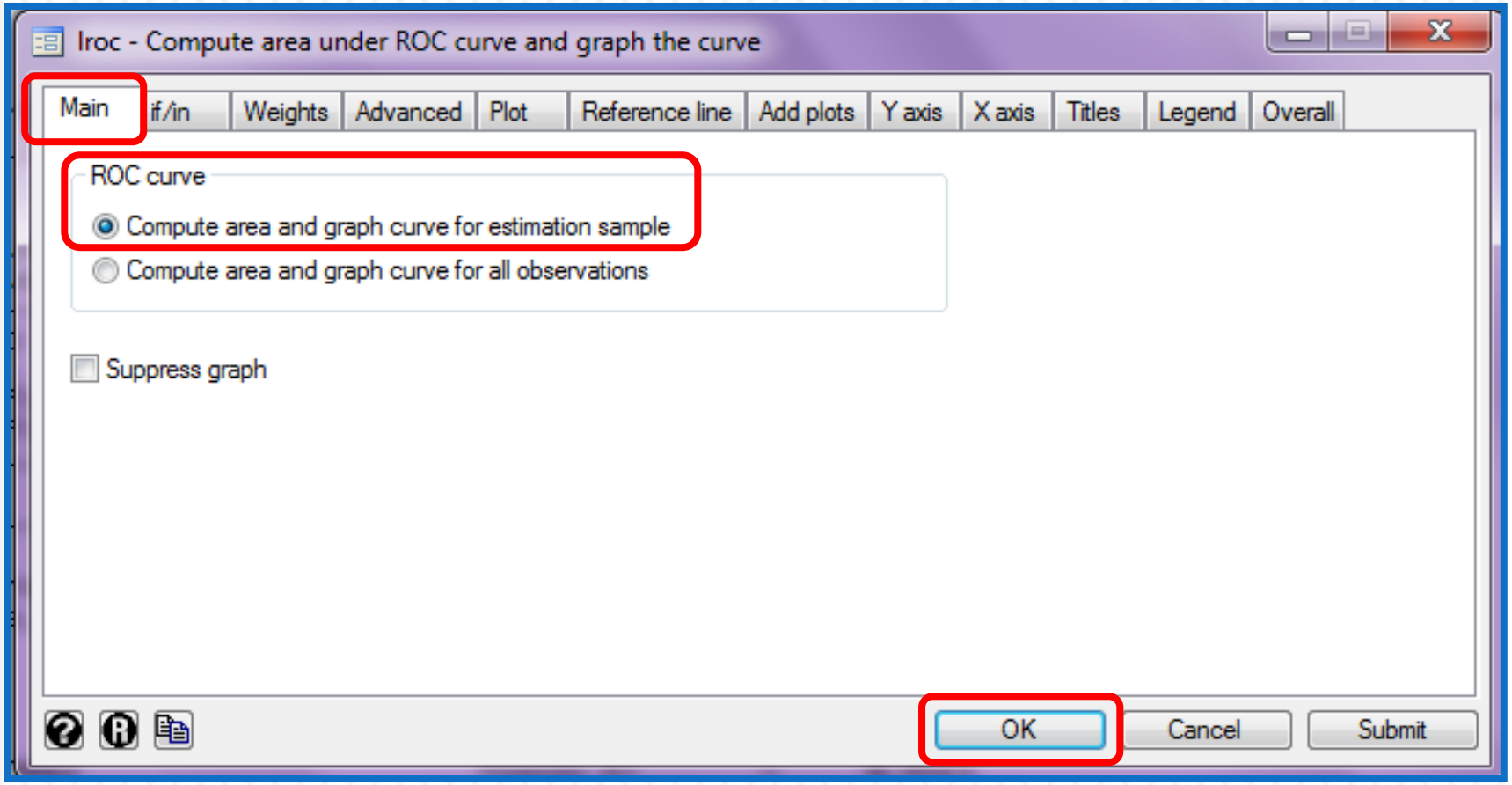
Area under the ROC curve

- Ranges from 0 to 1
- Able to assess the model discrimination
- A value of 0.5 means the model is useless for discrimination
- The recommended area under the ROC curve is at least 0.70
- Values near to 1 is better

1st method: Using lroc command
Disadvantage: 95% CI is not provided

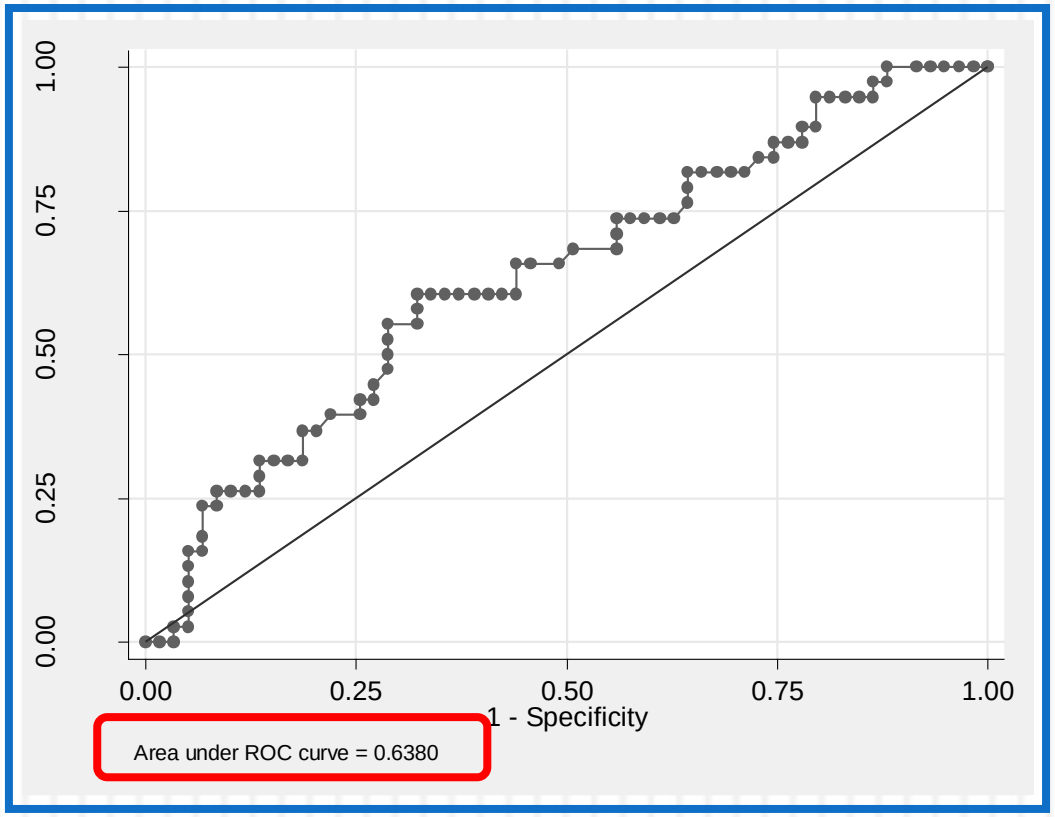


Double click ROC curve and area under ROC curve


Iroc - Compute area under ROC curve and graph the curve
Main
if/in
Weights
Advanced
Plot
Reference line
Add plots
Y axis
X axis
Titles
Legend
Overall
ROC curve
Compute area and graph curve for estimation sample
Compute area and graph curve for all observations
Suppress graph
OK
Cancel
Submit

command: lroc

```
. lroc

Logistic model for bwt4n

number of observations =       97
area under ROC curve   =   0.6380
```

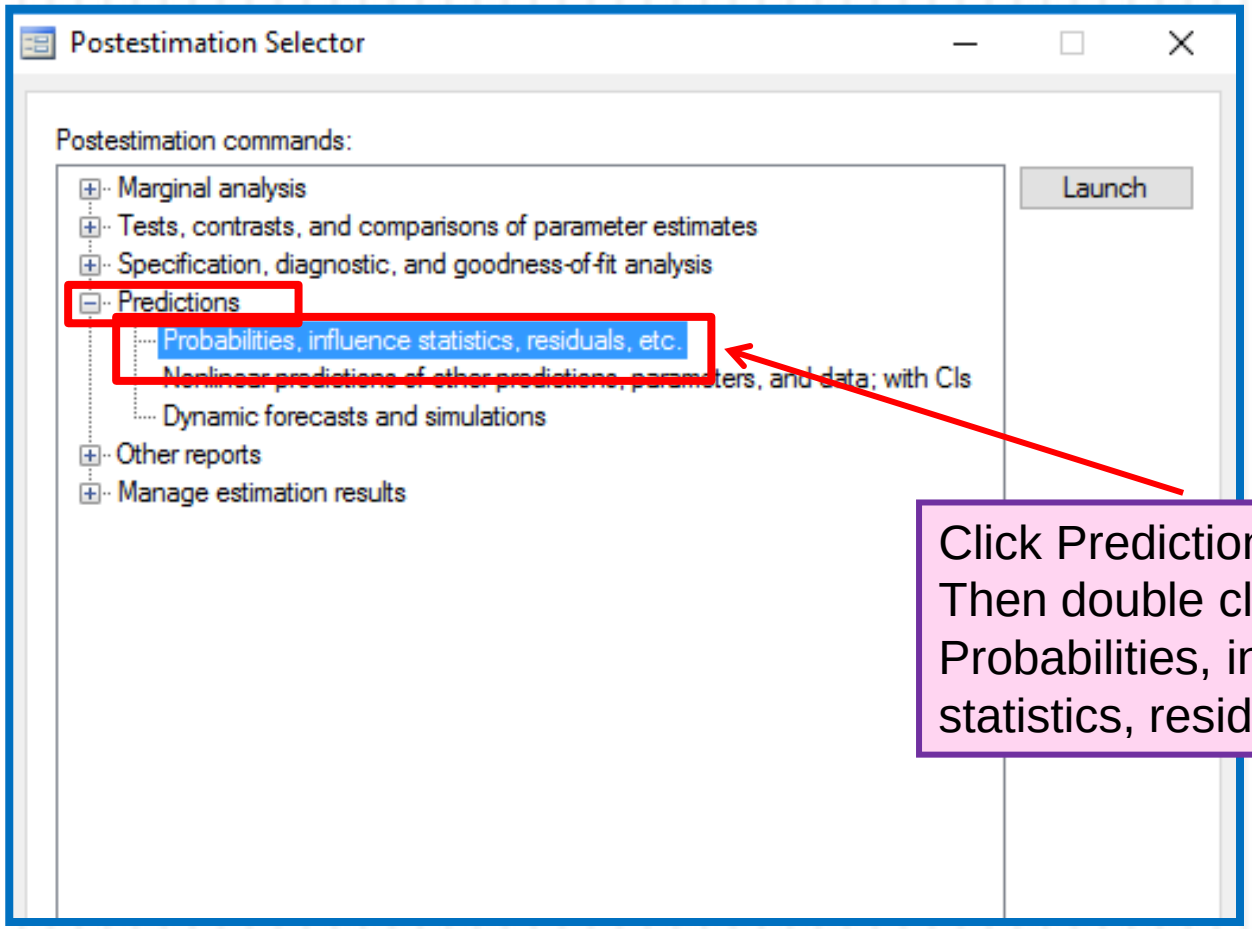


- Area under the ROC curve is 0.638
- It is significantly different from 0.5
- The model can accurately discriminate 63.8% of the cases

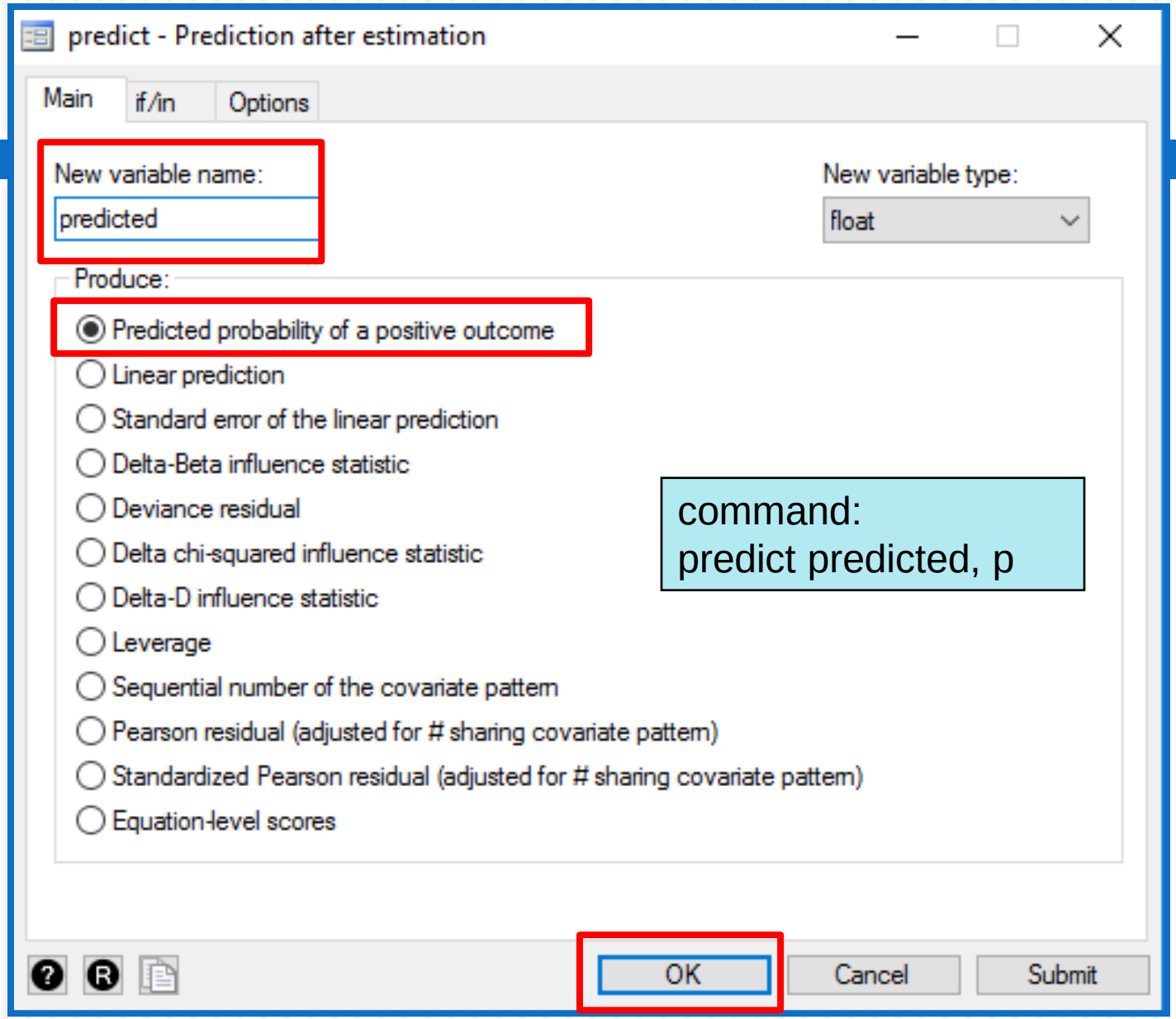
***Repeat for second and third binary logit model**

2nd method: Using roctab command
Advantage: 95% CI is provided

Create predicted value

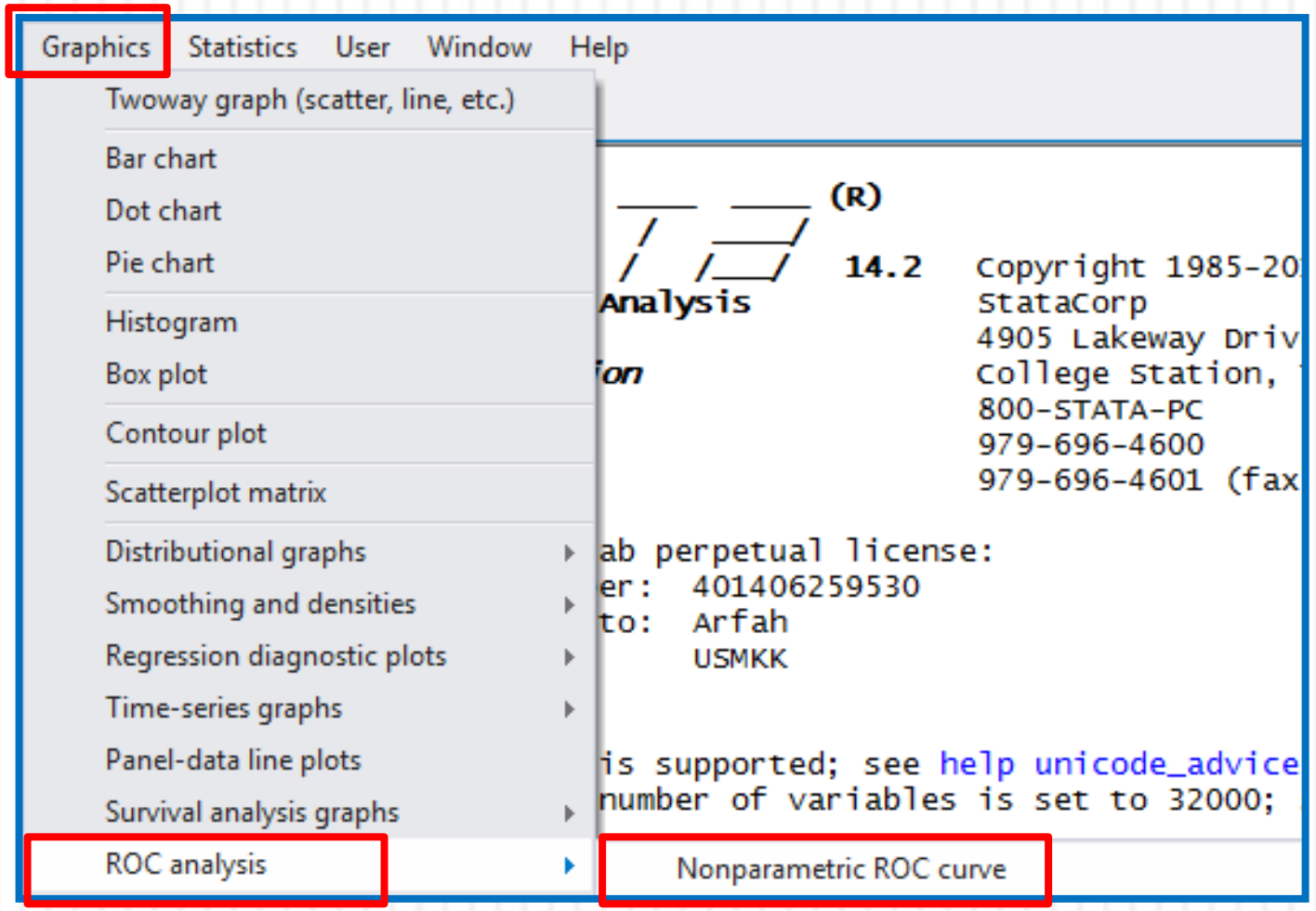


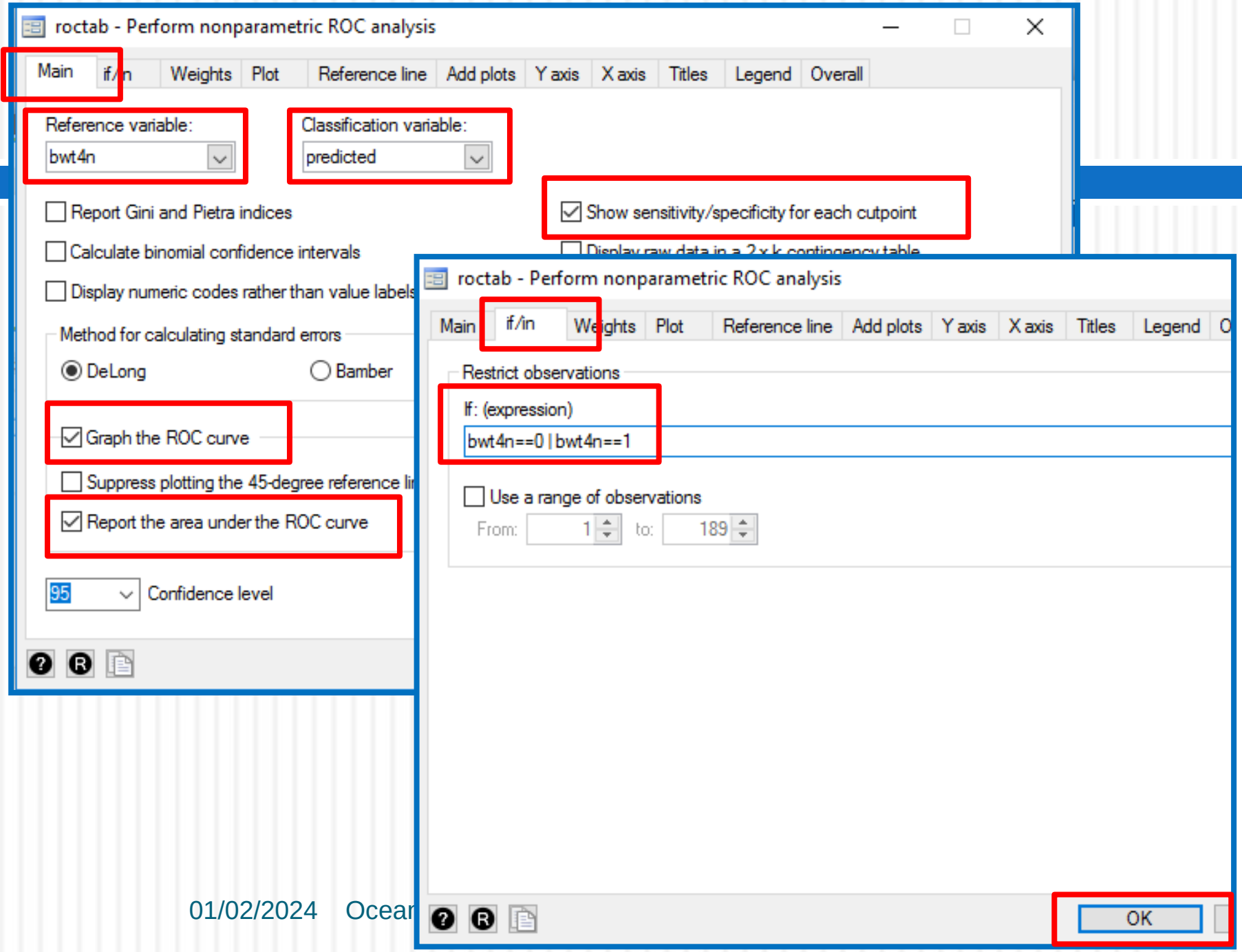
Click Predictions.
Then double click Probabilities, influence statistics, residuals, etc.


predict - Prediction after estimation
Main
if/in
Options
New variable name:
predicted
New variable type:
float
Produce:
Predicted probability of a positive outcome
Linear prediction
Standard error of the linear prediction
Delta-Beta influence statistic
Deviance residual
Delta chi-squared influence statistic
Delta-D influence statistic
Leverage
Sequential number of the covariate pattern
Pearson residual (adjusted for # sharing covariate pattern)
Standardized Pearson residual (adjusted for # sharing covariate pattern)
Equation-level scores
command:
predict predicted, p
OK
Cancel
Submit

Create ROC curve

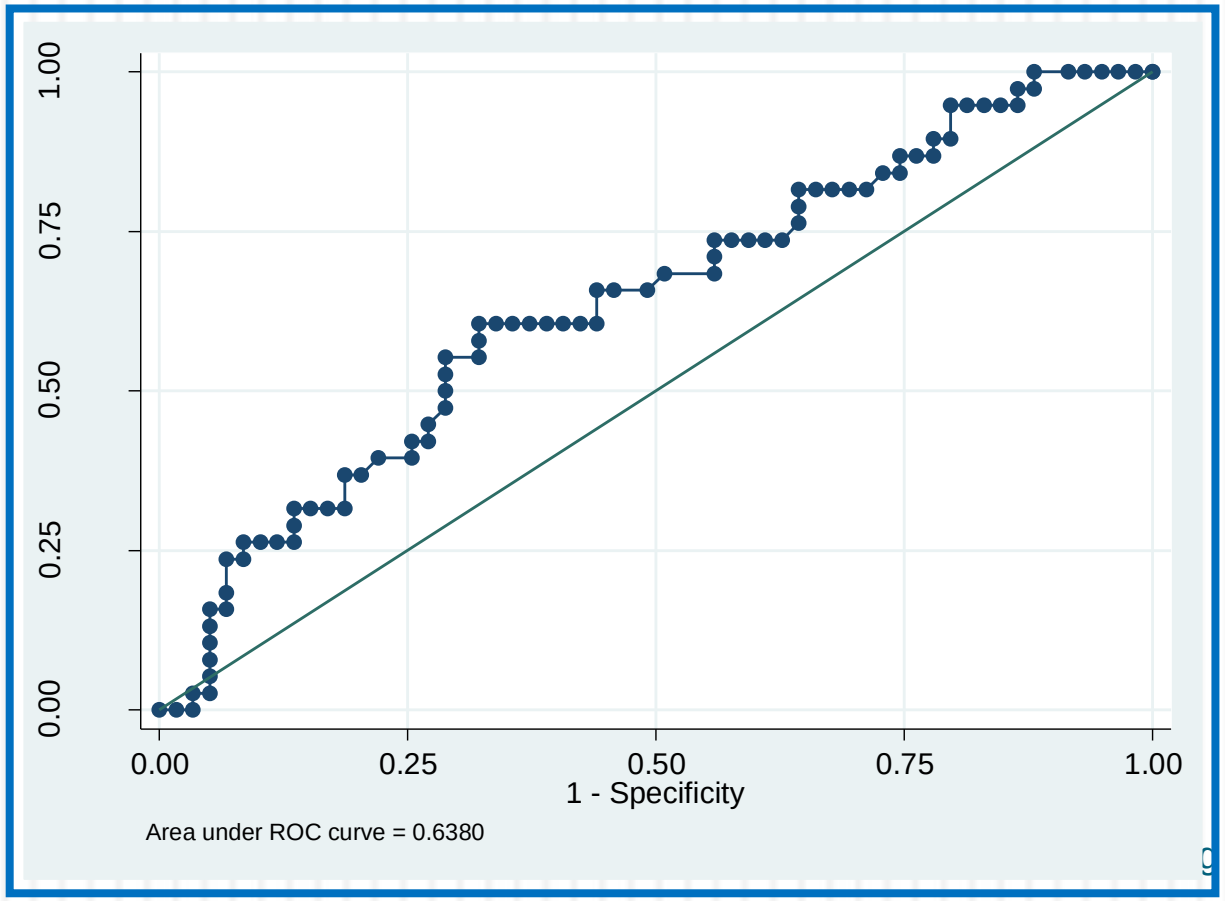
Graphics > ROC analysis > Nonparametric ROC curve




roctab - Perform nonparametric ROC analysis
Main
if/in
Weights
Plot
Reference line
Add plots
Y axis
X axis
Titles
Legend
Overall
Reference variable:
bwt4n
Classification variable:
predicted
Report Gini and Pietra indices
Calculate binomial confidence intervals
Display numeric codes rather than value labels
Show sensitivity/specificity for each cutpoint
Method for calculating standard errors
DeLong
Bamber
Graph the ROC curve
Suppress plotting the 45-degree reference li
Report the area under the ROC curve
95
Confidence level
roctab - Perform nonparametric ROC analysis
Main
if/in
Weights
Plot
Reference line
Add plots
Y axis
X axis
Titles
Legend
Restrict observations
If: (expression)
bwt4n==0 | bwt4n==1
Use a range of observations
From:
1
to:
189
OK

```
roctab bwt4n predicted if bwt4n==0 | bwt4n==1, detail graph summary
```

Obs	ROC Area	Std. Err.	Asymptotic Normal [95% Conf.	Interval]
97	0.6380	0.0579	0.52459	0.75151



***Repeat for second and third binary logit model**

- As a conclusion:

	1st binary logit model	2nd binary logit model	3rd binary logit model
Hosmer-Lemeshow test	***P***=0.892	***P***=0.686	***P***=0.186
Pearson chi-square test	***P***=0.229	***P***=0.144	***P***=0.216
Classification table: Overall correctly classified percentage	63.9%	64.8%	71.4%
Area under the ROC curve	0.64 (95% CI: 0.52, 0.75)	0.69 (95% CI: 0.58, 0.79)	0.85 95% CI: 0.78, 0.93

Step 9: Regression diagnostic (outliers & influential)

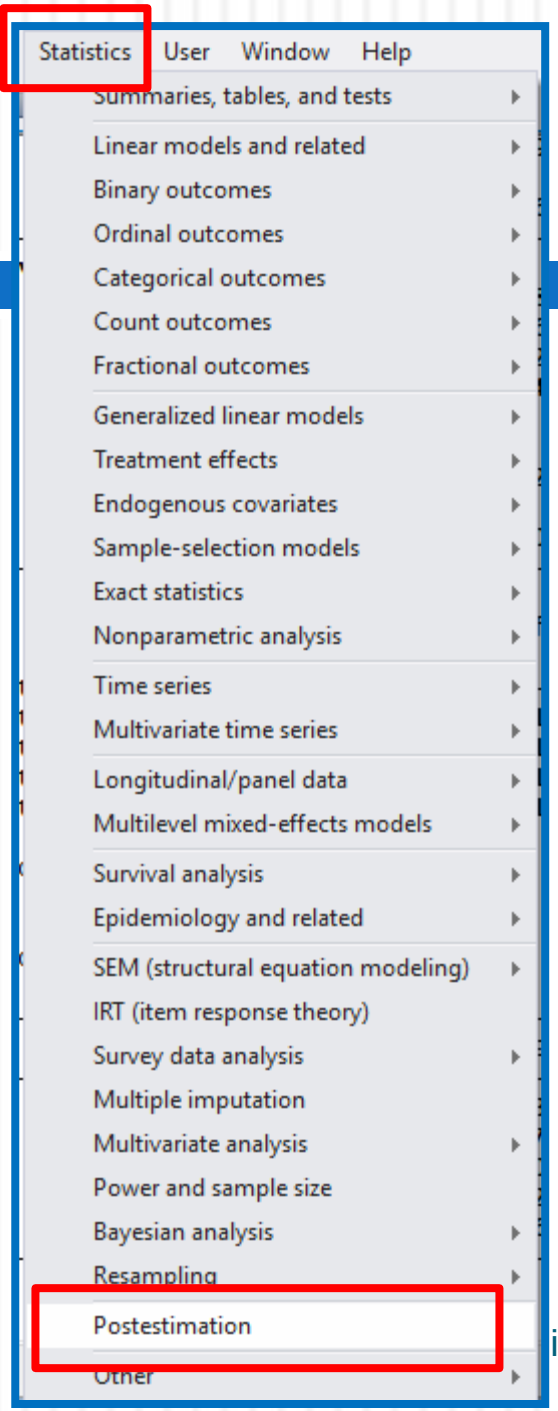
- Before concluding that the model “fits”, need to see if fit is supported over the entire set of covariate pattern
  - Estimated logistic probability (p)
  - Leverage (h)
  - Covariate pattern (n)
  - Hosmer and Lemeshow Delta chi-squared influence statistic (dx2)
  - Hosmer and Lemeshow Delta-D influence statistic (dd)
  - Pregibon Delta-Beta influence statistic (db)

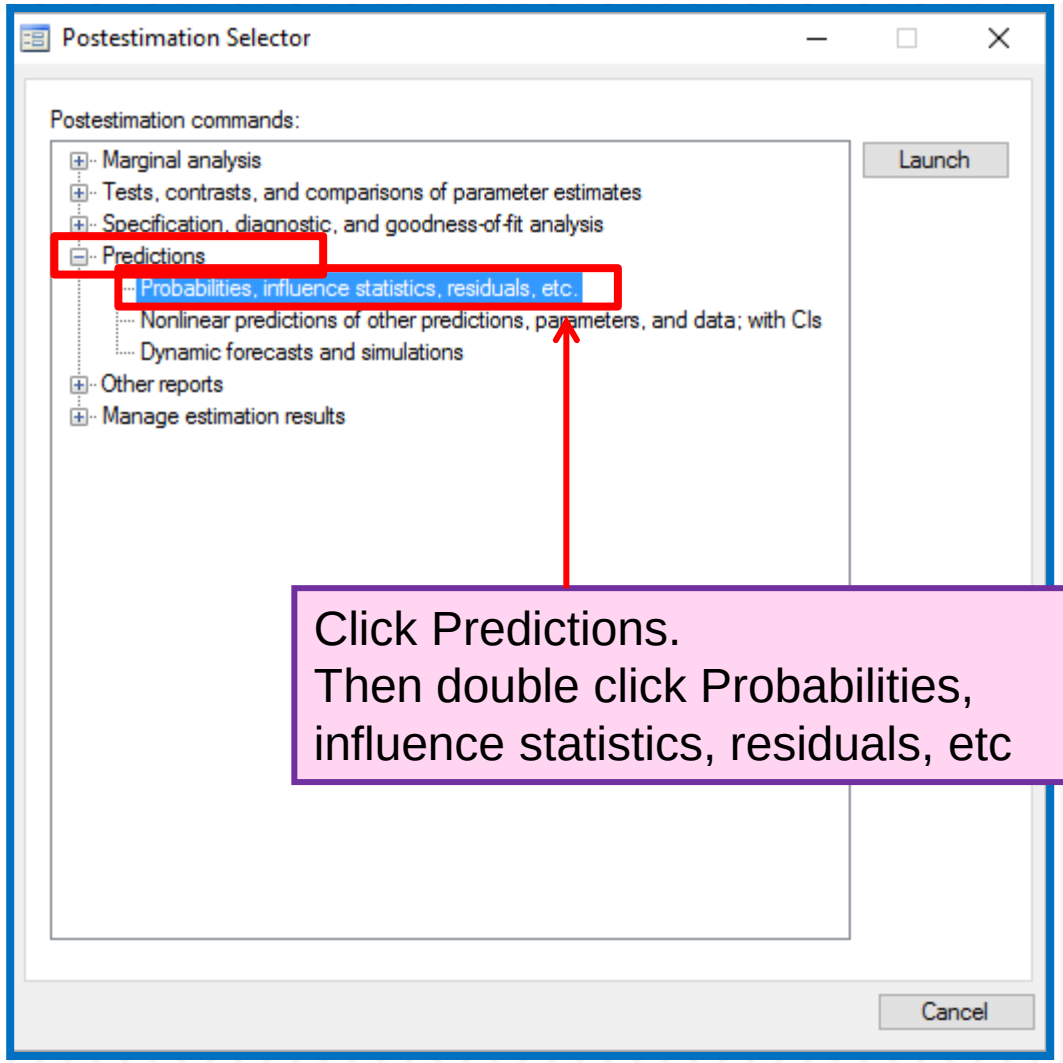
- Need to perform three separate full assessments of fit based on the binary logistic regressions of bwt4n=k vs. bwt4n=0

command: logit *[dependent] [independent]* if *[condition is fulfilled]*

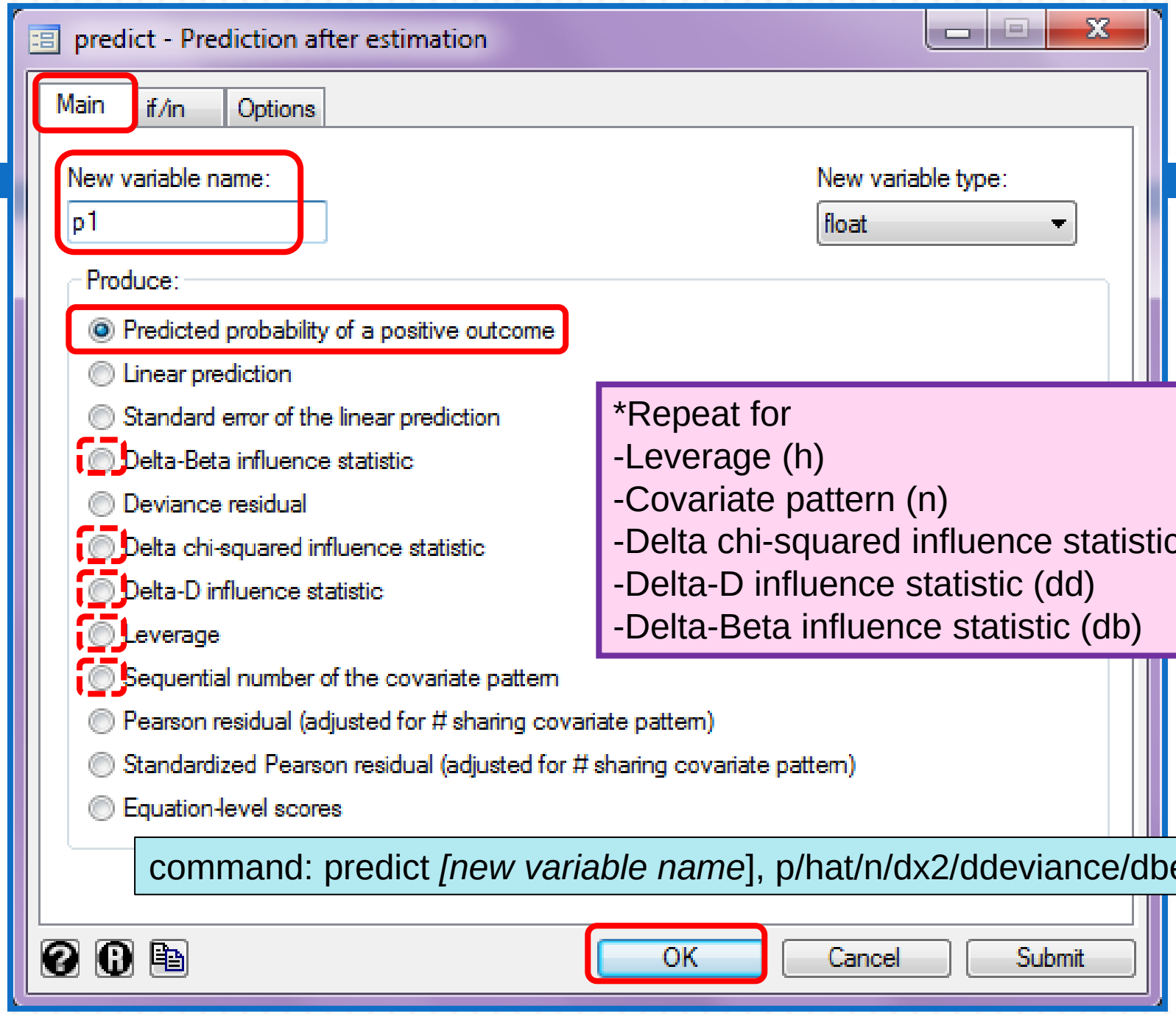
```
. logit bwt4n lwt _Irace_2 _Irace_3 smoke ui if bwt4n==0 | bwt4n==1

Iteration 0:   log likelihood = -64.943982
Iteration 1:   log likelihood = -61.615535
Iteration 2:   log likelihood = -61.600463
Iteration 3:   log likelihood = -61.600463

Logistic regression                             Number of obs     =         97
                                                LR chi2(5)        =       6.69
                                                Prob > chi2       =     0.2450
Log likelihood = -61.600463                     Pseudo R2         =     0.0515

------------------------------------------------------------------------------
       bwt4n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |  -.0012643   .0088442    -0.14   0.886    -.0185986    .0160701
    _Irace_2 |  -.7014273   .6362532    -1.10   0.270    -1.948461     .545606
    _Irace_3 |  -1.420067   .6400296    -2.22   0.027    -2.674502    -.165632
       smoke |  -1.127938   .5674118    -1.99   0.047    -2.240045    -.0158314
          ui |  -.2906456   .5406617    -0.54   0.591    -1.350323    .7690319
       _cons |   .9395066    1.29274     0.73   0.467    -1.594217    3.47323
------------------------------------------------------------------------------
```

Statistics > Postestimation





Click Predictions.
Then double click Probabilities, influence statistics, residuals, etc



*Repeat for
-Leverage (h)
-Covariate pattern (n)
-Delta chi-squared influence statistic (dx2)
-Delta-D influence statistic (dd)
-Delta-Beta influence statistic (db)

command: predict *[new variable name]*, p/hat/n/dx2/ddeviance/dbeta

- First binary logit model

p1	h1	n1	dx1	dd1	db1
.1139657	.06227	41	.1371659	.2580693	.0091085
.618811	.1099959	55	1.824003	2.167316	.2254292
.2448484	.2302428	79	.8424408	1.459353	.2519833
.3513046	.1112136	21	1.827961	2.921683	.2287323

- Second binary logit model

p2	h2	n2	dx2	dd2	db2
.2062928	.0760365	39	.2812995	.5001078	.0231492
.3913957	.1066747	49	.7198988	1.111772	.0859653
.4464501	.2170862	80	2.060309	3.021551	.5712821
.4564553	.0955082	18	2.78535	4.044106	.2941141

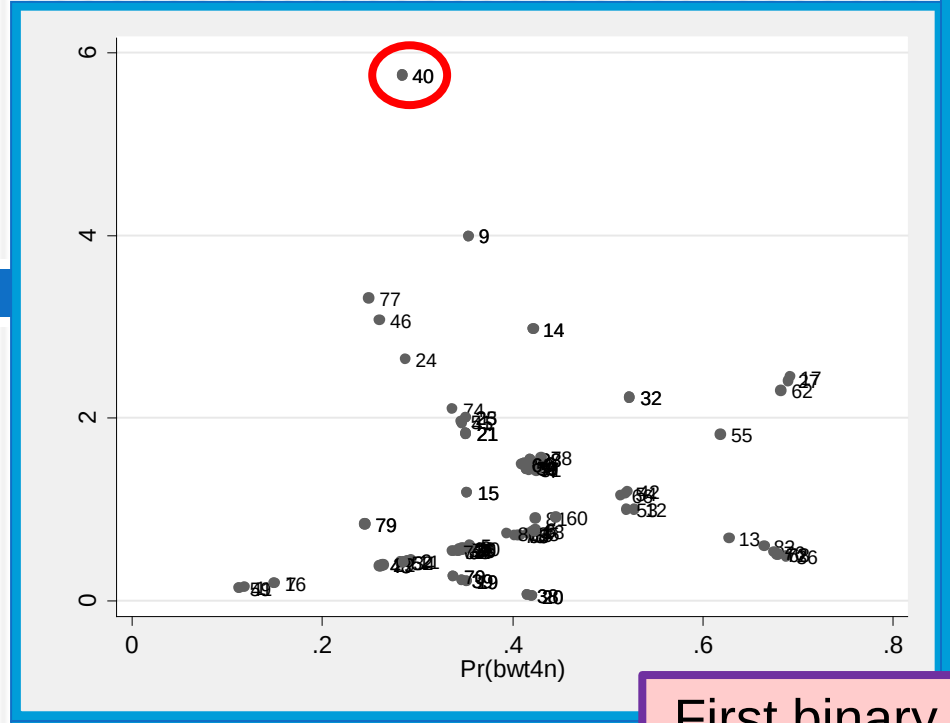
- Third binary logit model

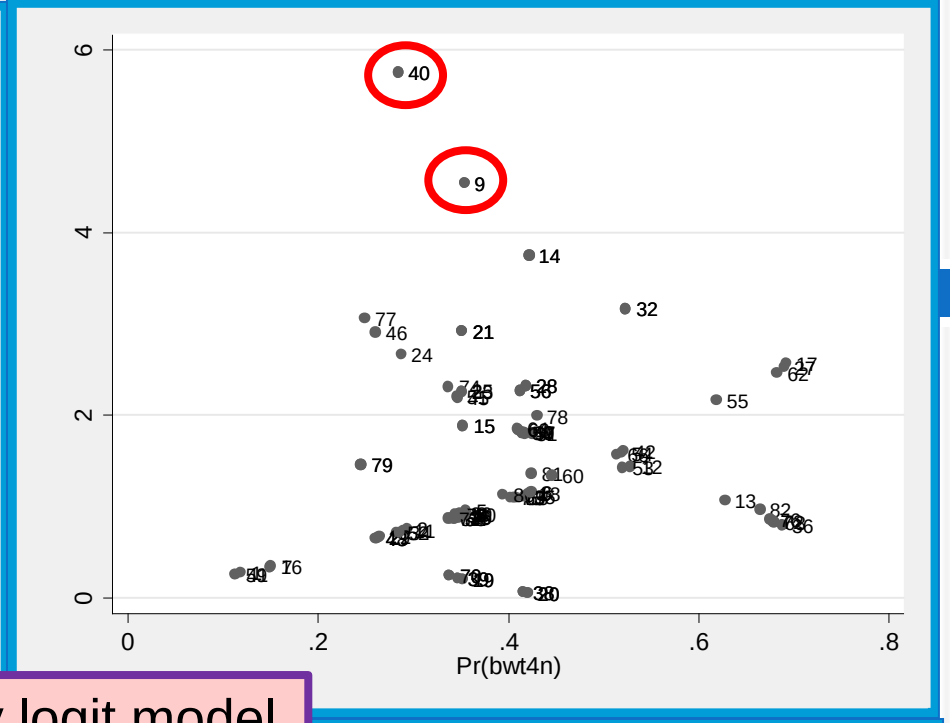
p3	h3	n3	dx3	dd3	db3
.0139223	.0187736	37	.014389	.0285767	.0002753
.5479807	.2160631	49	1.546419	2.025751	.4262129
.0758886	.1622093	81	.1960408	.376813	.0379565
.3148411	.1572768	18	.092511	.096497	.0172653

- Graphical presentation for each binary logit model
  - Scatter plot of dx2 vs p
  - Scatter plot of dd vs p
  - Scatter plot of db vs p
  - Scatter plot of h vs p
  - Scatter plot of dx2 vs p weighted with db (This plot allows us to ascertain the contributions of residual and leverage to db. The larger circle corresponds to the larger value of dx2)

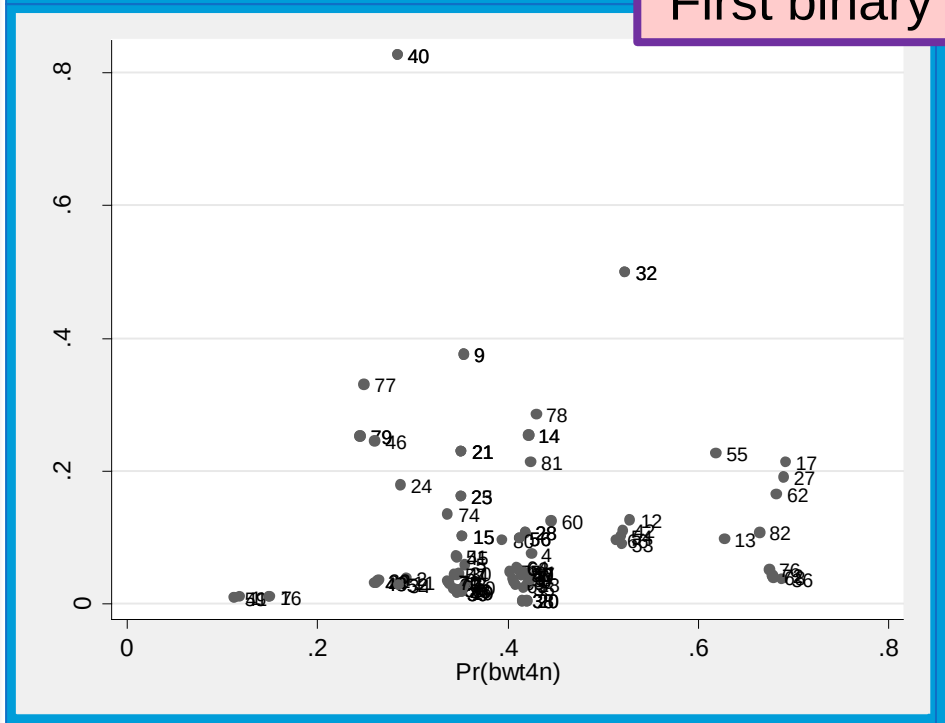
command: scatter *[y-axis] [x-axis],* mlabel(*covariate pattern*)
scatter *[dx2] [p]* [w=*db*], msymbol(oh)

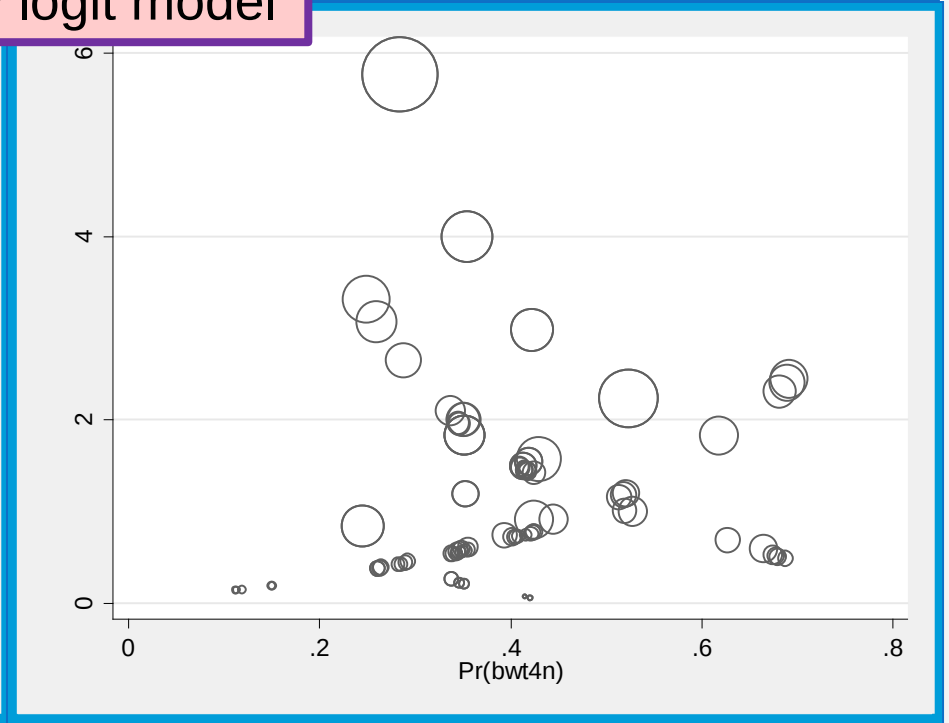
- How to determine outliers?
  - From the scatter plot, observe the cases with different covariate pattern that largely deviates from other cases
  - Delta Beta influential statistics (db) > 1
  - Delta D influential statistics (dd) > 4
  - Delta Chi-squared influential statistics (dx2) > 4
  - Leverage > 0.5

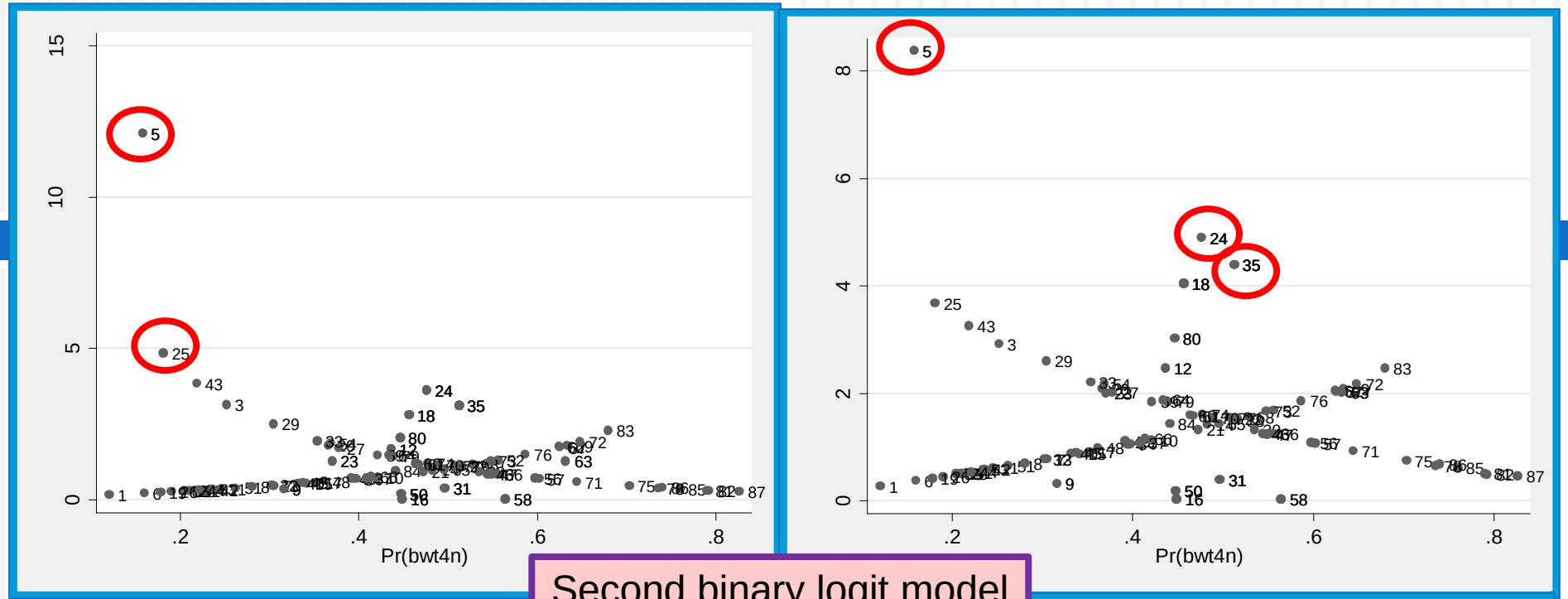




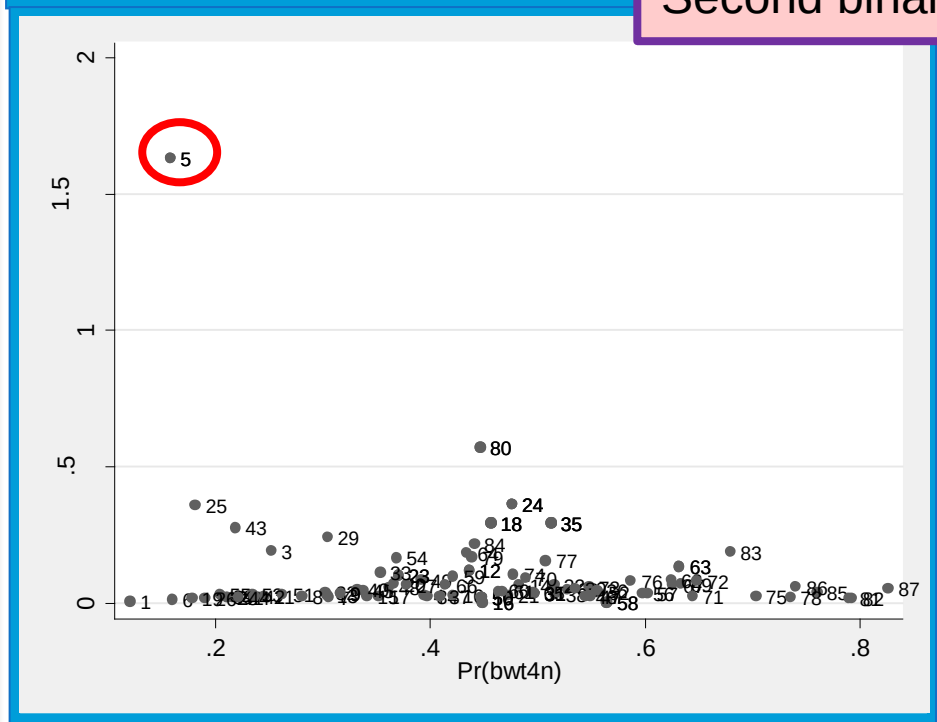
First binary logit model

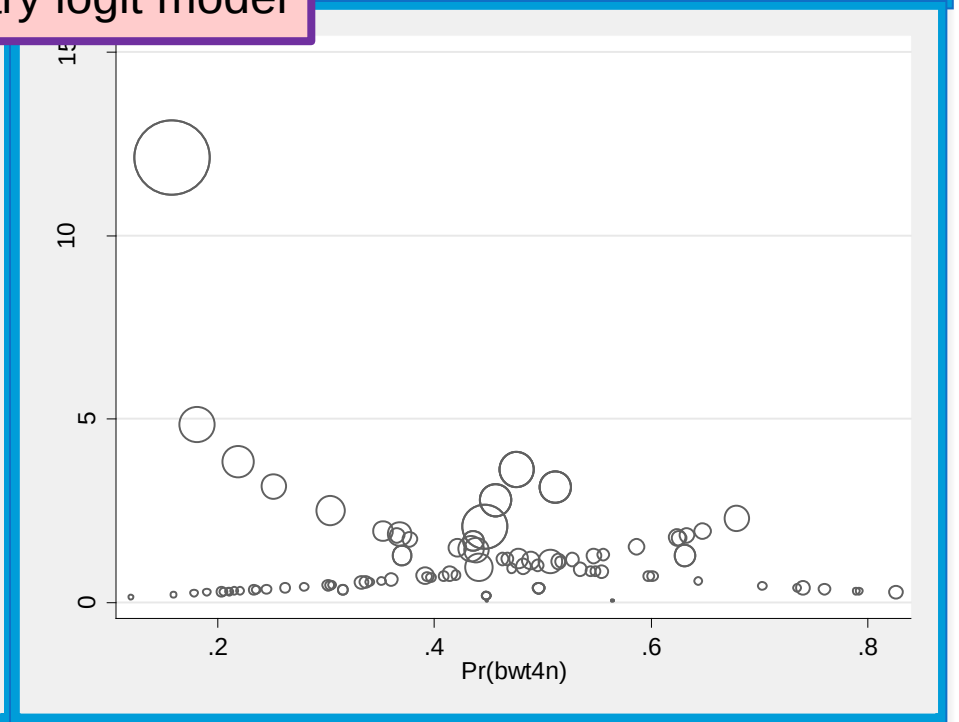




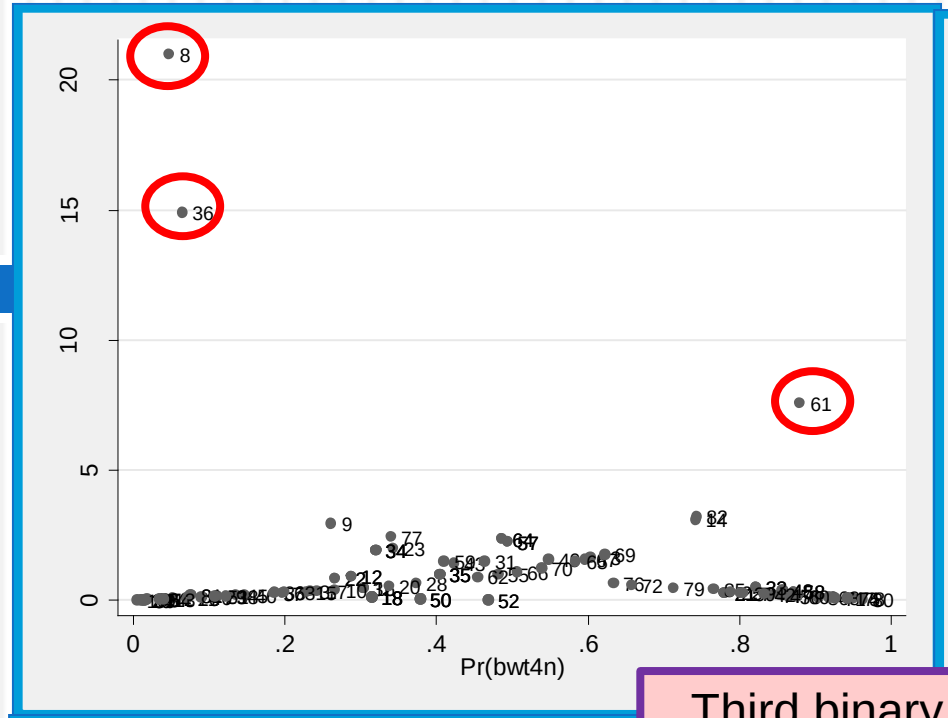


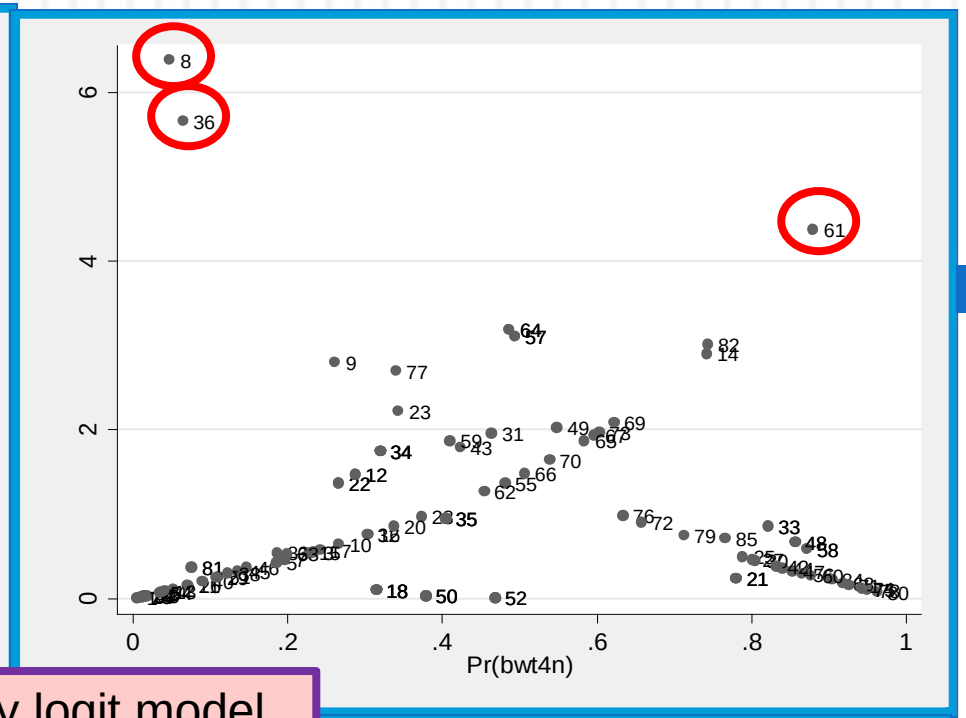
Second binary logit model

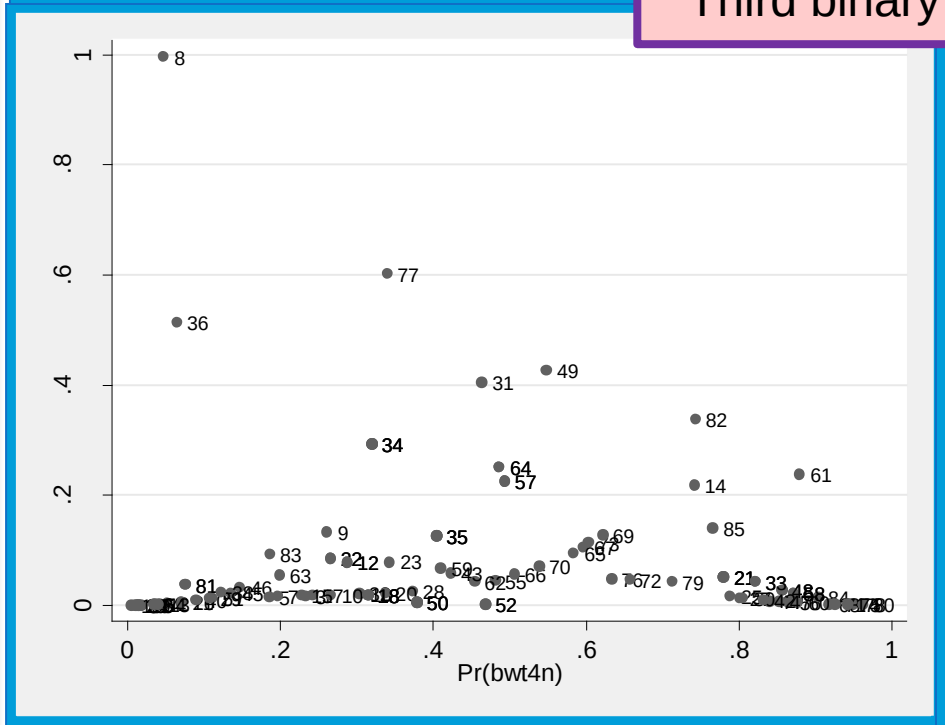


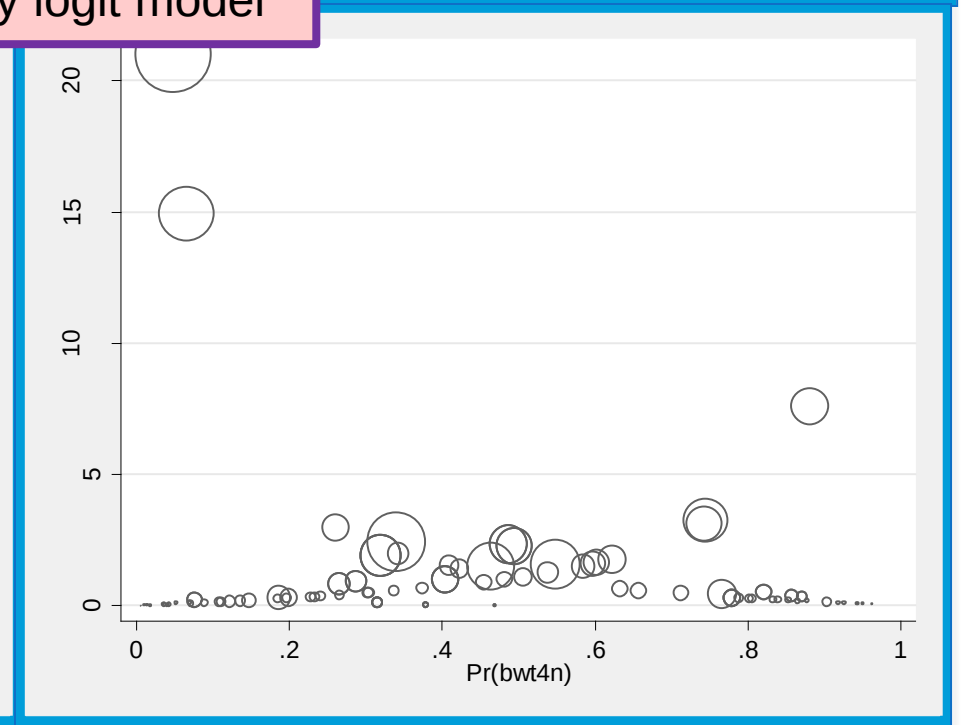


Third binary logit model









Checking influential statistics & leverage

command: list *[prediction]* if *[condition is fulfilled]*

First binary logit model

```
. list n1 h1 dx1 dd1 db1 if h1>0.5 | dx1>4 | dd1>4 | db1>1
```

	n1	h1	dx1	dd1	db1
70.	9	.0860416	3.989973	4.54252	.375623
72.	9	.0860416	3.989973	4.54252	.375623
78.	40	.1255865	5.756149	5.752438	.8267196
84.	40	.1255865	5.756149	5.752438	.8267196

- For first binary logit model, only dx1>4 and dd1>4 for the cases with covariate pattern of 9 and 40

Second binary logit model

```
. list n2 h2 dx2 dd2 db2 if h2>0.5 | dx2>4 | dd2>4 | db2>1
```

	n2	h2	dx2	dd2	db2
4.	18	.0955082	2.78535	4.044106	.2941141
32.	18	.0955082	2.78535	4.044106	.2941141
52.	18	.0955082	2.78535	4.044106	.2941141
104.	5	.1186738	12.10283	8.377453	1.629692
105.	5	.1186738	12.10283	8.377453	1.629692
115.	24	.0905434	3.625397	4.891823	.3609361
116.	25	.068779	4.850256	3.667814	.358235
119.	35	.0860869	3.122882	4.388952	.2941627
120.	35	.0860869	3.122882	4.388952	.2941627
121.	35	.0860869	3.122882	4.388952	.2941627
122.	24	.0905434	3.625397	4.891823	.3609361
142.	24	.0905434	3.625397	4.891823	.3609361

- For second binary logit model, dx2>4, dd2>4 and db2>1 for the cases with covariate pattern of 5, 18, 24, 25 and 35

Third binary logit model

```
. list n3 h3 dx3 dd3 db3 if h3>0.5 | dx3>4 | dd3>4 | db3>1
```

	n3	h3	dx3	dd3	db3
25.	61	.030424	7.569792	4.375112	.2375301
145.	36	.0333031	14.90084	5.657753	.5133401
153.	8	.0452701	20.99183	6.381906	.9953616

- For third binary logit model, dx3>4 and dd3>4 for the cases with covariate pattern of 8, 36 and 61

Step 10: Remedial Measures

- Percent changes in regression coefficient ≥ 20%

$$\frac{|\beta(\text{without outlier} - \beta(\text{with outlier})|}{\beta(\text{with outlier})} \times 100$$

command: logit *[dependent] [independent]* if *[condition is fulfilled]*

command: logit *[dependent] [independent]* if *[condition is fulfilled]* & n2!=25

Percent changes in regression coefficient

For covariate pattern 25 in second binary logit model

```
. logit bwt4n lwt _Irace_2 _Irace_3 smoke ui if bwt4n==0 | bwt4n==2, nolog

Logistic regression                             Number of obs     =        105
                                                LR chi2(5)        =      11.14
                                                Prob > chi2       =     0.0487
Log likelihood = -66.404182                     Pseudo R2         =     0.0774

------------------------------------------------------------------------------
       bwt4n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |   .0160355   .0074761     2.14   0.032     .0013825    .0306885
    _Irace_2 |  -.9187345   .6458722    -1.42   0.155    -2.184621    .3471518
    _Irace_3 |  -.1545177   .5027579    -0.31   0.759    -1.139905    .8308698
       smoke |  -.5910963   .4623199    -1.28   0.201    -1.497227     .315034
          ui |   -.822233   .5835252    -1.41   0.159    -1.965921    .3214554
       _cons |   -1.70383   1.087146    -1.57   0.117    -3.834596    .4269371
------------------------------------------------------------------------------
```

With outlier

```
. logit bwt4n lwt _Irace_2 _Irace_3 smoke ui if bwt4n==0 | bwt4n==2 & n2!=25, nolog

Logistic regression                             Number of obs     =        104
                                                LR chi2(5)        =      13.24
                                                Prob > chi2       =     0.0212
Log likelihood = -64.522427                     Pseudo R2         =     0.0930

------------------------------------------------------------------------------
       bwt4n |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         lwt |   .0163081   .0075372     2.16   0.030     .0015355    .0310808
    _Irace_2 |   -.963132   .6523186    -1.48   0.140    -2.241653     .315389
    _Irace_3 |  -.3041265   .5167651    -0.59   0.556    -1.316967    .7087145
       smoke |    -.74075   .4770581    -1.55   0.120    -1.675767    .1942668
          ui |  -1.086352   .6290782    -1.73   0.084    -2.319322    .1466192
       _cons |  -1.605179    1.09237    -1.47   0.142    -3.746186    .5358273
------------------------------------------------------------------------------
```

Without outlier

Table 3: Percent changes in regression coefficient for each variable in for covariate pattern 25

Variable	Percent changes (%)
lwt	1.88
race1	4.83
race2	96.83
smoke	25.33
ui	32.12

- From the table, percent changes in variable race, smoke and ui were more than 20%
- It is high possibility that the observations with covariate pattern 25 are influential outlier
- Decision to remove the cases depends on the researcher

Step 11: Interpretation, conclusion and presentation

- Run final regression model

command: ologit *[dependent] [independent]* if *[condition is fulfilled]*, nolog or

```
. xi:ologit bwt4n lwt i.race smoke ui, nolog or
i.race            _Irace_1-3          (naturally coded; _Irace_1 omitted)

Ordered logistic regression                     Number of obs   =        189
                                                LR chi2(5)      =      39.66
                                                Prob > chi2     =     0.0000
Log likelihood = -239.82339                     Pseudo R2       =     0.0764
```

bwt4n	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
lwt	1.011999	.0046509	2.60	0.009	1.002925	1.021156
_Irace_2	.2196947	.0916313	-3.63	0.000	.0970055	.4975569
_Irace_3	.3923495	.1288091	-2.85	0.004	.2061683	.7466625
smoke	.3361977	.1035037	-3.54	0.000	.1838819	.6146818
ui	.3810796	.1489093	-2.47	0.014	.1771753	.8196494
/cut1	-.4963771	.666113			-1.801935	.8091804
/cut2	.4722462	.6689457			-.8388633	1.783356
/cut3	1.742281	.6832494			.403137	3.081425

Coef.
.0119279
-1.515517
-.9356021
-1.090056
-.9647471

Table 4: Associated factors of heavier versus a lighter birth weight babies

Variables	Regression coefficient (b)	Adjusted Odds Ratio (95% CI)	p-value
lwt	0.01	1.01 (1.003, 1.021)	0.009
race			
white	0	1	
black	-1.52	0.22 (0.097, 0.498)	<0.001
others	-0.94	0.39 (0.206, 0.747)	0.004
smoke			
no	0	1	
yes	-1.09	0.34 (0.184, 0.615)	<0.001
ui			
no	0	1	
yes	-0.96	0.38 (0.177, 0.820)	0.014

Footnote:

- Backward stepwise logistic regression was applied
- Linearity of continuous variable was checked and reported to be linear
- Multicollinearity and interactions were unlikely
- Assumptions of similarity between proportional model and unconstrained baseline logit, proportional odds assumption and parallel regression assumption were checked and found to be fulfilled
- Overall fit of the model was checked and reported to be Hosmer-Lemeshow test (first model: p=0.892; second model: p=0.686; third model: p=0.186), Pearson chi-square test (first model: p=0.229; second model: p=0.144; third model: p=0.216), correctly classified percentage (first model: 63.9%; second model: 64.8%; third model: 71.4%), Area under the ROC curve (first model: 0.64 (95% CI: 0.52, 0.75); second model: 0.69 (95% CI: 0.58, 0.79) ; third model: 0.85 (95% CI: 0.78, 0.93))
- Regression diagnostic was performed by estimated logistic probability (p), Leverage (h), covariate pattern (n), Hosmer and Lemeshow Delta chi-squared influence statistic (dx2), Hosmer and Lemeshow Delta-D influence statistic (dd) and Pregibon Delta-Beta influence statistic (db)
- Influential outliers were identified by checking percent changes in regression coefficient set at 20%

Interpretation & conclusion

- The estimate of the OR for a heavier versus lighter weight baby with black mothers compared to white mothers was adjusted OR=0.22 (95% CI 0.097, 0.498), ***P***<0.001. The odds of a heavier versus lighter weight baby was 78% less for black mothers compared to white mothers
- The estimate of the OR for a heavier versus lighter weight baby with mothers of other races compared to white mothers was adjusted OR=0.39 (95% CI 0.206, 0.747), ***P***=0.004. The odds of a heavier versus lighter weight baby was 61% less for mothers of other races compared to white mothers

Interpretation and conclusion

- The estimate of the OR for a heavier versus lighter weight baby with mother who smoked compared to those who did not smoke was adjusted OR=0.34 (95% CI 0.184, 0.615), ***P***<0.001. The odds of a heavier versus lighter weight baby was 66% less for smoking mothers compared to non-smoking mothers
- The estimate of the OR for a heavier versus lighter weight baby with mothers with uterine irritability compared to mothers without uterine irritability was adjusted OR=0.38 (95% CI 0.177, 0.820), ***P***=0.014. The odds of a heavier versus lighter weight baby was 62% less for mothers with uterine irritability compared to mothers without uterine irritability

Interpretation and conclusion

- 95% confidence interval of adjusted OR=1.01 does not include one (1.003, 1.021), variable weight of mother is significant to the model (***P***=0.009). There is a 1% increase in the odds of a heavier baby per one pound increase in weight of mother

THANK YOU