BIOSTATISTICS II

Statistical Modeling for Biomedical Researchers

William D. Dupont, Ph.D.

Volume I, Sections 1-4

M.P.H. Program
Vanderbilt University School of Medicine
March 2011

I. REVIEW OF BIOSTATISTICS 1 AND SIMPLE LINEAR REGRESSION

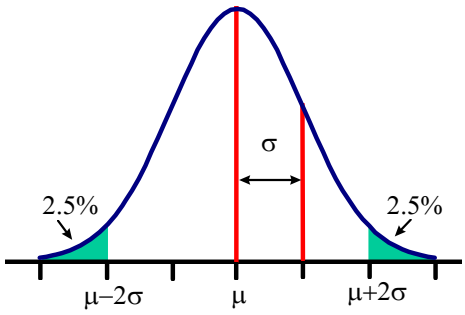
- Distinction between a parameter and a statistic
- The normal distribution
- Inference from a known sample about an unknown target population
- Simple linear regression: Assessing simple relationships between two continuous variables
- Interpreting the output from a linear regression program. Analyzing data with Stata
- Plotting linear regression lines with confidence bands
- Making inferences from simple linear regression models
- Lowess regression and residual plots. How do you know you have the right model?
- Transforming data to improve model fit
- Comparing slopes from two independent linear regressions

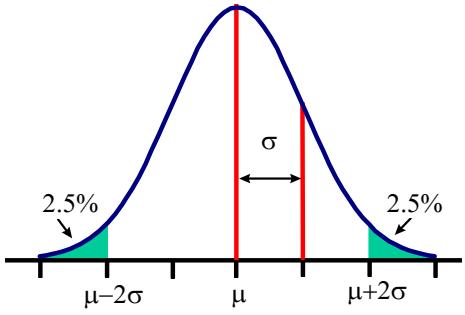
I. REVIEW OF BIOSTATISTICS 1 AND SIMPLE LINEAR REGRESSION

1. Parameters, Statistics, and Statistical Notation

a) Normal distribution

This distribution is defined by two **parameters:**





b) **Mean** μ

c) **Variance** σ^2

= mean squared distance of a statistic from its mean.

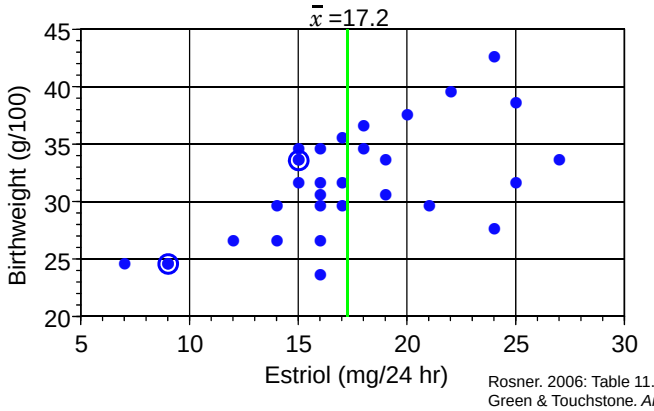
d) **Standard deviation** $\sigma = \sqrt{\sigma^2}$

e) **Sample mean and central tendency**

Given a sample $x_1, x_2, \ldots, x_n = \{x_i : i = 1,2,\ldots,n\}$, the

sample mean $\bar{x} = \sum x_i / n$ is a measure of **central tendency.**

$\bar{x} \to \mu$ as $n \to \infty$ if all members of the population have an equal chance of being sampled.



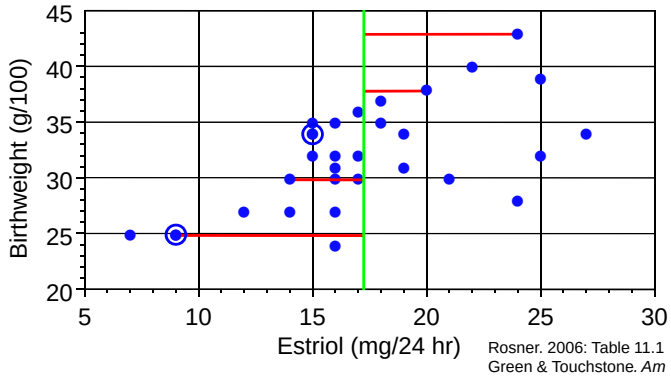
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

f) Residuals and dispersion

The i^{th} **residual** $= x_i - \bar{x}$

The average residual length = $\sum |x_i - \bar{x}| / n$ is an intuitive measure of **dispersion:** the extent to which observations vary from each other.

It is rarely used because it is difficult to work with mathematically.



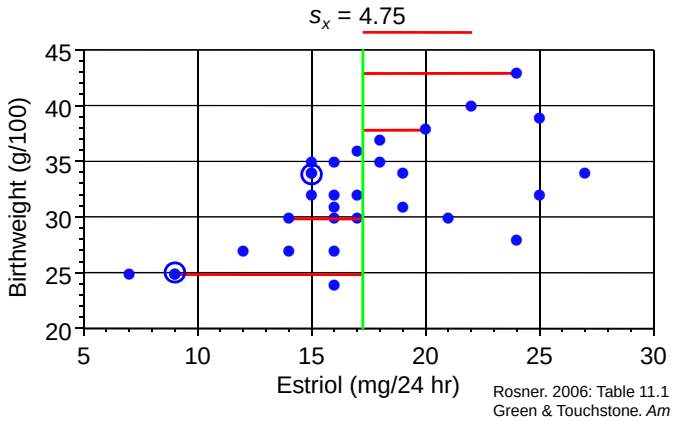
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

g) Sample variance

$$s^2 = \sum (x_i - \bar{x})^2 / (n-1) \qquad \{1.1\}$$

h) Sample standard deviation

$s = \sqrt{s^2}$ is the most common measure of dispersion.



Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

i) Expected value

The **expected value** of a statistic is its average value from a very large number of experiments.

The expected value of both x_i and $\bar{x}$ is μ.

The expected value of s^2 is σ^2.

We write $E(\bar{x})=\mu$ $E(s^2)=\sigma^2$

j) Unbiased estimate of a parameter

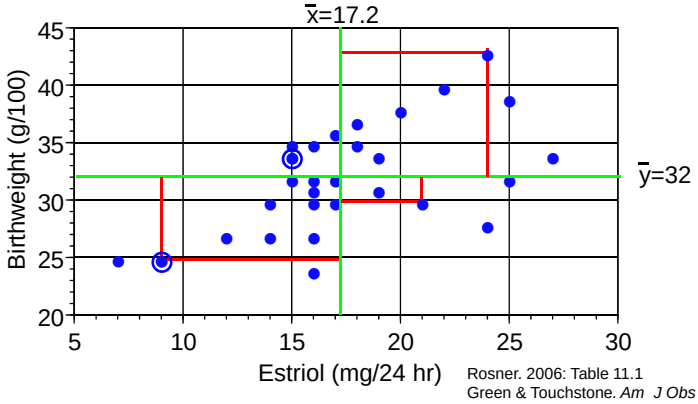
A statistic is **unbiased estimate** of a parameter if its expected value equals the parameter.

$\bar{x}$ is an unbiased estimate of μ since $E(\bar{x})=\mu$

The denominator of $s^2=\sum(x_i-\bar{x})^2/(n-1)$ is n-1 rather than n in order to make s^2 an unbiased estimate of σ^2.

2. Elementary Statistics for Continuous Bivariate Data

a) Covariance



Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

Given paired observations (x,y), the **covariance**

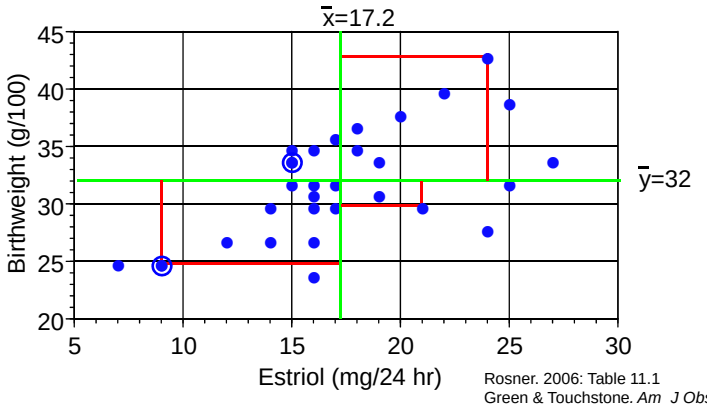
σ_{xy} = expected mean product of paired residuals

b) Sample covariance

Given a sample of size n, σ_{xy} is estimated by the **sample covariance**

$$s_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) / (n-1)$$

$$E(s_{xy}) = \sigma_{xy}$$

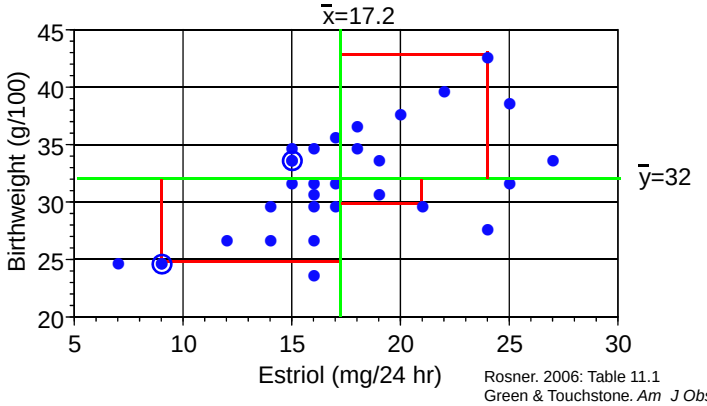


Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

Note: **i)** Product of paired residuals may be positive or negative.

ii) If x and y are independent these terms cancel each out and $\sigma_{xy} = 0$

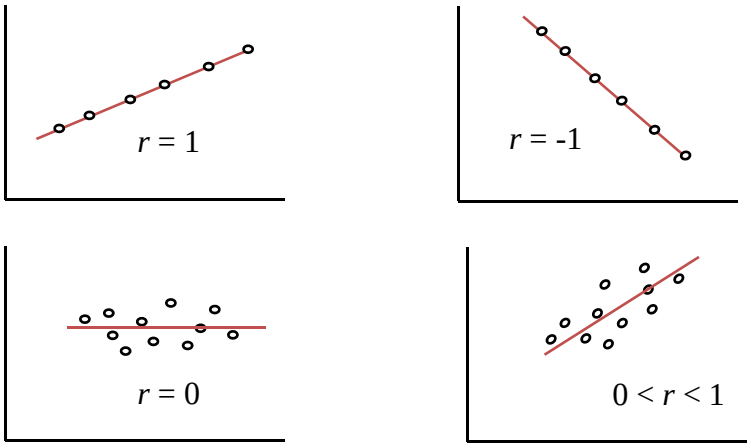
iii) $\sigma_{xy} > 0$ if most pairs fall in the upper right or lower left quadrant of the figure.

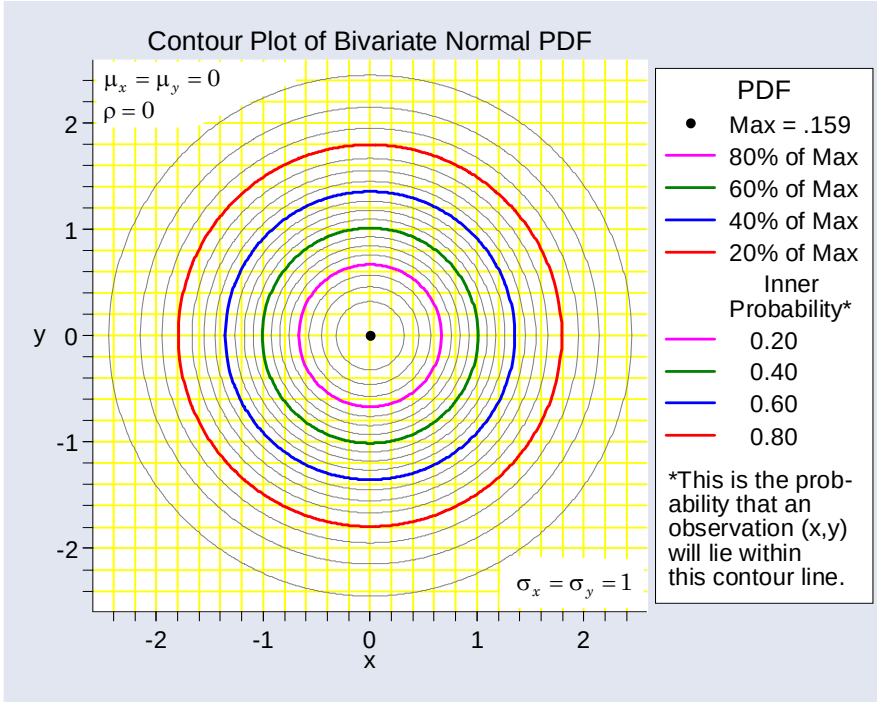


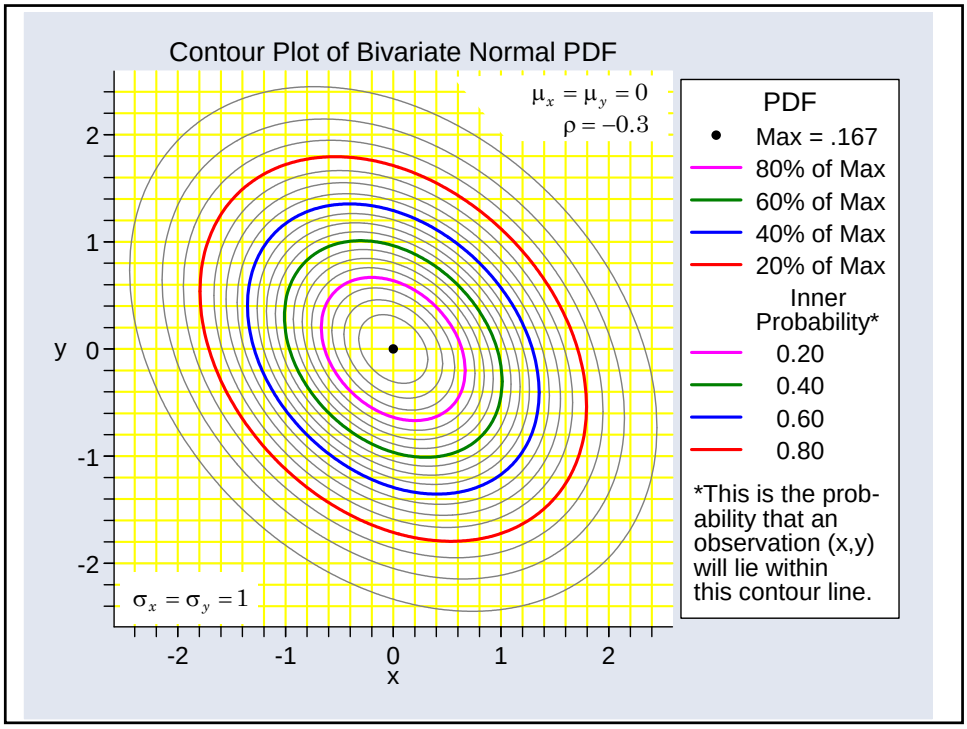
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

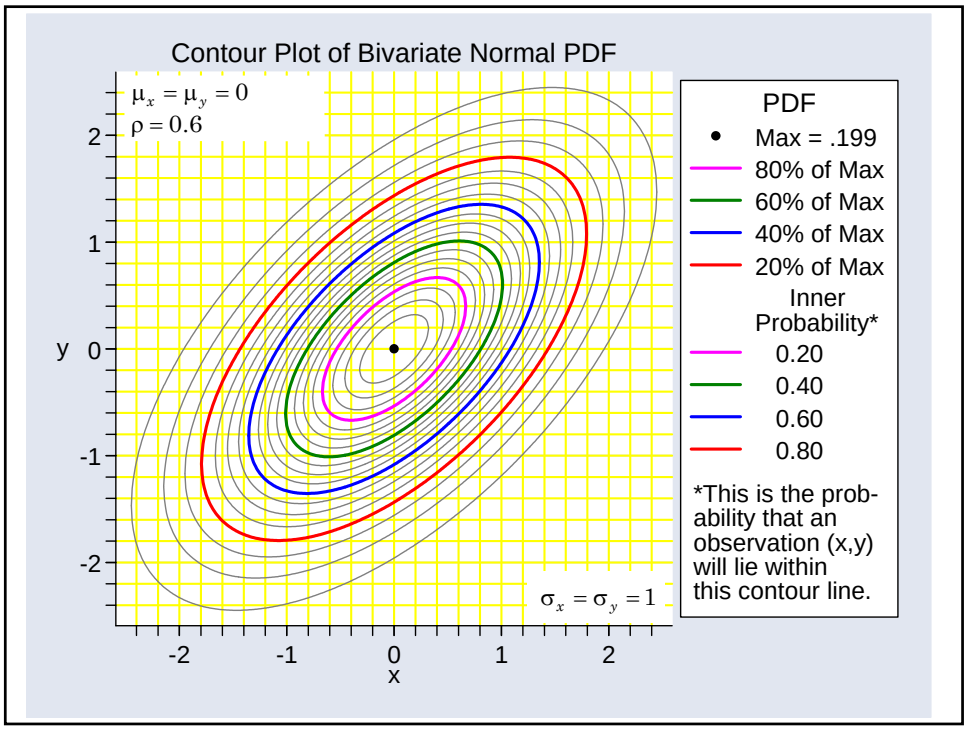
c) Correlation coefficient ρ

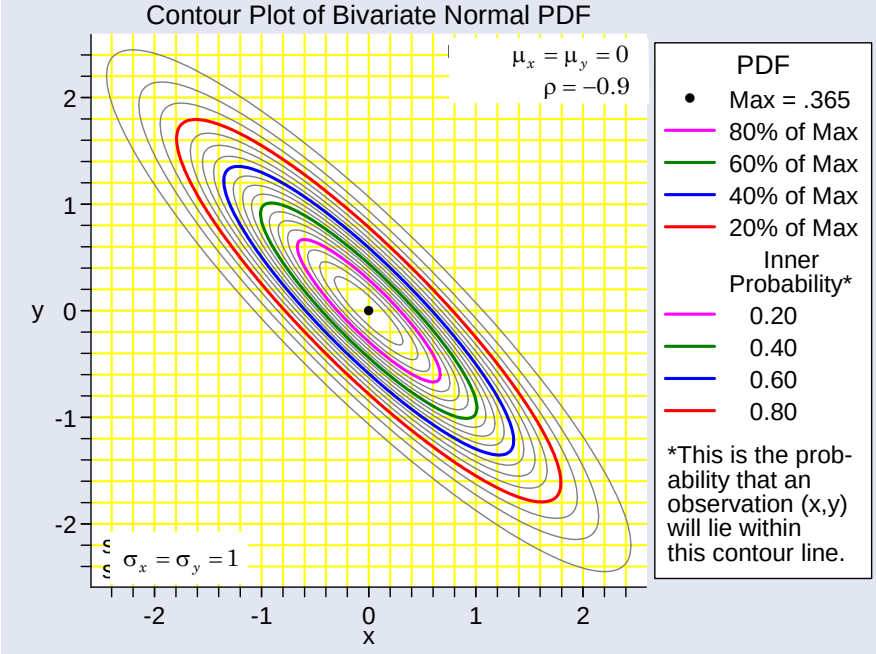
$\rho = \sigma_{xy} / (\sigma_x \sigma_y)$ is estimated by $r = s_{xy} / (s_x s_y)$ **{1.2}**






Contour Plot of Bivariate Normal PDF
$\mu_x = \mu_y = 0$
$\rho = -0.3$
$\sigma_x = \sigma_y = 1$
y
x
PDF
Max = .167
80% of Max
60% of Max
40% of Max
20% of Max
Inner Probability*
0.20
0.40
0.60
0.80
*This is the probability that an observation (x,y) will lie within this contour line.


Contour Plot of Bivariate Normal PDF
$\mu_x = \mu_y = 0$
$\rho = 0.6$
$\sigma_x = \sigma_y = 1$
y
x
PDF
Max = .199
80% of Max
60% of Max
40% of Max
20% of Max
Inner Probability*
0.20
0.40
0.60
0.80
*This is the probability that an observation (x,y) will lie within this contour line.



3. Simple Linear Regression

a) The Model

We assume that

$$y_i = \alpha + \beta x_i + \varepsilon_i$$

Where x_i is a variable observed on the i^{th} patient

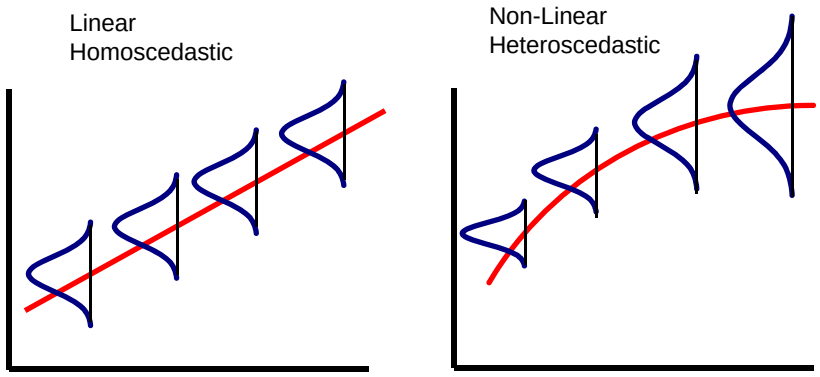
ε_i is assumed to be normally and independently distributed with mean 0 and standard deviation σ

y_i is the response from the i^{th} patient

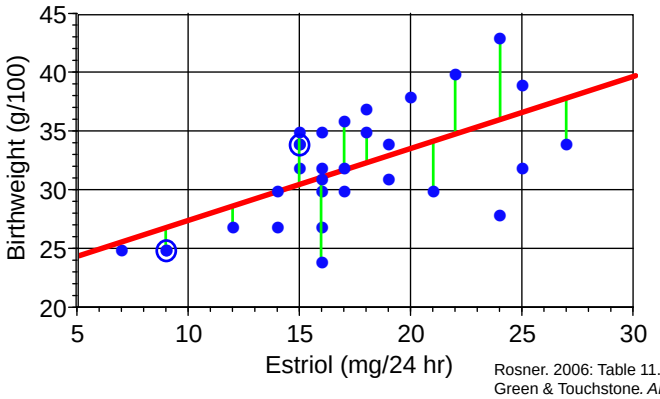
α and β are model parameters.

The expected value of y_i is $E(y_i) = \alpha + \beta x_i$.

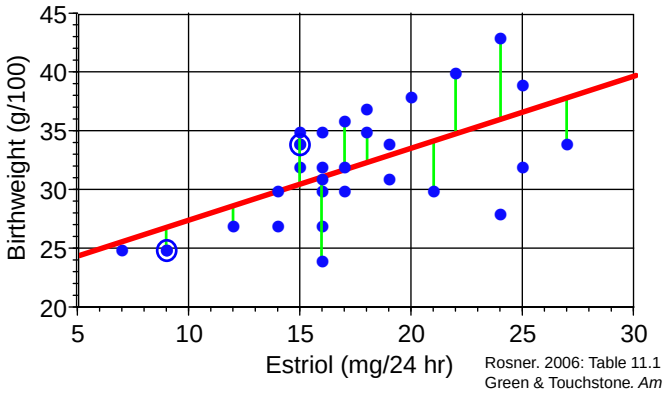
b) Implications of linearity and homoscedasticity.



We estimate α and β by minimizing the sum of the squared residuals



Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*



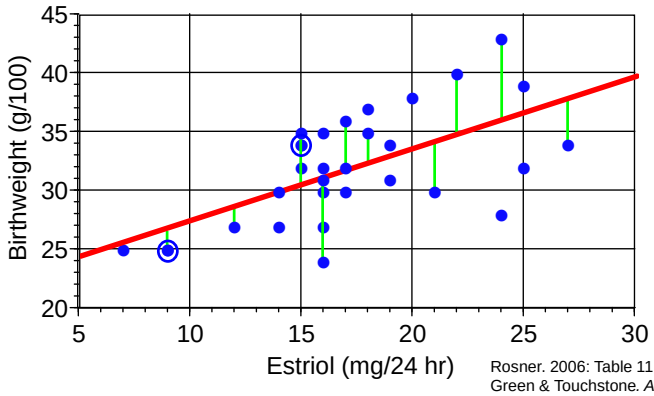
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

c) Slope parameter estimate

β is estimated by $b = r\, s_y / s_x$ **{1.3}**

d) Intercept parameter estimate

α is estimated by $a = \bar{y} - b\bar{x}$ **{1.4}**



Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

e) Least squares estimation

$\hat{y} = a + bx$ is the **least squares estimate** of $\alpha + \beta x$.

Note: **i)** $\hat{y}$ is an **unbiased** estimate of $\alpha + \beta x$.

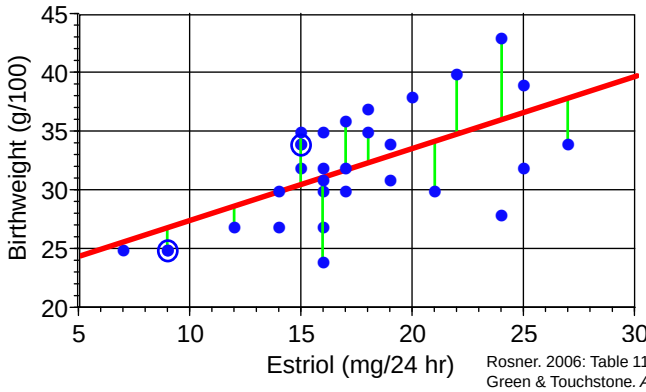
ii) Since $a = \bar{y} - b\bar{x}$

$\hat{y} = a + bx$ can be rewritten $\hat{y} - \bar{y} = b(x - \bar{x})$.

Hence the regression line passes through $(\bar{x}, \bar{y})$

iii) Since $b = r\, s_y / s_x$

$b \rightarrow 0$ as $r \rightarrow 0$ and $b \rightarrow s_y / s_x$ as $r \rightarrow 1$

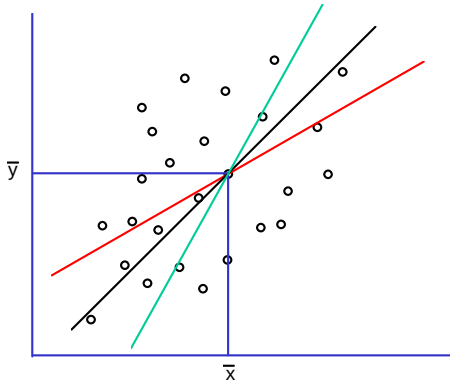


Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

4. Historical Trivia: Origin of the Term *Regression*

If $s_x = s_y$ then $b = r$ and hence $\hat{y} - \bar{y} = r(x - \bar{x}) < (x - \bar{x})$

if $1 > r > 0$ and $x > \bar{x}$



Francis Galton, a 19th century pioneer of statistics who was interested in eugenics studied patterns of inheritance of all sorts of attributes and found that, for example, the sons of tall men tended to be shorter than their fathers. He called this regression toward the mean, which is where the term linear regression comes from.

5. The Stata Statistical Software Package

Stata is an excellent tool for the analysis of medical data. It is far easier to use than other software of similar sophistication. However, we will be using Stata for some complex analyses and you may be puzzled by some of its responses. If so please ask. I would very much like to minimize the time you spend struggling with Stata and maximize the time you spend learning statistics. I will be available to answer questions at most times on days, evenings and weekends.

If you have not used Stata since Biometry I you are probably very rusty. Here are a few reminders and aids that may help.

a) Punctuation

Proper punctuation is mandatory. If Stata gives a confusing error message, the first thing to check is your punctuation. Stata commands are modified by **qualifiers** and **options**. Qualifiers precede options and there must be a comma before the first option.

For example

```
table age if treat==1 ,by(sex)
```

Might produce a table showing the number of men and women of different ages receiving treatment 1.

`if treat==1` is a qualifier and `by(sex)` is an option.

Without the comma, Stata will not recognize `by(sex)` as a valid option to the table command.

Some command prefixes must be followed by a colon.

b) Capitalization

Stata **variables** and commands are **case sensitive**. That is, Stata considers `age` and `Age` to be two distinct variables. In general, I recommend that you always use lower case variables. Sometimes Stata will create variables for you that contain upper case letters. You must use the correct capitalization when referring to these variables.

c) Command summary

At the end of the text book is a summary of most of the commands that are needed in this course. These may be helpful in answering your class exercises.

d) GUI interface

You can avoid learning Stata syntax by using their pull down menus. These menus generate rather complex syntax but feel free to use them if it makes the exercises easier.

This interface is extensively documented in my text. See Section 1.3.8 on page 15.

d) Data files and log files

You may download the Stata data files, log files,do files and these lecture notes that you will need from this course from the web at:
http://biostat.mc.vanderbilt.edu/BiostatIILectureNotes

then click on the desired links for data files, Stata log files, do files or lecture notes. Pages for the class schedule, student names and exercises are password protected. The username and password for these pages is the same as for the other MPH courses except that the second letter of the username must be lower case while the other letters must be upper case. (This is due to a camel-case requirement for usernames on the Biostatistics wiki that I can't get around.)

The class exercises are very similar to the examples discussed in class. You can save yourself time by cutting and pasting commands from these log files into your Stata Command window, or by modifying these do files.

e) Other things you should know

It is important that you can do the following by the end of today.

- Open, close and save Stata data files.
- Review and modify data in the Stata editor.
- Open and close Stata log files.
- Paste commands from the Review window into the Command window.
- Copy and paste Stata graphics into your word processor.
- Copy and paste commands from your external text editor into the command window.

6. Color Coding Conventions for Stata Log Files

a) Introductory Example

```
. * RosnerTable11.1.log                                                    {1}
. *
. * Examine the Stata data set from Table 11.1 of Rosner, p. 554          {2}
. * See Green & Touchstone 1963
. *
. use "C:\MyDocs\MPH\LectureNotes\rostab11.dta", clear                     {3}
. * Data > Describe data > Describe data in memory
. describe
Contains data from \\PMPC158\mph\analyses\linear_reg\stata\rostab11.dta
  obs:            31
 vars:             3                          10 Nov 1998 12:34
 size:           496 (99.9% of memory free)
------------------------------------------------------------------------------
   1. id          float  %9.0g
   2. estriol     float  %9.0g                 Estriol (mg/24 hr)
   3. bweight     float  %9.0g                 Birth Weight (g/100)
------------------------------------------------------------------------------
Sorted by:  estriol
```

{1} You will find this and other Stata log files discussed in this course at: http://biostat.mc.vanderbilt.edu/BiostatisticsTwoClassPage

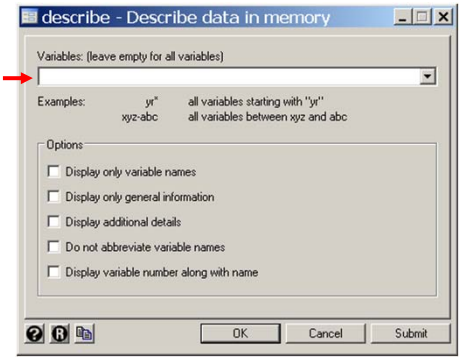
and clicking on Example Logs and Data from Lecture Notes.

Most of the class exercises may be completed by performing Stata sessions that are similar to those discussed in class.

{2} I have adopted the following color coding conventions for Stata log files throughout these notes.

- **Red** is used for Stata comment statements. Also, red numbers in braces in the right margin refer to comments at the end of the program and are not part of the programming code. **Red** text on Stata output is my annotation rather than text printed by Stata.
- Stata command words, qualifiers and options are written in **blue**
- Variables and data set names are written in **black**, as are algebraic or logical formulas
- Stata output is written in **green**

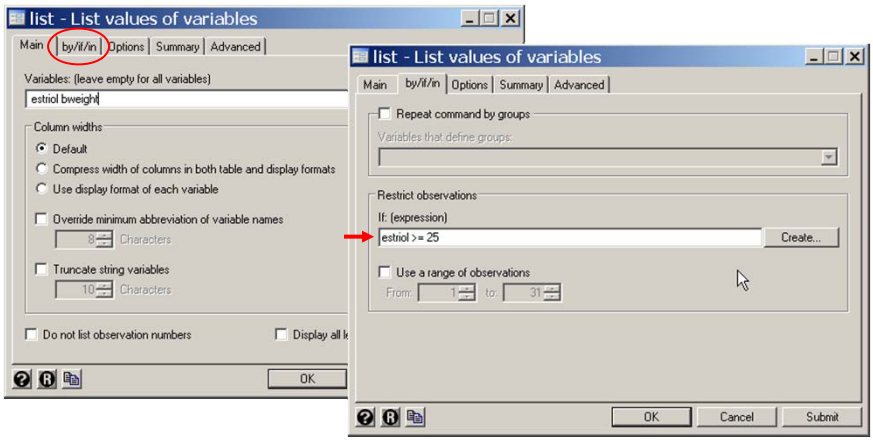
{3} The use command reads the Rosner data into memory. The previous contents of memory are purged.



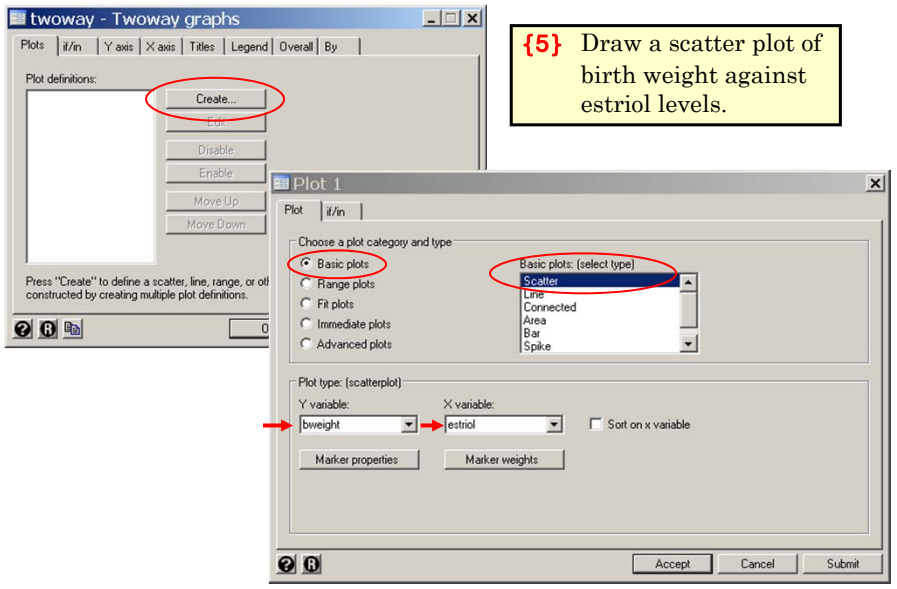
```
. * Data > Describe data > List data
. list estriol bweight if estriol >= 25                              {4}

      estriol    bweight

 29.       25         39
 30.       25         32
 31.       27         34
```

{4} List the variables of estriol and birthweight for those patients whose estriol values are at least 25.

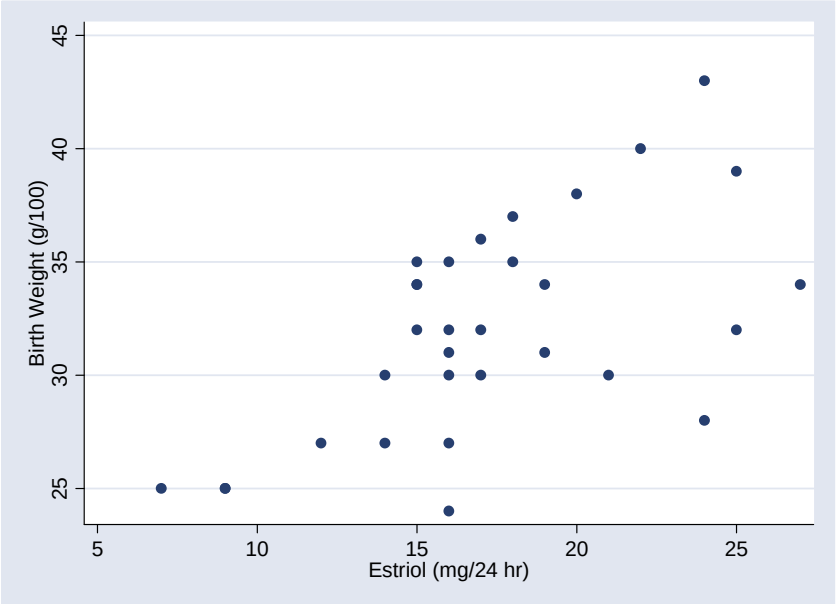
There is no obvious distinction between command modifiers like *if* and command options. In general, modifiers apply to most Stata commands while options are specific to a given class of commands. See you manuals and command summary in the text.



```
. * Graphics > Twoway graph (scatter, line, etc.)
. twoway scatter bweight estriol                                    {5}
```



{5} Draw a scatter plot of birth weight against estriol levels.



7. Linear Regression with Stata

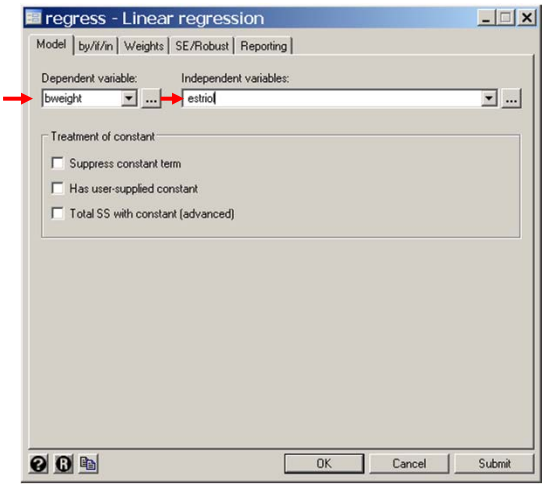
a) Running a simple linear regression program

```
. *  Birth_Weight.LR.log
. *
. *  Linear regression of birth weight on estriol
. *  See Rosner, Table 11.1, p554, and Green & Touchstone 1963
. *
. use C:\MyDocs\MPH\ANALYSES\LINEAR_REG\Stata\rostab11.dta , clear
. * Statistics > Linear models and related > Linear regression
. regress bweight estriol                                          {1}
```

{1} This command analyses the following model.

$$E(\text{bweight}) = \alpha + \text{estriol} * \beta$$

Of particular interest is to estimate the **slope parameter β** and to test that the null hypothesis that $\beta = 0$.



{2} The **Total Sum of Squares** (TSS) = 674 can be shown to equal $\sum(y_i - \bar{y})^2$, the total squared variation in y.

{3} The **Model Sum of Squares** (MSS) = 250.6 equals $\sum(\hat{y}_i - \bar{y})^2$, which is the squared variation explained by the model.

```
  Source |       SS          df          MS            Number of obs =      31
---------+------------------------------------         F(  1,    29) =   17.16
   Model |  250.574476        1     250.574476         Prob > F      =  0.0003  {3}
Residual |  423.425524 n-2 = 29 s² = 14.6008801         R-squared     =  0.3718
---------+------------------------------------         Adj R-squared =  0.3501
   Total |      674.00       30      22.4666667         Root MSE  = s =  3.8211  {2}

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.    t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117   4.14  P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417    8.21    0.000       16.16407    26.88278
------------------------------------------------------------------------------
```

{4} The **Residual** or **Error Sum of Squares** (ESS) = 423.4 = $\sum(y_i - \hat{y}_i)^2$

It can be shown that SST = $\sum(y_i - \bar{y})^2 = \sum(\hat{y}_i - \bar{y})^2 + \sum(y_i - \hat{y}_i)^2$

= SSM + SSE

{5} **R-squared** is the square of the correlation coefficient. It also equals MSS/TSS and hence measures the **proportion of the total variation in bweight** that is explained by the model. When R-squared =1, s^2=0 and the data points fall on the straight line

```
  Source |       SS          df          MS            Number of obs =      31
---------+------------------------------------         F(  1,    29) =   17.16
   Model |  250.574476        1     250.574476         Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s² = 14.6008801 {4}    R-squared     =  0.3718  {5}
---------+------------------------------------         Adj R-squared =  0.3501
   Total |      674.00       30      22.4666667         Root MSE  = s =  3.8211

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.    t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117   4.14  P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417    8.21    0.000       16.16407    26.88278
------------------------------------------------------------------------------
```

b) Interpreting output from a simple linear regression program

i) The estimates a and b of α and β and their standard errors are as shown.

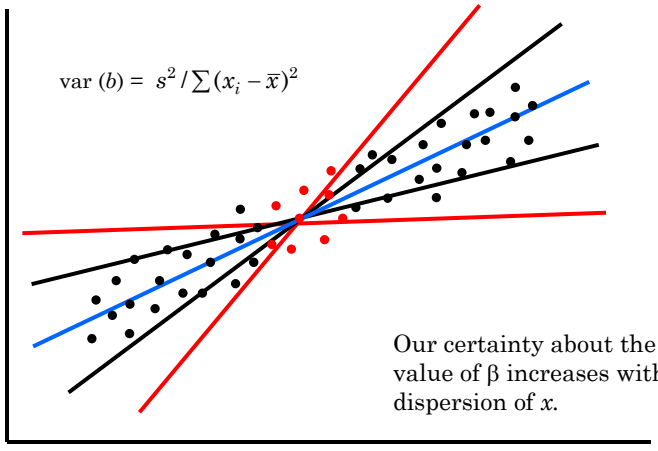
ii) The null hypothesis that β = 0 can be rejected with P< 0.0005

iii) $s^2 = \sum(y_i - (a + bx_i))^2/(n-2)$ {1.5}
estimates σ^2

iv) s^2 is often called the mean sums of squares for error, or MSE.

v) s is often called the Root MSE.

```
  Source |       SS          df          MS            Number of obs =      31
---------+------------------------------------         F(  1,    29) =   17.16
   Model |  250.574476         1     250.574476        Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s²  = 14.6008801       R-squared     =  0.3718
---------+------------------------------------         Adj R-squared =  0.3501
   Total |      674.00        30      22.4666667       Root MSE  = s =  3.8211

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.    t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117    4.14 P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417     8.21   0.000      16.16407    26.88278
------------------------------------------------------------------------------
```



b) Interpreting output from a simple linear regression program

viii) $se(b) = \sqrt{var(b)}$

ix) $b/se(b) = .6081905/ .1468117 = 4.143$ {1.7}

has a t distribution with n-2 degrees of freedom when $\beta = 0$

If n is large then the t distribution converges to a standard normal z distribution.

For large n the these coefficients will be significantly different from 0 if the ratio of the coefficient to its standard error is greater than 2.

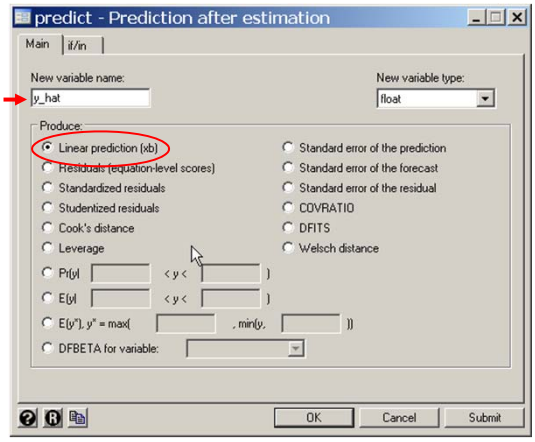
```
  Source |       SS           df          MS              Number of obs =      31
---------+------------------------------------           F(  1,    29) =   17.16
   Model |  250.574476         1     250.574476          Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s²  = 14.6008801         R-squared     =  0.3718
---------+------------------------------------           Adj R-squared =  0.3501
   Total |      674.00        30      22.4666667         Root MSE  = s =  3.8211

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.   t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117   4.14 P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417    8.21   0.000       16.16407    26.88278
------------------------------------------------------------------------------
```

8. Plotting a Linear Regression with Stata

```
. *  Statistics > Postestimation > Predictions, residuals, etc.
. predict y_hat , xb                                                    {1}
```

{1} **predict** is a post estimation command that can estimate a variety of statistics after a regression or other estimation command is run. The ***xb*** option causes a new variable (in this example *y_hat*) to be set equal to each child's **expected birth weight** $\hat{y}(x) = a + bx$, where x is the mother's estriol level and a and b are the parameter estimates of the linear regression.

N.B. Calculations by the predict command always are based on the most recently executed regression command.

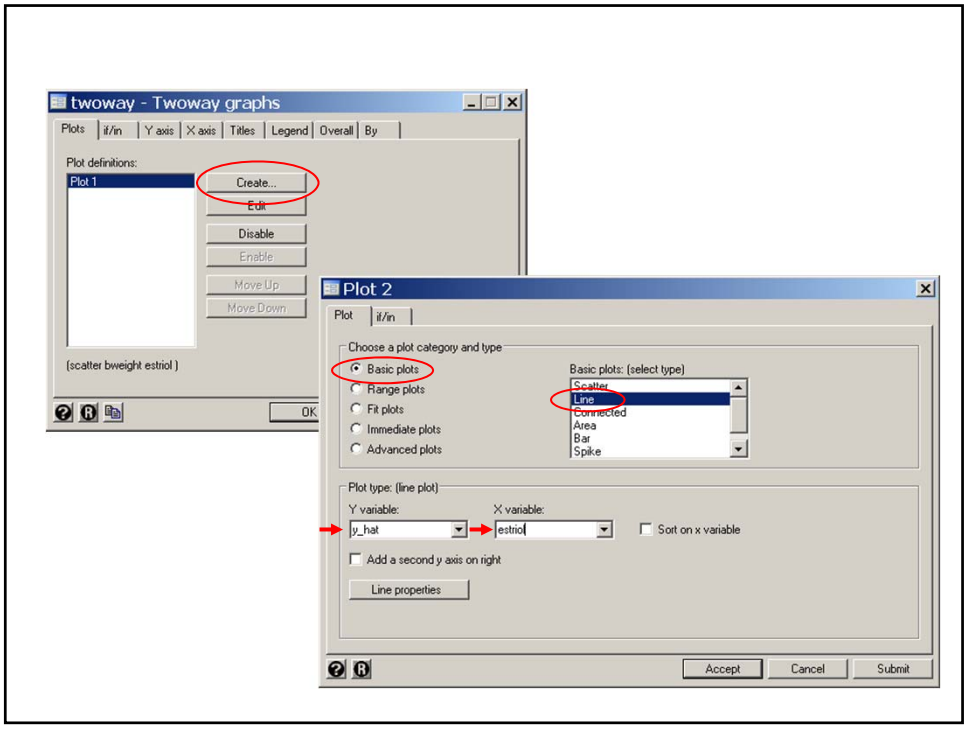


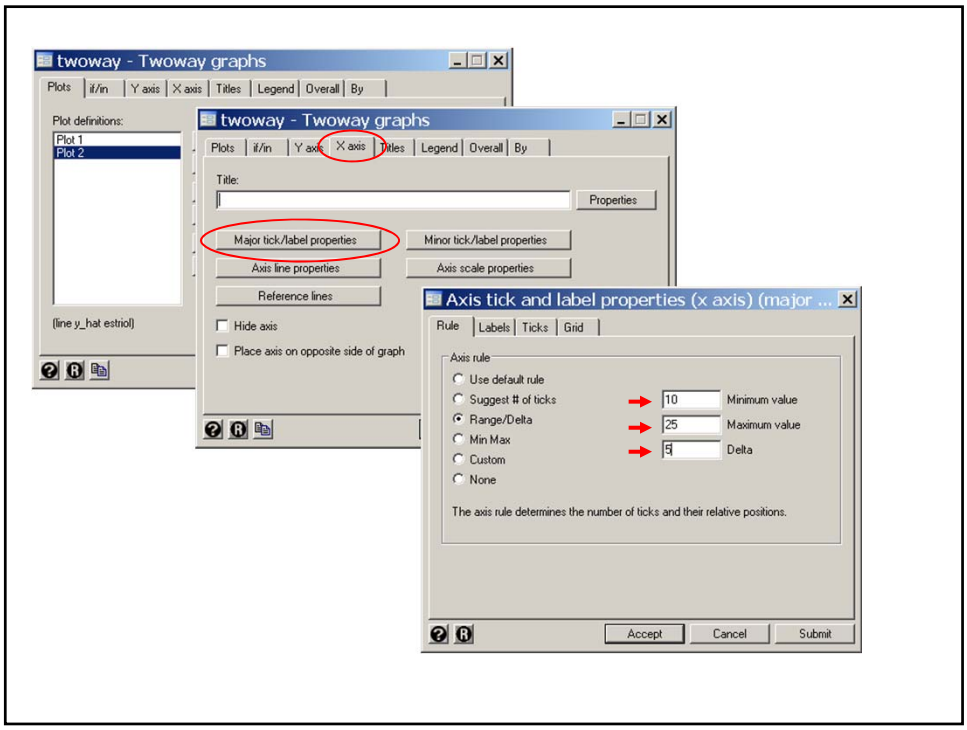

```
. * Graphics > Twoway graph (scatter, line, etc.)
. twoway scatter bweight estriol                                    ///    {2}
>     || line y_hat estriol                                         ///    {3}
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")   {4}
```

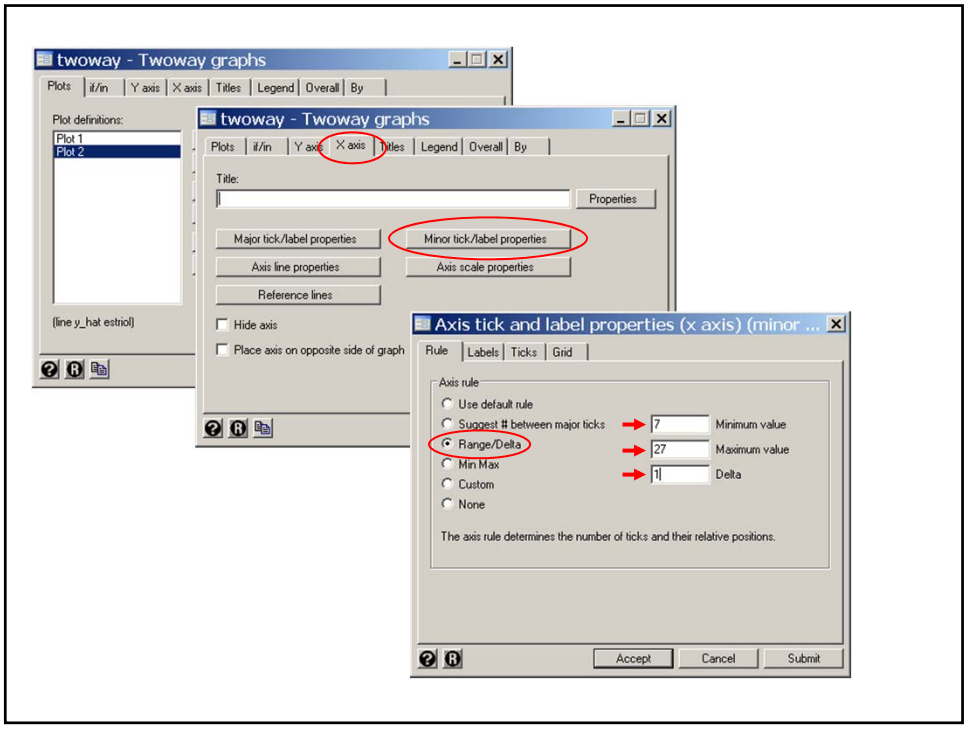
{2} This command is written over several lines to improve legibility. The three slashes (///) indicate that this command continues on the next line. These slashes are permitted in do files but not in the Command window.

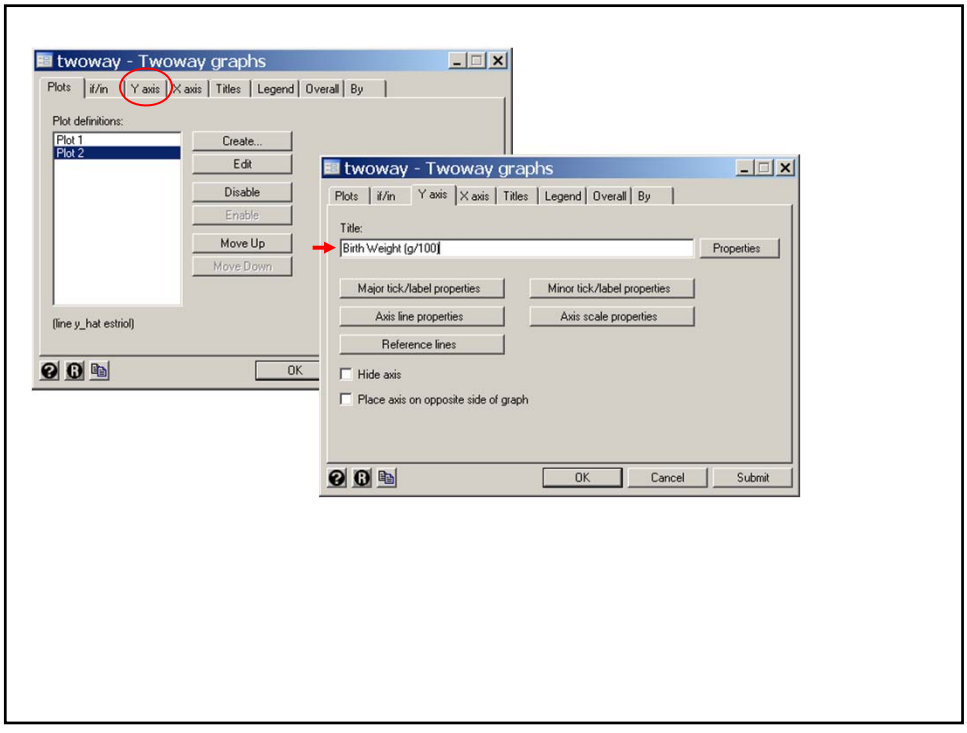
{3} A double bar (||) indicates that another graph is to be overlaid on top of the preceding one. *line y_hat estriol* indicates that *y_hat* is to be plotted against *estriol* with the points connected by a straight line.

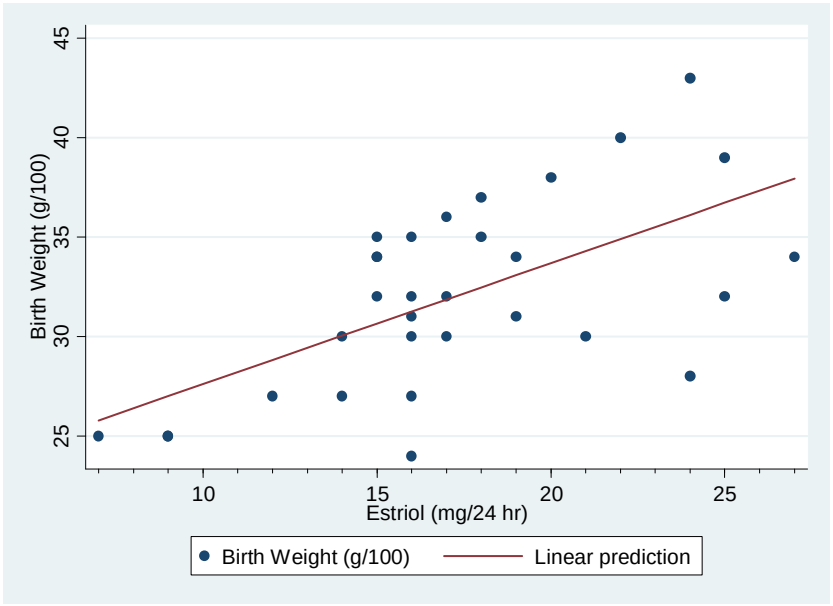
{4} This *xlabel* option labels the x-axis from 10 to 25 in steps of 5. *xmtick* adds tick marks from 7 to 27 in unit steps. *ytitle* gives a title to the y-axis


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
Move Up
Move Down
(scatter bweight estriol)
OK
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Area
Bar
Spike
Plot type: (line plot)
Y variable:
X variable:
y_hat
estriol
Sort on x variable
Add a second y axis on right
Line properties
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
(line y_hat estriol)
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis tick and label properties (x axis) (major ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
10
25
5
Minimum value
Maximum value
Delta
The axis rule determines the number of ticks and their relative positions.
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
(line y_hat estriol)
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis tick and label properties (x axis) (minor ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # between major ticks
Range/Delta
Min Max
Custom
None
7
27
1
Minimum value
Maximum value
Delta
The axis rule determines the number of ticks and their relative positions.
Accept
Cancel
Submit

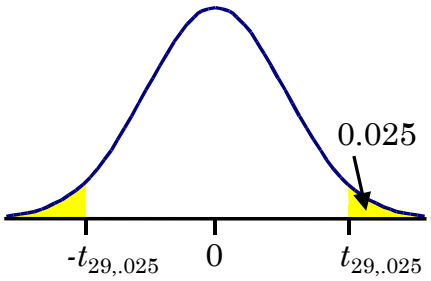

twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
Edit
Disable
Enable
Move Up
Move Down
(line y_hat estriol)
OK
Title:
Birth Weight (g/100)
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines
Hide axis
Place axis on opposite side of graph
OK
Cancel
Submit



9. 95% Confidence Interval (CI) for β

$$
\begin{aligned}
b \pm t_{n\text{-}2,.025}\text{se}(b) &= 0.608 \pm t_{29,.025} \times 0.147 \\
&= 0.608 \pm 2.045 \times 0.147 \\
&= 0.608 \pm 0.300 \\
&= (0.308,\ 0.908)
\end{aligned}
$$

```
  Source |       SS          df          MS            Number of obs =      31
---------+------------------------------------         F(  1,    29) =   17.16
   Model |  250.574476         1     250.574476        Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s² = 14.6008801        R-squared     =  0.3718
---------+------------------------------------         Adj R-squared =  0.3501
   Total |      674.00        30      22.4666667       Root MSE  = s =  3.8211

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.   t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117  4.14 P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417   8.21   0.000       16.16407    26.88278
------------------------------------------------------------------------------
```



10. 95% (CI) for $\alpha + \beta x$

Let $\hat{y}(x) = a + bx$

The variance of $\hat{y}(x)$ is

$$\text{var}(\hat{y}(x)) = [s^2/n] + (x - \bar{x})^2 \text{var}(b) \qquad \{1.8\}$$

The 95% confidence interval for $\hat{y}(x)$ is

$$\hat{y} \pm t_{n-2,0.025}\sqrt{\text{var}(\hat{y}(x))} \qquad \{1.9\}$$

```
  Source |       SS          df          MS            Number of obs =      31
---------+------------------------------------        F(  1,    29) =   17.16
   Model |  250.574476         1     250.574476        Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s²  = 14.6008801       R-squared     =  0.3718
---------+------------------------------------        Adj R-squared =  0.3501
   Total |      674.00        30      22.4666667       Root MSE  = s =  3.8211

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.   t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117  4.14  P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417   8.21    0.000       16.16407    26.88278
------------------------------------------------------------------------------
```

11. Plotting a 95% Confidence Region for the Expected Response

The listing of *Birth_Weight.LR.log* continues as follows

```
. predict std_p, stdp                                                    {1}
. * Data > Create or change data > Create new variable
. generate ci_u = y_hat + invttail(_N-2,0.025)*std_p                     {2}
. generate ci_l = y_hat - invttail(_N-2,0.025)*std_p
```

{1} The *stdp* option of the **predict** command defines *std_p* to be the error of *y_hat*.

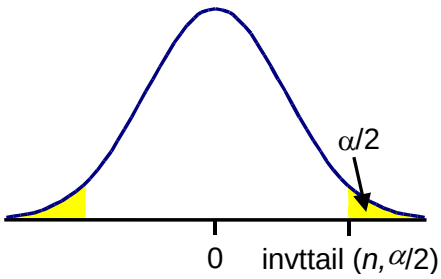
That is $std\_p = \sqrt{\text{var}(\hat{y}(x))}$

{2} **invttail** calculates a critical value of size α for a t distribution with n degrees of freedom.

```
. generate ci_u = y_hat + invttail(_N-2,0.025)*std_p
```

_N denotes the number of variables in the data set, which in this example is 31. Thus *invttail*(_N-2, 0.025) =

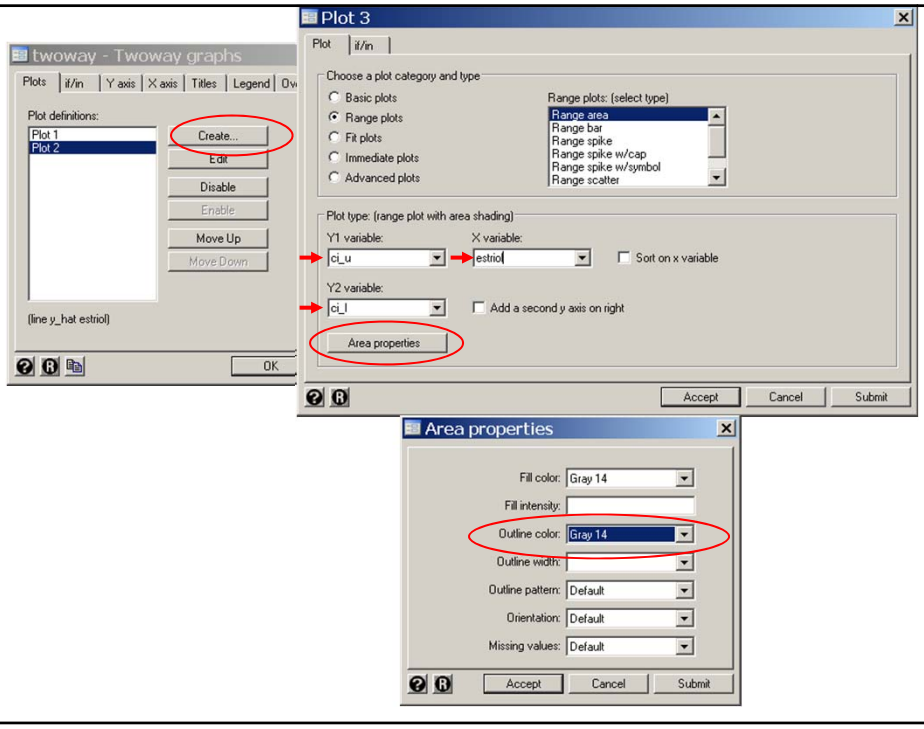
invttail(29, 0.025) = $t_{29,0.025} = 2.045$

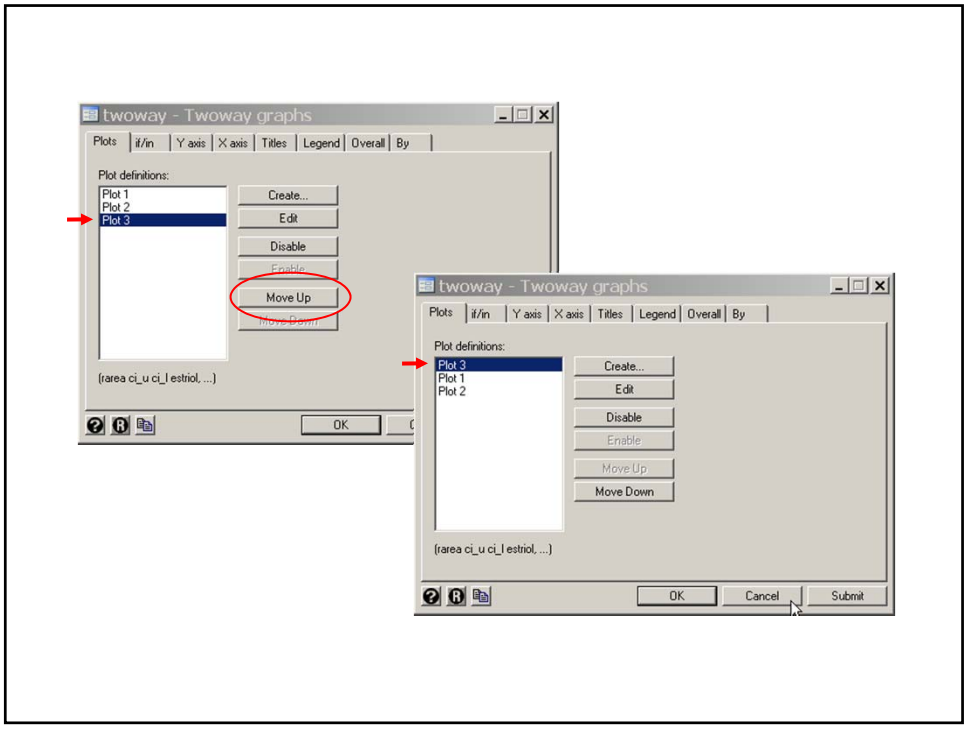


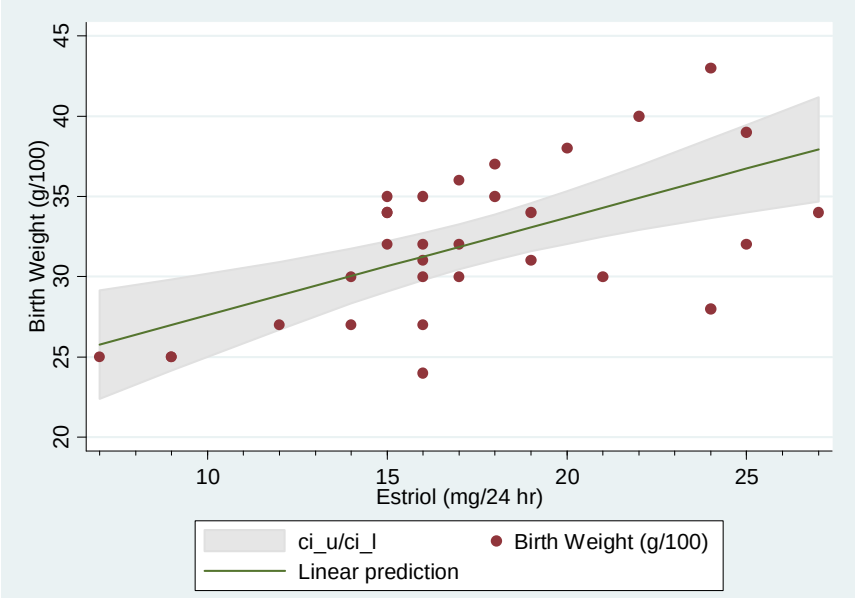
```
. twoway rarea   ci_u ci_l estriol, color(gs14)                    /// {3,4}
>     || scatter bweight estriol                                     ///
>     || line y_hat  estriol                                         ///
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")
```

{3} ***twoway rarea*** shades the region between *ci_u* and *ci_l*

{4} ***color*** selects the color of the shaded region. *gs14* is a gray scale. Gray scales vary from *gs0* which is black through *gs16* which is white.



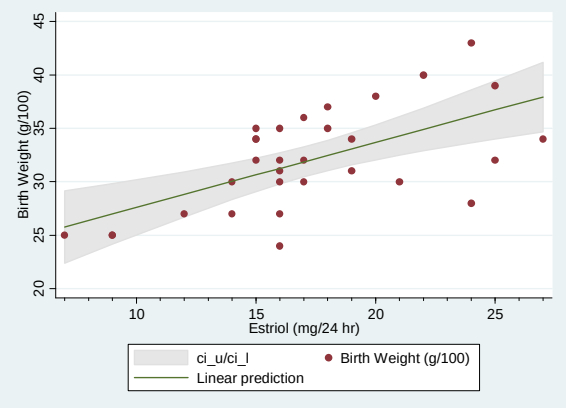




N.B. More than 5% of the observations lie outside of the 95% CI band. Band is for the regression line not the observations.

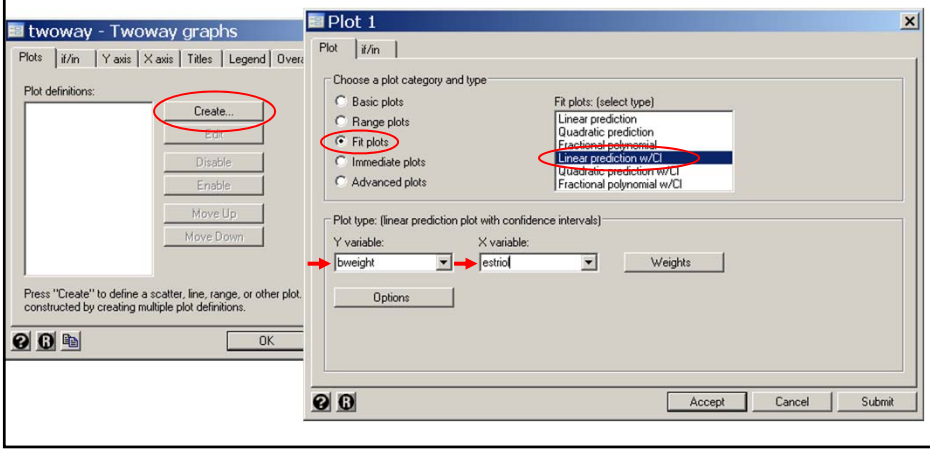
$$\hat{y} \pm t_{n-2,0.025}\sqrt{\text{var}(\hat{y}(x))}$$ Is the 95% confidence band for $\hat{y}(x)$

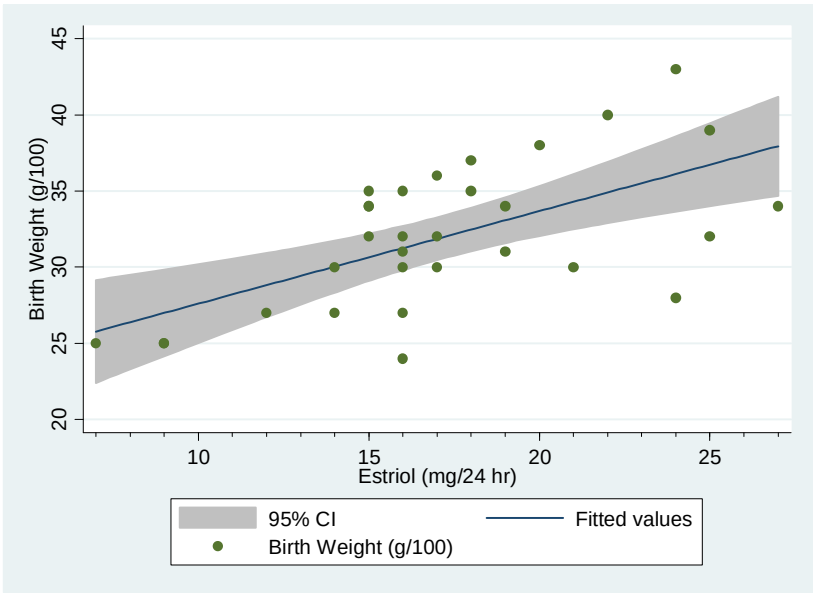
```
. predict std_p , stdp
. generate ci_u = y_hat + invttail(_N-2,0.25)*std_p
. generate ci_l = y_hat - invttail(_N-2,0.25)*std_p
. twoway rarea ci_u ci_l estriol, bcolor(gs14)
>     || scatter bweight estriol                                          ///
>     || line y_hat  estriol                                              ///
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")
```




```
. *
. *  The preceding graph could also have been generated without explicitly
. *  calculating yhat, ci_u or ci_l as follows
. *
. twoway lfitci  bweight estriol                                          /// {1}
>     || scatter bweight estriol                                          ///
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")
```

{1} **lfitci** plots both the regression line and the 95% confidence interval for *bweight* against *estriol*.



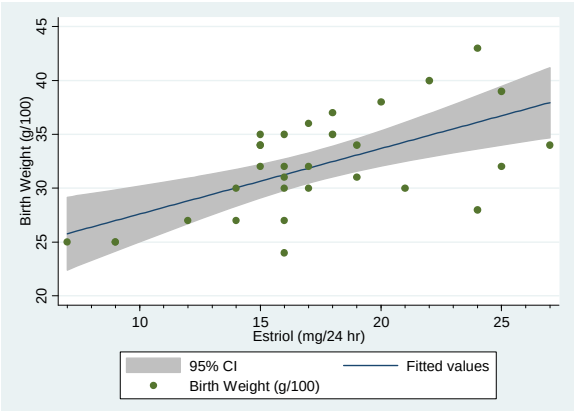


The variance of $\hat{y}(x)$ is

$$\text{var}(\hat{y}(x)) = [s^2/n] + (x - \bar{x})^2 \text{var}(b)$$

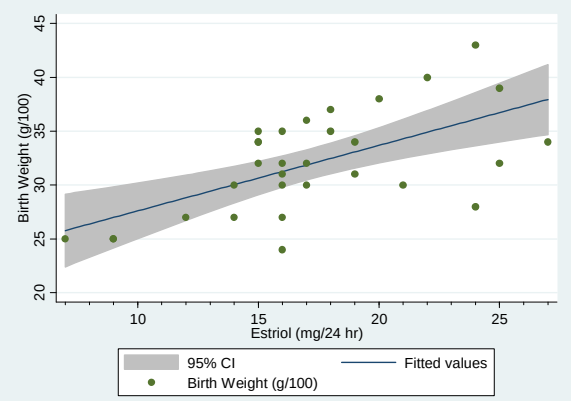
The 95% confidence interval for $\hat{y}(x)$ is

$$\hat{y} \pm t_{n-2,0.025}\sqrt{\text{var}(\hat{y}(x))}$$



$$\text{var}(\hat{y}(x)) = [s^2/n] + (x - \bar{x})^2 \text{var}(b)$$

- Decreases as n increases
- Increases as s increases
- Increases as x diverges from $\bar{x}$
- Increases as var(b) increases



12. Distribution of the Sum of Independent Variables.

Suppose that x has mean μ_x and variance σ_x^2

y has mean μ_y and variance σ_y^2 and

x and y are independent.

Then $x + y$ has mean $\mu_x + \mu_y$ and variance $\sigma_x^2 + \sigma_y^2$

13. The 95% CI for the Forecasted Response of a New Patient

This response is $y = \alpha + \beta x + \varepsilon_i \cong \hat{y}(x) + \varepsilon_i$

The variance of $y = \text{var}(\hat{y}(x)) + s^2$

Therefore, a 95% confidence interval for y is **{1,10}**

$$\hat{y} \pm t_{n-2,0.025}\sqrt{\text{var}(\hat{y}(x)) + s^2}$$

$$\text{var}(\hat{y}(x)) = [s^2/n] + (x - \bar{x})^2 \text{var}(b)$$

If $x = 22$ then

$\text{var}(\hat{y}(22)) = 14.6009/31 + (22 - 17.226)^2 \times 0.1468^2 = 0.9621$

$\hat{y} = 21.523 + 0.6082 \times 22 = 34.903$

95% C.I. for $\hat{y}(x) = 34.903 \pm t_{29,.025} \times \sqrt{0.9621}$

$= 34.903 \pm 2.045 \times 0.981$

$= (32.9, 36.9)$

```
  Source |       SS          df          MS            Number of obs =      31
---------+------------------------------------         F(  1,    29) =   17.16
   Model |  250.574476        1     250.574476         Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s²  = 14.6008801       R-squared     =  0.3718
---------+------------------------------------         Adj R-squared =  0.3501
   Total |      674.00       30      22.4666667        Root MSE  = s =  3.8211

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.   t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117   4.14 P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417   8.21   0.000       16.16407    26.88278
------------------------------------------------------------------------------
```

95% C.I. for y at $x = \hat{y} \pm t_{n-2,0.025}\sqrt{\text{var}(\hat{y}(x)) + s^2}$

If $x = 22$ then

$\text{var}(\hat{y}(22)) = 0.9621$

$\hat{y} = 34.903$

95% C.I. for y at $x = 34.903 \pm 2.045 \times \sqrt{0.9621 + 14.6009}$

$= (26.8, 43.0)$

```
  Source |       SS          df          MS            Number of obs =      31
---------+------------------------------------         F(  1,    29) =   17.16
   Model |  250.574476        1     250.574476         Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s²  = 14.6008801       R-squared     =  0.3718
---------+------------------------------------         Adj R-squared =  0.3501
   Total |      674.00       30      22.4666667        Root MSE  = s =  3.8211

------------------------------------------------------------------------------
 bweight |      Coef.       Std. Err.   t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
 estriol |b= .6081905 se(b)= .1468117   4.14 P=0.000       .3079268    .9084541
   _cons |a= 21.52343       2.620417   8.21   0.000       16.16407    26.88278
------------------------------------------------------------------------------
```

14. Plotting 95% Confidence Intervals for Forecasted Responses

The listing of *Birth_Weight*.LR.*log* continues as follows

```
. predict std_f , stdf                                                    {1}
. generate ci_uf = y_hat + invttail(_N-2,0.025)*std_f
. generate ci_lf = y_hat - invttail(_N-2,0.025)*std_f
. twoway lfitci bweight estriol
>     || scatter bweight estriol                                  ///
>     || line ci_uf estriol, color(red)                           ///  {2}
>     || line ci_lf estriol, color(red)                           ///
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27)                          ///
>        ytitle("Birth Weight (g/100)") legend(off)                   {3}
```

{1} **predict** defines *std_f* to be the forecasted error of y.

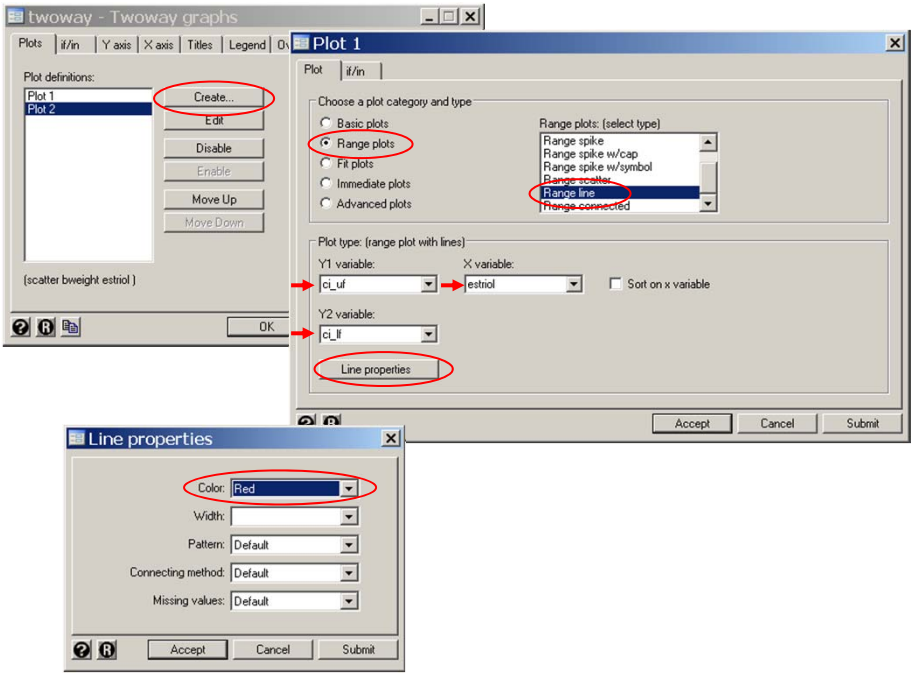
That is *std_f* = $\sqrt{\text{var}(\hat{y}) + s^2}$

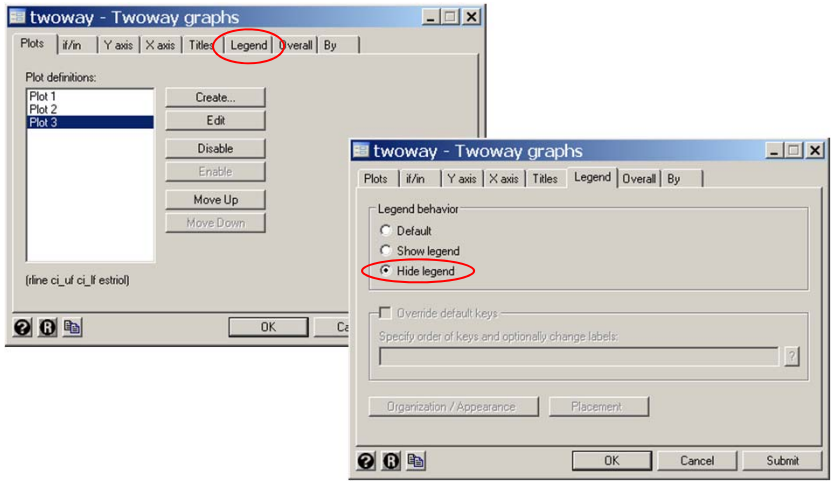
{2} **rline** plots both *ci_uf* and *ci_lf* against *estriol*.

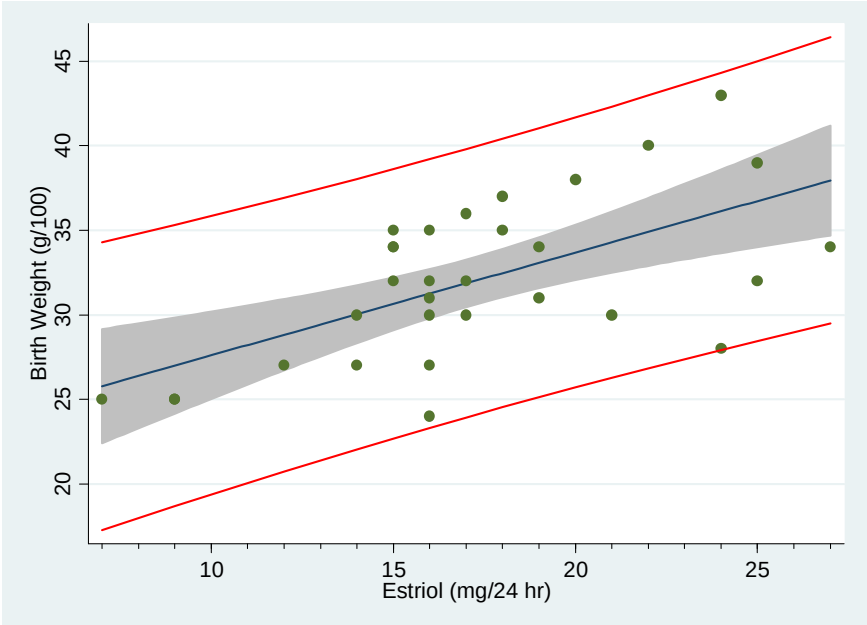
color specifies the color of the plotted lines.

When plotting non-linear lines, it is important that the data be sorted by the x-variable.

{3} **legend(off)** eliminates the legend from the graph




twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Create...
Edit
Disable
Enable
Move Up
Move Down
(lfine ci_uf ci_lf estriol)
OK
twoway - Twoway graphs
Legend behavior
Default
Show legend
Hide legend
Override default keys
Specify order of keys and optionally change labels:
Organization / Appearance
Placement
OK
Cancel
Submit


Birth Weight (g/100)
20
25
30
35
40
45
10
15
20
25
Estriol (mg/24 hr)

15. Lowess Regression

Linear regression is a useful tool for describing a relationship that is linear, or approximately linear. It has the disadvantage that the linear relationship is assumed *a priori*. It is often useful to fit a line through a scatterplot that does not make any model assumptions. One such technique is **lowess regression**, which stands for locally weighted scatterplot smoothing. The idea is that for each observation (x_i, y_i) is fitted to a **separate linear regression** line based on adjacent observations. These points are **weighted** so that the farther away the x value is from x_i, the less effect it has on determining the estimate of $\hat{y}_i$. The proportion of the total data set considered for each $\hat{y}_i$ is called the **bandwidth**. In Stata the default bandwidth is 0.8, which works well for small data sets. For **larger data sets** a bandwidth of **0.3 or 0.4** usually works better.

On large data sets lowess is computationally intensive.

16. Plotting a Lowess Regression Curve in Stata

We can now compare the lowess and linear curves.

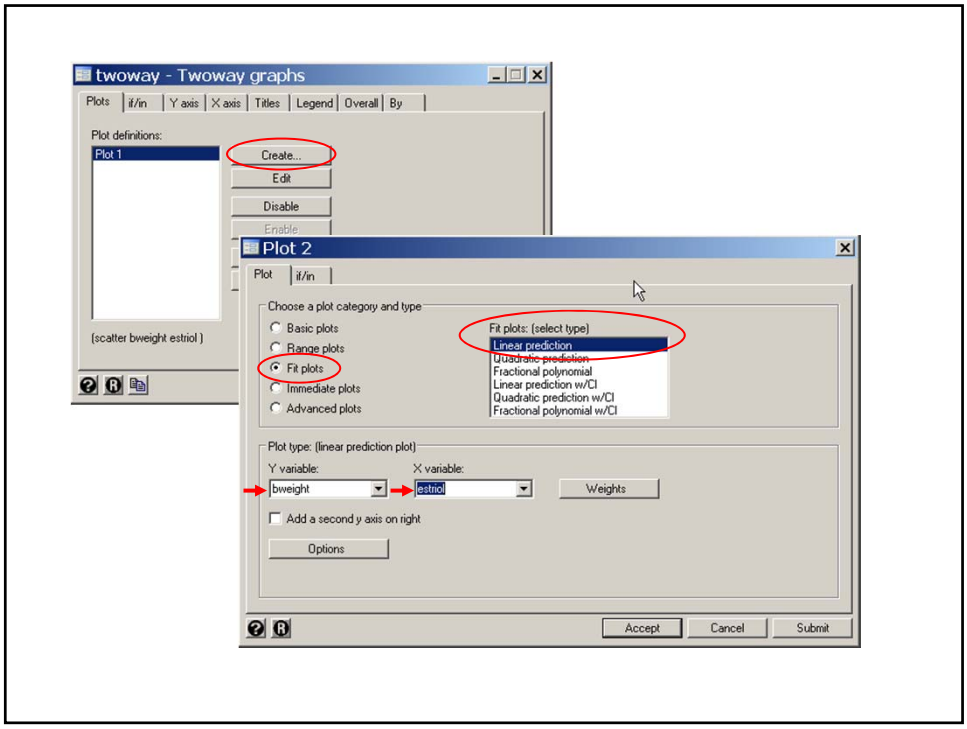
The listing of *Birth_Weight*.LR.*log* continues as follows

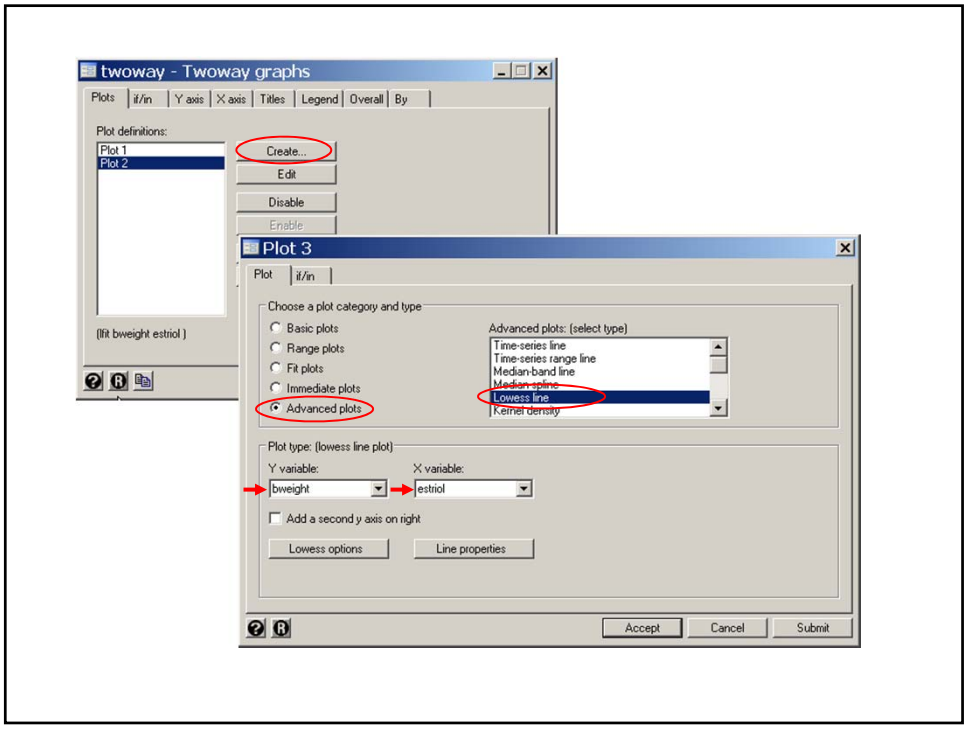
```
.twoway scatter bweight estriol                                    ///
>     || lfit bweight estriol                                      /// {1}
>     || lowess bweight estriol , bwidth(.8)                       /// {2}
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")
```

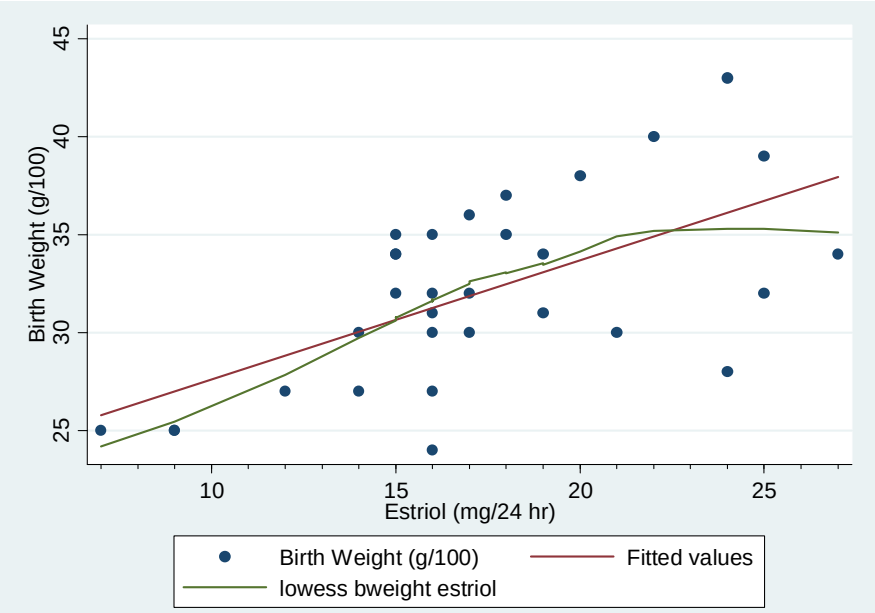
{1} This **lfit** command plots the linear regression line for *bweight* against *estriol*.

{2} This **lowess** command plots the lowess regression line for *bweight* against *estriol*. *bwidth(.8)* specifies a bandwidth of .8

This comparison suggests that the linear model is fairly reasonable.


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
(scatter bweight estriol)
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Fit plots: (select type)
Linear prediction
Quadratic prediction
Fractional polynomial
Linear prediction w/CI
Quadratic prediction w/CI
Fractional polynomial w/CI
Plot type: (linear prediction plot)
Y variable:
X variable:
bweight
estriol
Weights
Add a second y axis on right
Options
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
Edit
Disable
Enable
(lfit bweight estriol)
Plot 3
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Advanced plots: (select type)
Time-series line
Time-series range line
Median-band line
Median-spline
Lowess line
Kernel density
Plot type: (lowess line plot)
Y variable:
X variable:
bweight
estriol
Add a second y axis on right
Lowess options
Line properties
Accept
Cancel
Submit



17. Residuals, Standard Error Of Residuals, And Studentized Residuals

The residual $e_i = y_i - (a + bx_i)$
has variance $\text{var}(e_i) = s^2 - \text{var}(\hat{y}(x))$

The Standardized residual corresponding to the point

(x_i, y_i) is $e_i / \sqrt{s^2 - \text{var}(\hat{y}(x))}$ **{1.11}**

A problem with Equation {1.11} is that a single large residual can

- Inflate the value of s^2
- Decrease the size of the standardized residuals
- Can pull $\hat{y}(x_i)$ towards y_i

To avoid this problem, we usually calculate the studentized residual

$$t_i = e_i / s_{(i)}\sqrt{1-h_i} \qquad \{1.12\}$$

where $s_{(i)}$ denotes the root MSE estimate of σ with the i^{th} case deleted and

$$h_i = \mathrm{var}\left(\hat{y}(x)\right) / s^2_{(i)}$$

is the variance of $\hat{y}(x)$ measured in units of s_i^2. h_i is called the leverage of the i^{th} observation.

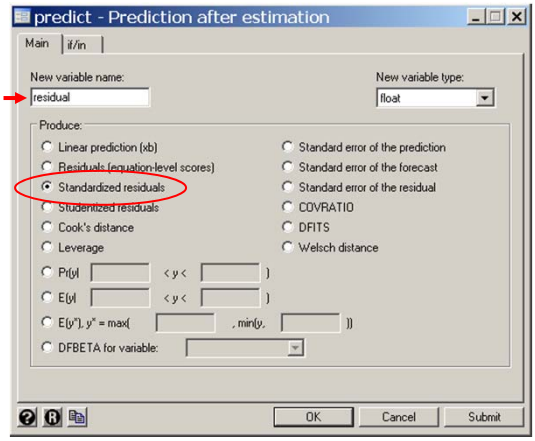
t_i is sometimes referred to as the jackknife residual

Plotting these studentized residuals against x_i assesses the homoscedasticity assumption.

The listing of *Birth_Weight.LR.log* continues as follows

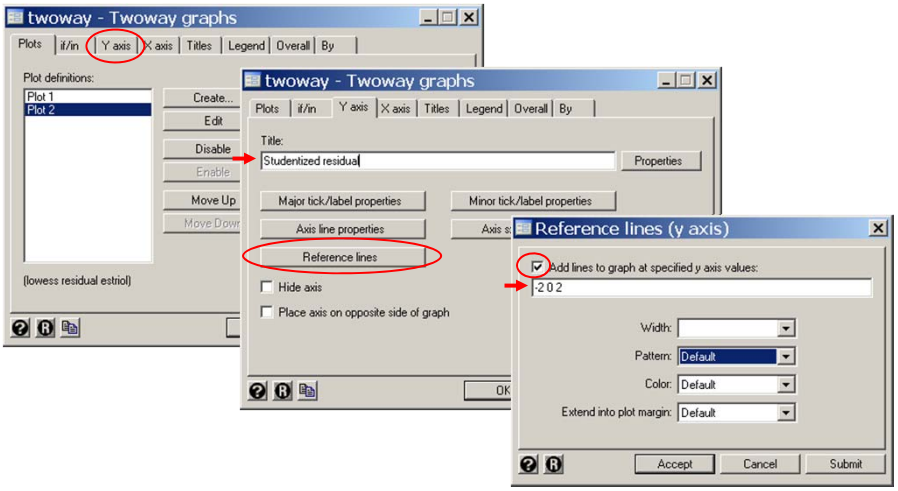
```
. predict residual, rstudent                                    {1}
```

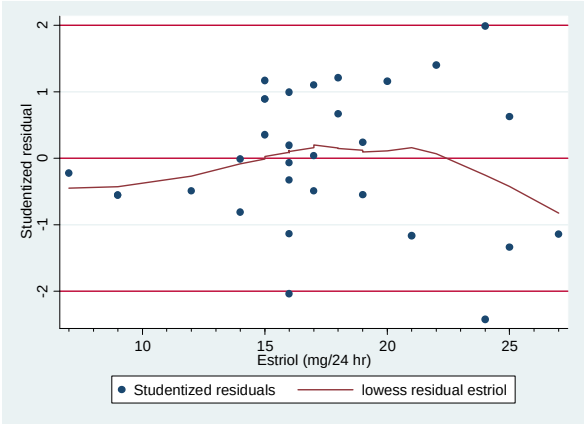
{1} Define *residual* to be a new variable that equals the studentized residual.



```
. twoway scatter residual estriol                                    ///
>     || lowess residual estriol                                     ///
>      , ylabel(-2 (1) 2) yline(-2 0 2) xlabel(10 (5) 25)            ///  {2}
>        xmtick(7 (1) 27)
```

{2} The *yline* option adds horizontal grid lines at specified values on the y-axis.





This graph hints at a slight **departure from linearity** for large estriol levels. There is also a slight suggestion of **increasing variance** with increasing estriol levels, although the dispersion is fairly uniform if you disregard the patients with the lowest three values of estriol. Three of 31 residuals **(9.6%)** have a magnitude of 2 or greater. A large data set from a true linear model should have **5%** of their residuals outside this range.

18. Variance Stabilizing Transformations

a) **Square root transform**

Useful when the residual variance is proportional to the expected value.

b) **Log transform**

Useful when the residual standard deviation is proportional to the expected value.

N.B. Transformations that stabilize variance may cause non-linearity. In this case it may be necessary to use a non-linear regression technique.

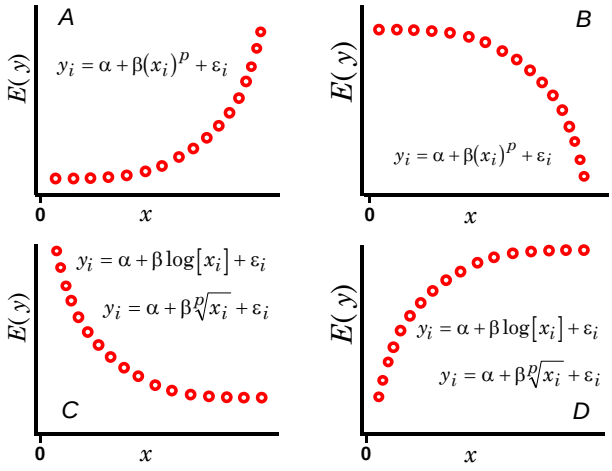
19. Normalizing the Data Distribution

For skewed data we can often improve the quality of our model fit by transforming the data to give the residuals a more normal distribution.

For example, log transforms can normalize data that is right skewed.

The figure shows common patterns on non-linearity between x and y variables.

20. Correcting for Non-linearity



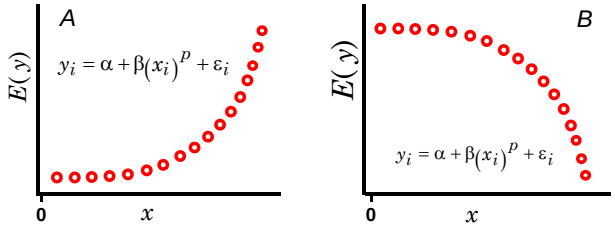
If x is positive then models of the form

$$y_i = \alpha + \beta(x_i)^p + \varepsilon_i \quad \{1.13\}$$

$$y_i = \alpha + \beta\log[x_i] + \varepsilon_i \quad \{1.14\}$$

$$y_i = \alpha + \beta\sqrt[p]{x_i} + \varepsilon_i \quad \{1.15\}$$

should be considered for some $p > 1$.



Data similar to panels A and B of this figure may be modeled with equation {1.13}

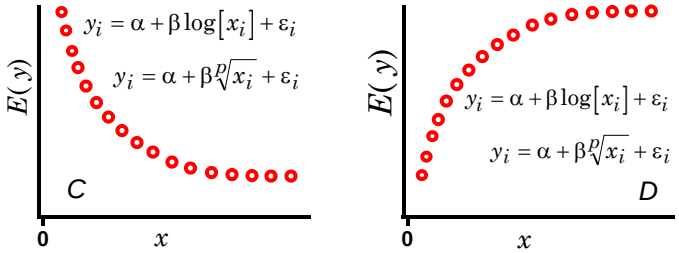
If x is positive then models of the form

$$y_i = \alpha + \beta(x_i)^p + \varepsilon_i \quad \{1.13\}$$

$$y_i = \alpha + \beta\log[x_i] + \varepsilon_i \quad \{1.14\}$$

$$y_i = \alpha + \beta\sqrt[p]{x_i} + \varepsilon_i \quad \{1.15\}$$

should be considered for some $p > 1$.



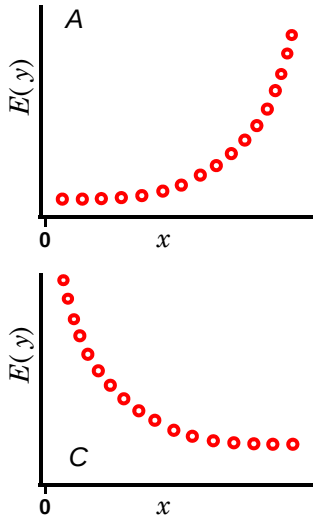
Data similar to panels C and D may be modeled with equations {1.14} and {1.15}.

The best value of p is found empirically.

Alternately, data similar to panels A or C may be modeled with

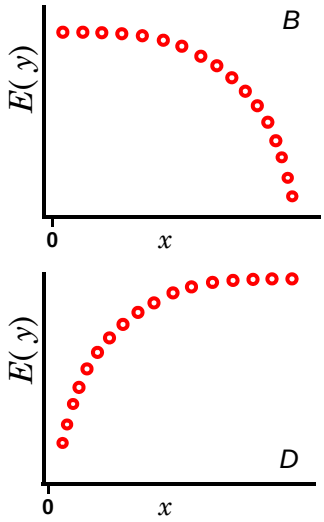
$$\log[y_i] = \alpha + \beta x_i + \varepsilon_i \qquad \{1.16\}$$

or $$\sqrt[p]{y_i} = \alpha + \beta x_i + \varepsilon_i \qquad \{1.17\}$$



Data similar to panels B or D may be modeled with

$$y_i^p = \alpha + \beta x_i + \varepsilon_i \qquad \{1.18\}$$



21. Example: Research Funding and Morbidity for 29 Diseases

Gross et al. (1999) studied the relationship between NIH research funding for 29 different diseases and disability-adjusted person-years of life lost due to these illnesses.

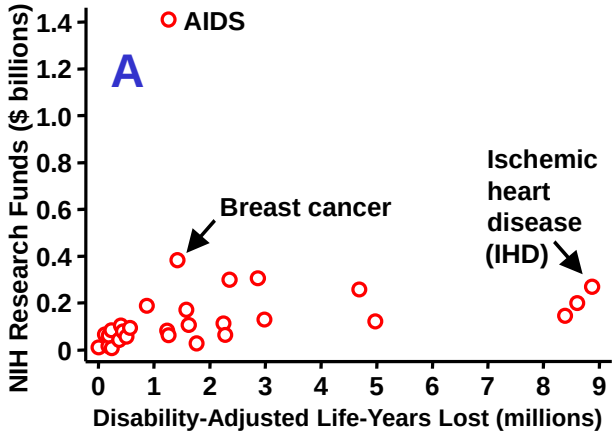


Figure A shows the untransformed scatter plot. Funding for AIDS is 3.7 times higher than for any other disease

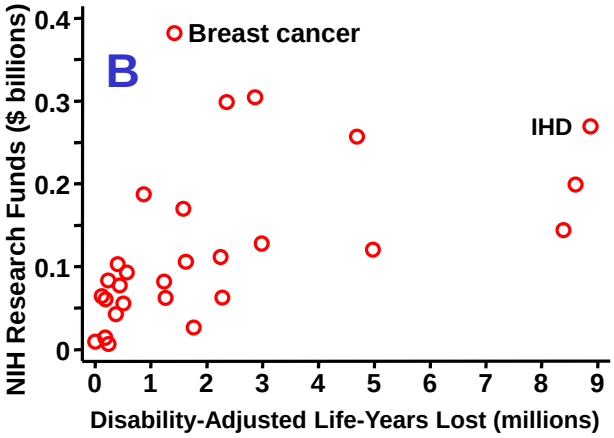


Figure B is similar to panel A except the AIDS data has been deleted and the *y*-axis has been re-scaled.

This scatter plot has a concave shape, which suggest using a log or power transform (equations 1.14 or 1.15).

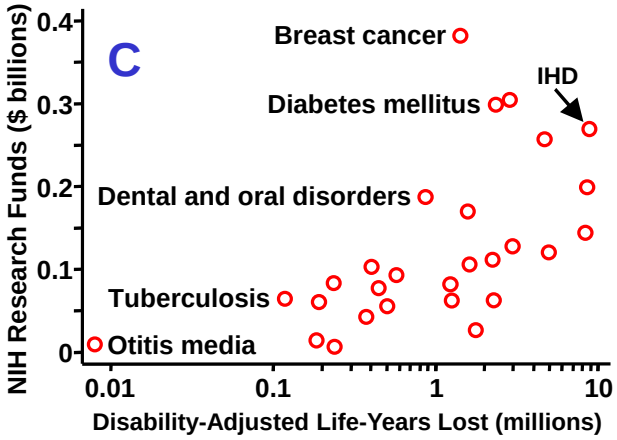
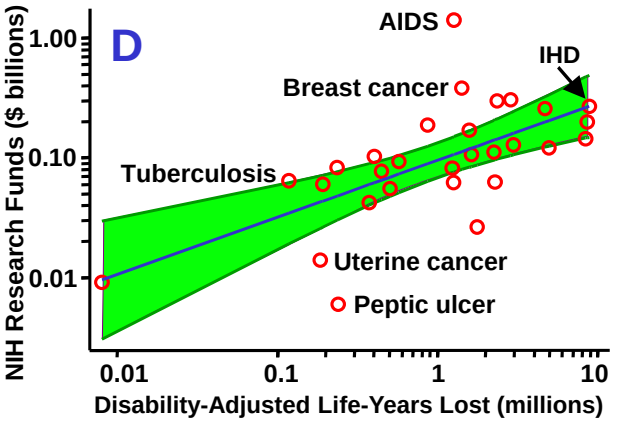


Figure C shows funding plotted against log disability-adjusted life-years.

The resulting scatter plot has a convex shape.

This suggest either using a less concave transform of the *x*-axis or using a log transform of the *y*-axis.

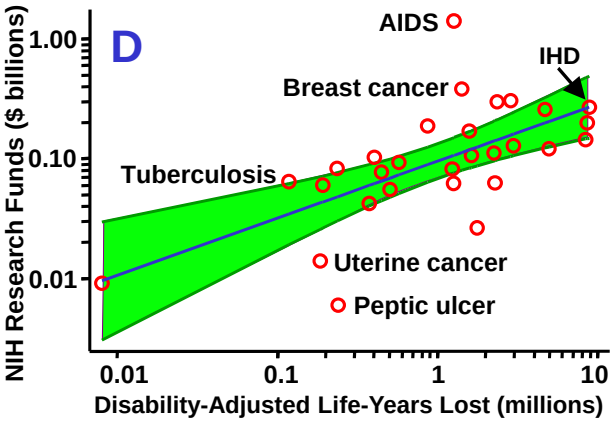


In **Figure D** we plot log funding against log disability.

The relationship between these transformed variables is now quite linear.

AIDS remains an outlier but is far less discordant with the other diseases than it is in panel A.

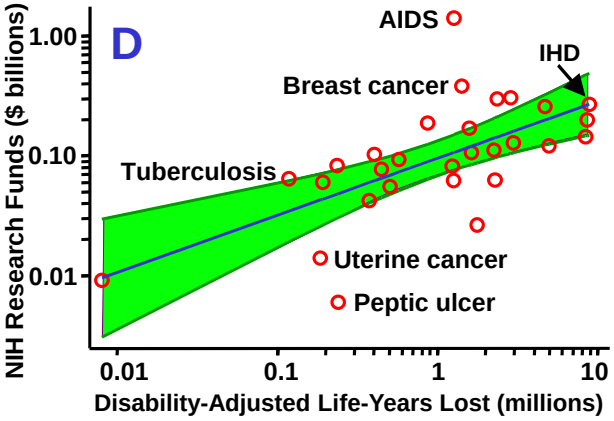
The linear regression line and associated 95% confidence intervals are shown in this panel.



The model for this linear regression is

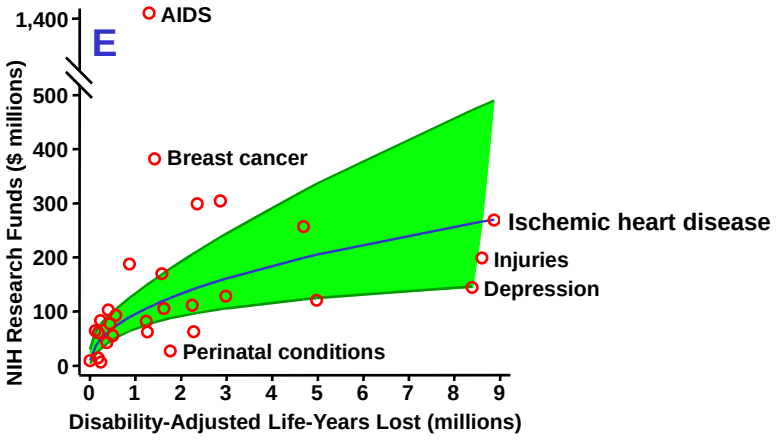
$$E[\log[y_i]]=\alpha+\beta\log[x_i] \qquad \{1.19\}$$

where y_i and x_i are the research funds and disability-adjusted life-years lost for the i^{th} disease, respectively.



The slope estimate is $\beta = 0.48$, which differs from zero with overwhelming statistical significance.

Gross et al. (1999) published a figure that is similar to panel D.



The relationship between funding and life-years lost is more easily understood in **Panel E**, which uses the untransformed data.

If $\log[\hat{y}_i]=a+b\log[x_i]$ is the estimated regression line for the model specified by Equation (1.19) then the predicted funding level for the i^{th} disease is $\hat{y}_i = e^a x_i^b$

To draw this graph we would use the **twoway rarea** command

22. Testing the Equality of Regression Slopes

Suppose that $y_{i1} = \alpha_1 + \beta_1 x_{i1} + \varepsilon_{i1}$ and

$$y_{i2} = \alpha_2 + \beta_2 x_{i2} + \varepsilon_{i2}$$

where ε_{ij} is assumed to be normally and independently distributed with mean 0 and standard deviation σ.

We wish to test the null hypothesis that $\beta_1 = \beta_2$.

The pooled estimate of σ^2 is

$s^2 = [\sum_1 (y_{i1} - \hat{y}_1(x_{i1}))^2 + \sum_2 (y_{i2} - \hat{y}_2(x_{i2}))^2] / (n_1 + n_2 - 4)$

$= (s_1^2(n_1 - 2) + s_2^2(n_2 - 2)) / (n_1 + n_2 - 4)$ **{1.20}**

Since $s_1^2 = \sum_1 \left(y_{i1} - \hat{y}_1(x_{i1})\right)^2 / (n_1 - 2)$ and $s_2^2 = \sum_2 \left(y_{i2} - \hat{y}_2(x_{i2})\right)^2 / (n_2 - 2)$

$$\text{var}(b_1 - b_2) = s^2 \left\{ \frac{1}{\sum_1 (x_{i1} - \bar{x}_1)^2} + \frac{1}{\sum_2 (x_{i2} - \bar{x}_2)^2} \right\} \quad \textbf{\{1.21\}}$$

But $\text{var}(b_1) = s_1^2 / \sum_1 (x_{i1} - \bar{x}_1)^2$ and hence

$\sum_1 (x_{i1} - \bar{x}_1)^2 = s_1^2 / \text{var}(b_1)$

Therefore $\text{var}(b_1 - b_2) = s^2(\text{var}(b_1)/s_1^2 + \text{var}(b_2)/s_2^2)$

$t = (b_1 - b_2)/\sqrt{\text{var}(b_1 - b_2)}$ **{1.22}**

has a t distribution with $n_1 + n_2 - 4$ degrees of freedom.

A 95% CI for $\beta_1 - \beta_2$ is $(b_1 - b_2) \pm t_{n_1+n_2-4,0.025}\sqrt{\text{var}(b_1 - b_2)}$

To compute the preceding statistic we run two separate linear regressions on group 1 and 2.

For group 1, s_1 is the root MSE and $\sqrt{\text{var}(b_1)}$ is the standard error of the slope estimate.

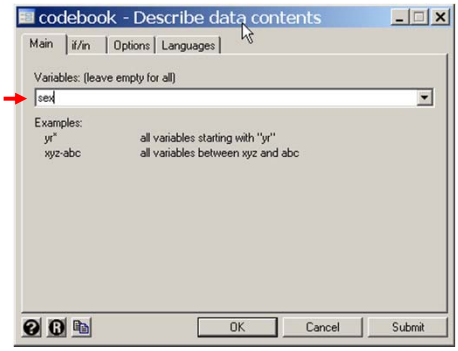
s_2 and $\text{var}(b_2)$ are similarly defined for group 2.

Substituting these values into the formulas from the previous slide gives the required test.

23. Comparing Linear Regression Slopes with Stata

```
. *  FramSBPbmiSex.log
. *
. *     Regression of systolic blood pressure against
. *     body mass index for men and women in the
. *     Framingham Heart Study.  (Levy 1999)
. *
. use "c:\WDDtext\2.20.Framingham.dta", clear
. codebook sex                                                    {1}

sex ---------------------------------------------------- sex
            type:  numeric (float)
           label:  sex

           range:  [1,2]                units:  1
   unique values:  2             coded missing:  0 / 4699

           tabulation:  Freq.   Numeric  Label
                        2049          1  Men
                        2650          2  Women
```

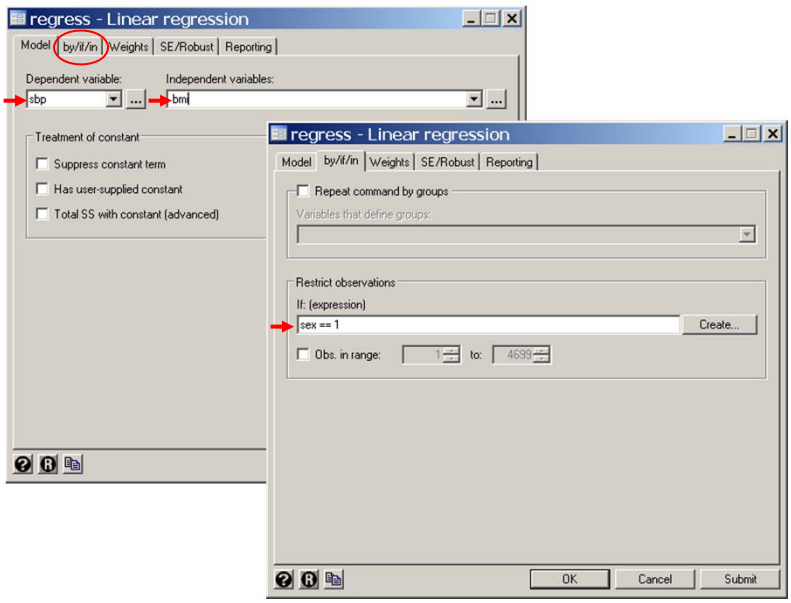
{1} The *codebook* command provides a summary of variables. *sex* takes numeric values 1 and 2 that denote men and women, respectively.



```
. regress sbp bmi if sex == 1                                          {2}

  Source |       SS       df       MS              Number of obs =    2047
---------+------------------------------           F(  1,  2045) =  121.09
   Model |  44504.0296     1  44504.0296            Prob > F      =  0.0000
Residual |  751572.011  2045  367.516876            R-squared     =  0.0559
---------+------------------------------           Adj R-squared =  0.0554
   Total |  796076.041  2046  389.088974            Root MSE      =  19.171

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   1.375953   .1250382     11.004   0.000       1.130738    1.621168
   _cons |   96.43061   3.272571     29.466   0.000       90.01269    102.8485
------------------------------------------------------------------------------

. predict yhatmen, xb
(9 missing values generated)
```

Comment **{2}**

This command regresses *sbp* against *bmi* in men only. Note that Stata distinguishes between $a = 1$, which assigns the value 1 to a, and $a==1$ which is a logical expression that is true if the a equals 1 and is false otherwise.

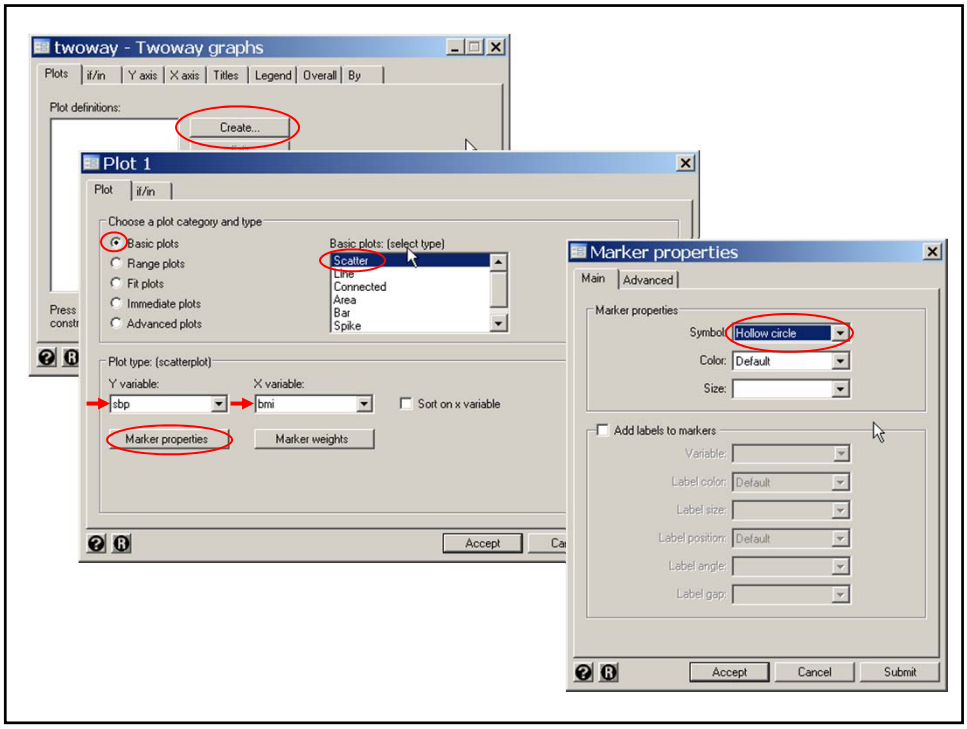


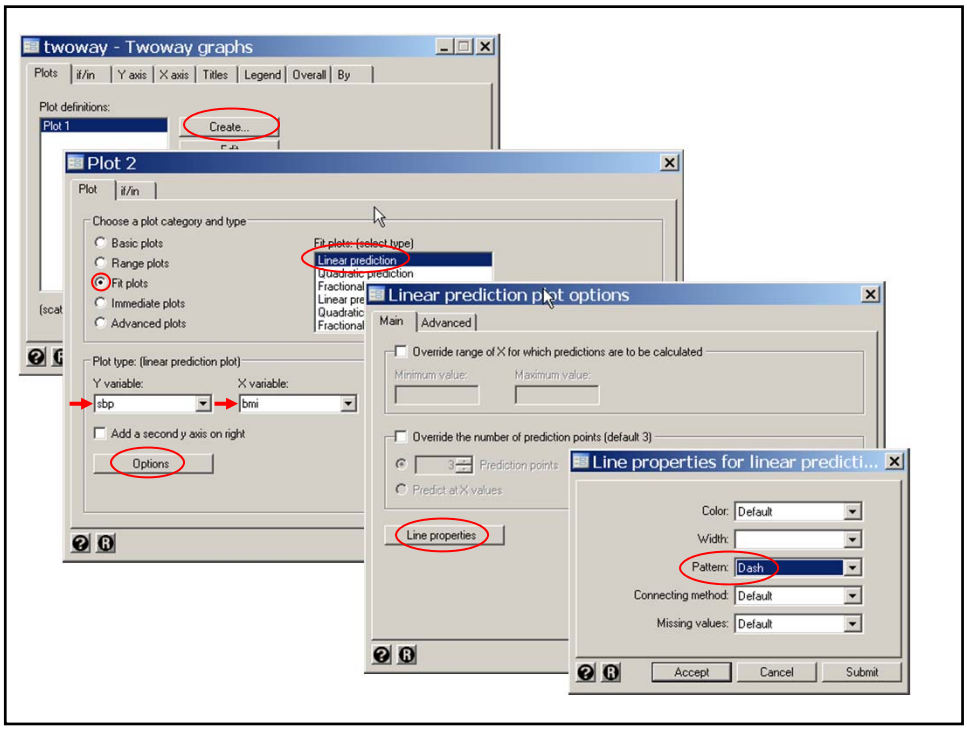
```
. regress sbp bmi if sex == 2

  Source |       SS       df       MS              Number of obs =    2643
---------+------------------------------           F(  1,  2641) =  428.48
   Model |  229129.452     1  229129.452           Prob > F      =  0.0000
Residual |  1412279.85  2641  534.751932           R-squared     =  0.1396
---------+------------------------------           Adj R-squared =  0.1393
   Total |  1641409.31  2642  621.275286           Root MSE      =  23.125

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   2.045966   .0988403     20.700   0.000       1.852154    2.239779
   _cons |   81.30435   2.548909     31.898   0.000       76.30629    86.30241
------------------------------------------------------------------------------

. predict yhatwom, xb
(9 missing values generated)
```

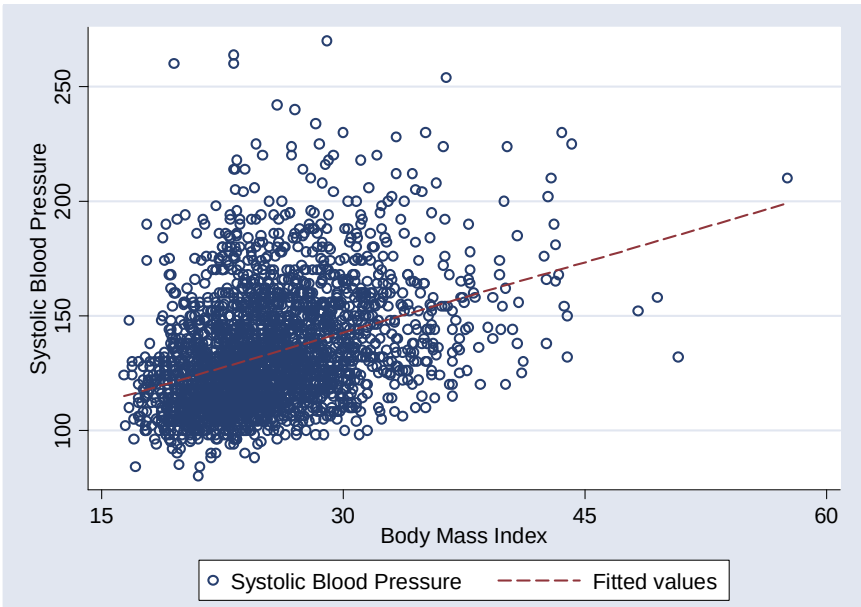
```
. sort bmi

. *
. *  Scatter plot with expected sbp for women
. *
. scatter sbp  bmi if sex==2 , symbol(Oh)                                  /// {1}
>     || lfit sbp bmi if sex==2                                            ///
>      , lpattern(dash)  xlabel(15 (15) 60)                                /// {2}
>        ylabel( 100 (50) 250 ) ytitle(Systolic Blood Pressure)
```

{1} **symbol** specifies the marker symbol used to specify individual observations. **Oh** indicates that a hollow circle is to be used.

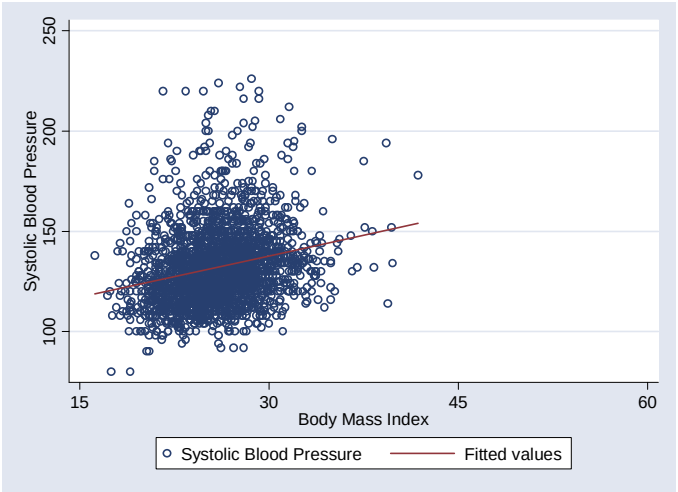
{2} **lpattern(dash)** specifies that a dashed line is to be drawn.


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Create...
Plot 1
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Area
Bar
Spike
Plot type: (scatterplot)
Y variable:
X variable:
sbp
bmi
Sort on x variable
Marker properties
Marker weights
Accept
Marker properties
Main
Advanced
Marker properties
Symbol:
Hollow circle
Color:
Default
Size:
Add labels to markers
Variable:
Label color:
Label size:
Label position:
Label angle:
Label gap:
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Fit plots: (select type)
Linear prediction
Quadratic prediction
Plot type: (linear prediction plot)
Y variable:
X variable:
sbp
bmi
Add a second y axis on right
Options
Linear prediction plot options
Main
Advanced
Override range of X for which predictions are to be calculated
Minimum value:
Maximum value:
Override the number of prediction points (default 3)
Prediction points
Predict at X values
Line properties
Line properties for linear predicti...
Color:
Default
Width:
Pattern:
Dash
Connecting method:
Default
Missing values:
Default
Accept
Cancel
Submit




```
. *
. *  Scatter plot with expected logsbp for men
. *
. scatter sbp  bmi if sex==1 , symbol(Oh)                                    ///
>     || lfit sbp bmi if sex==1                                              ///
>      , clpat(dash)  xlabel(15 (15) 60)                                     ///
>        ylabel( 100 (50) 250 ) ytitle(Systolic Blood Pressure)
```



```
. scatter sbp bmi , symbol(Oh) color(gs10)                                  /// {1}
>     ||  line yhatmen bmi if sex == 1, lwidth(medthick)                    /// {2}
>     ||  line yhatwom bmi, lwidth(medthick) lpattern(dash)                 ///
>     ||, by(sex) ytitle(Systolic Blood Pressure) xsize(8)                  /// {3}
>         ylabel( 100 (50) 250) ytick(75 (25) 225)                          ///
>         xtitle(Body Mass Index) xlabel( 15 (15) 60) xmtick(20 (5) 55)      ///
>         legend(order(1 "Observed" 2 "Expected, Men" 3                     /// {4}
>             "Expected, Women") rows(1))                                       {5}
```

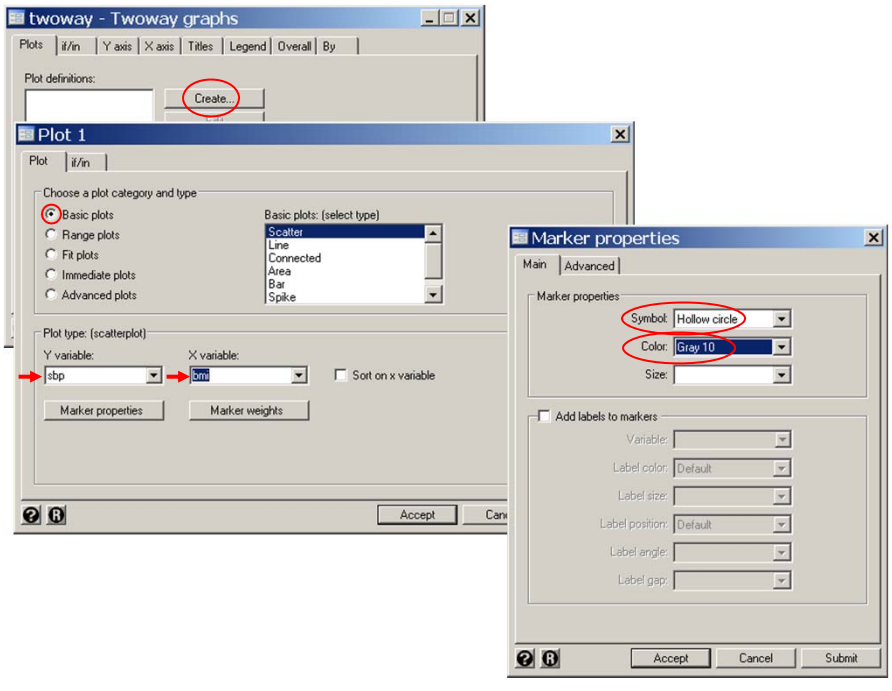
{1} **color** specifies the marker color used to specify individual observations. **gs10** indicates a gray is to be used. A gray scale of 0 is black; 16 is white.

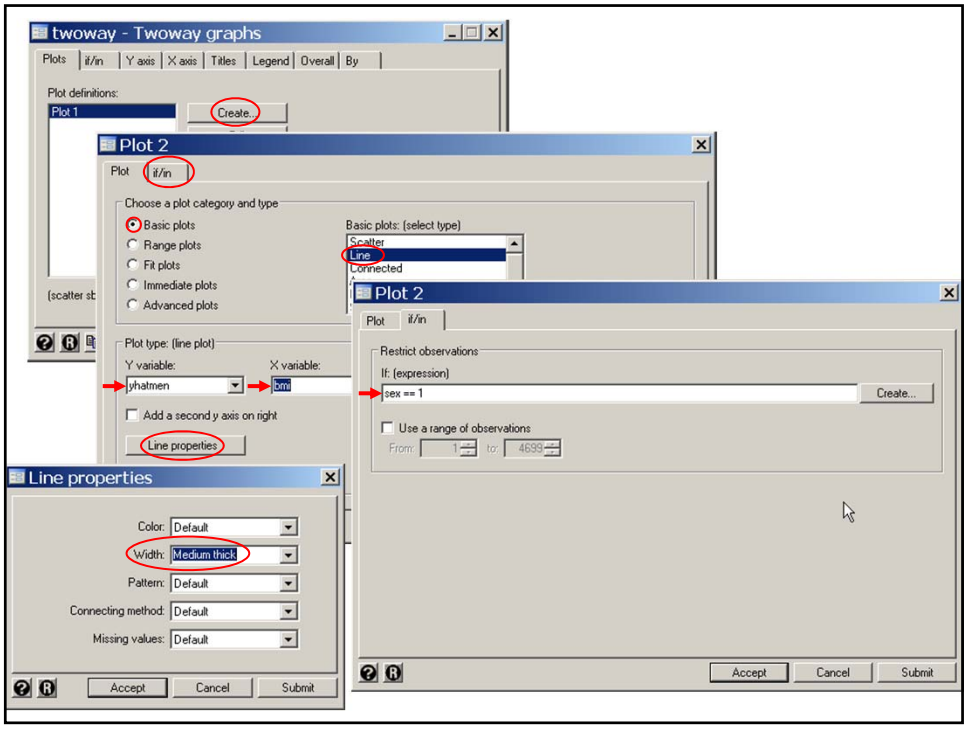
{2} **lwidth(medthick)** specifies width for the line

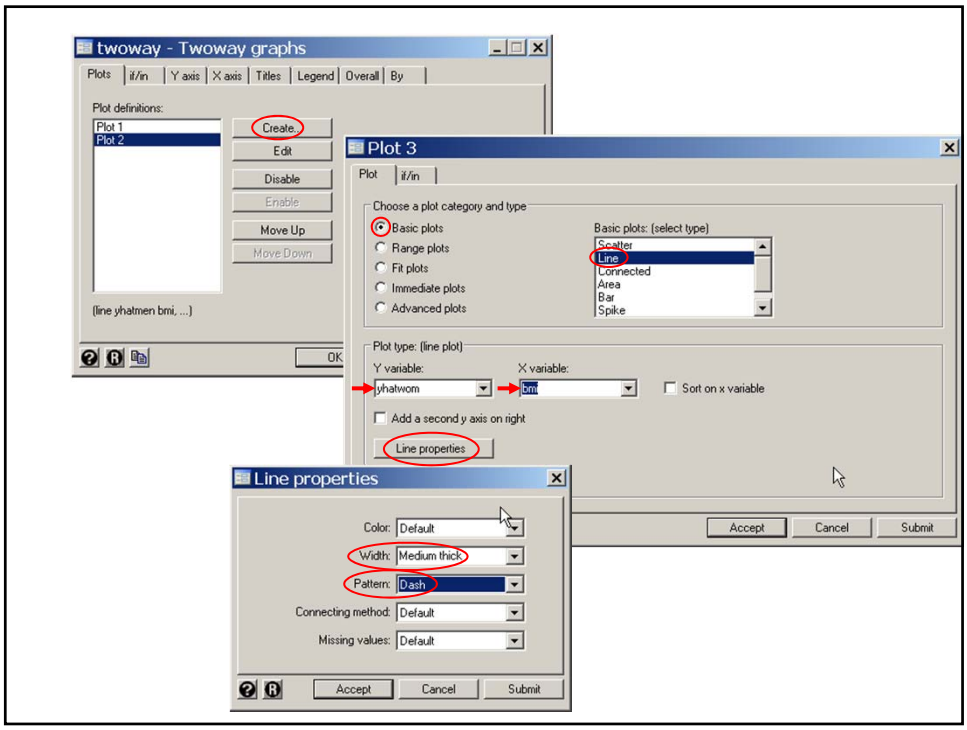
{3} **xsize** specifies the width of the graph in inches

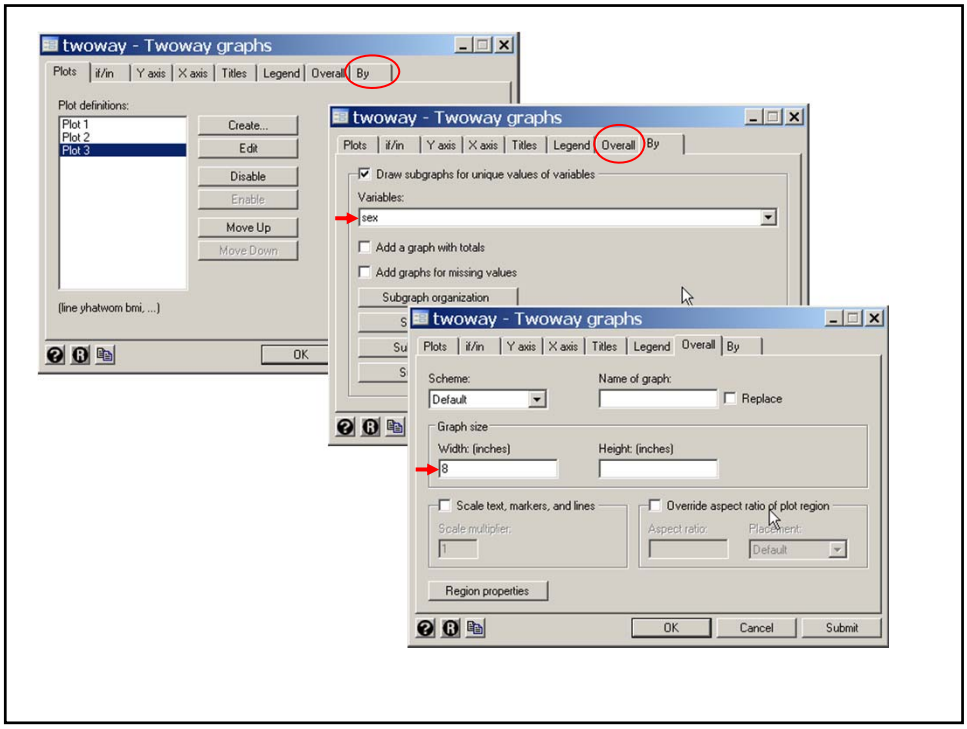
{4} **legend** controls the graph legend. **order** specifies the legend keys to be displayed and the labels assigned to these keys. The keys are indicated by the numbers 1, 2, 3, etc that are assigned in the order that they are defined.

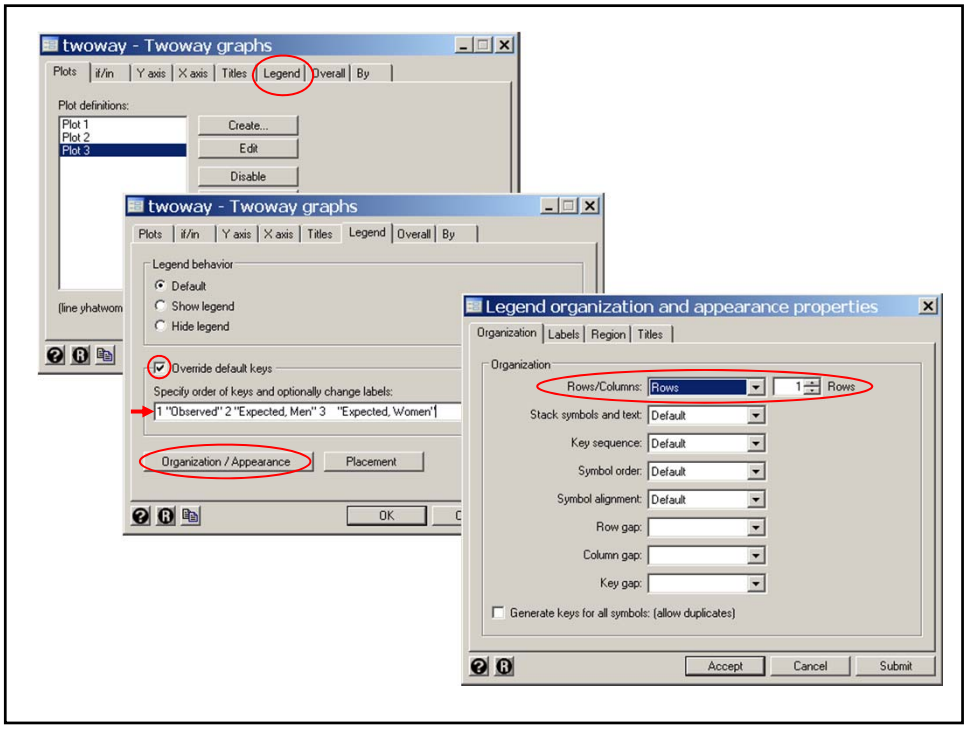
{5} **rows(1)** specifies that we want the legend displayed in a single row

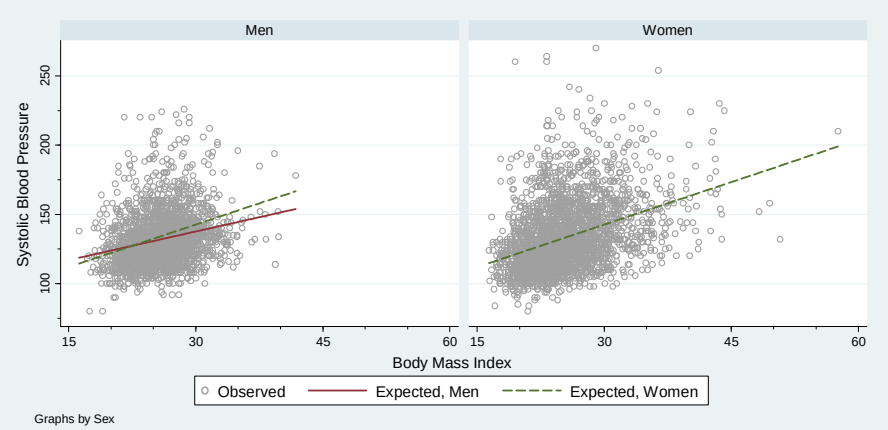



twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Plot type: (line plot)
Y variable:
X variable:
yhatmen
bmi
Add a second y axis on right
Line properties
Plot 2
Plot
if/in
Restrict observations
If: (expression)
sex == 1
Create...
Use a range of observations
From:
to:
Line properties
Color:
Default
Width:
Medium thick
Pattern:
Default
Connecting method:
Default
Missing values:
Default
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
Edit
Disable
Enable
Move Up
Move Down
(line yhatmen bmi, ...)
OK
Plot 3
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Area
Bar
Spike
Plot type: (line plot)
Y variable:
X variable:
yhatwom
bmi
Sort on x variable
Add a second y axis on right
Line properties
Line properties
Color:
Default
Width:
Medium thick
Pattern:
Dash
Connecting method:
Default
Missing values:
Default
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Create...
Edit
Disable
Enable
Move Up
Move Down
(line yhatwom bmi, ...)
OK
Draw subgraphs for unique values of variables
Variables:
sex
Add a graph with totals
Add graphs for missing values
Subgraph organization
Scheme:
Default
Name of graph:
Replace
Graph size
Width: (inches)
8
Height: (inches)
Scale text, markers, and lines
Scale multiplier:
Override aspect ratio of plot region
Aspect ratio:
Placement:
Region properties
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Create...
Edit
Disable
Legend behavior
Default
Show legend
Hide legend
Override default keys
Specify order of keys and optionally change labels:
1 "Observed" 2 "Expected, Men" 3 "Expected, Women"
Organization / Appearance
Placement
OK
Legend organization and appearance properties
Organization
Labels
Region
Titles
Rows/Columns:
Rows
Stack symbols and text:
Key sequence:
Symbol order:
Symbol alignment:
Row gap:
Column gap:
Key gap:
Generate keys for all symbols: (allow duplicates)
Accept
Cancel
Submit



For a body mass index of 40, the expected SBPs for men and women are 96.43 + 40 x 1.38 = **152** *and* 81.30 + 40 x 2.05 = **163**, respectively.

Is it plausible that the effect of BMI on blood pressure is different in men and women?

To test the equality of the slopes from the standard and experimental preparations, we note that

$$s^2 = (s_1^2(n_1 - 2) + s_2^2(n_2 - 2)) / (n_1 + n_2 - 4)$$

$$= \frac{367.52 \times (2047 - 2) + 534.75 \times (2643 - 2)}{2047 + 2643 - 4}$$

$$= 461.77$$

Men

```
  Source |       SS       df       MS              Number of obs =    2047
---------+------------------------------           F(  1,  2045) =  121.09
   Model |  44504.0296     1  44504.0296           Prob > F      =  0.0000
Residual |  751572.011  2045  367.516876           R-squared     =  0.0559
---------+------------------------------           Adj R-squared =  0.0554
   Total |  796076.041  2046  389.088974           Root MSE      =  19.171

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   1.375953   .1250382     11.004   0.000       1.130738    1.621168
   _cons |   96.43061   3.272571     29.466   0.000       90.01269    102.8485
------------------------------------------------------------------------------
```

To test the equality of the slopes from the standard and experimental preparations, we note that

$$s^2 = (s_1^2(n_1 - 2) + s_2^2(n_2 - 2)) / (n_1 + n_2 - 4)$$

$$= \frac{367.52 \times (2047 - 2) + 534.75 \times (2643 - 2)}{2047 + 2643 - 4}$$

$$= 461.77$$

Women

```
  Source |       SS       df       MS              Number of obs =    2643
---------+------------------------------           F(  1,  2641) =  428.48
   Model |  229129.452     1  229129.452           Prob > F      =  0.0000
Residual |  1412279.85  2641  534.751932           R-squared     =  0.1396
---------+------------------------------           Adj R-squared =  0.1393
   Total |  1641409.31  2642  621.275286           Root MSE      =  23.125

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   2.045966   .0988403     20.700   0.000       1.852154    2.239779
   _cons |   81.30435   2.548909     31.898   0.000       76.30629    86.30241
------------------------------------------------------------------------------
```

$$\sum_1 (x_{i1} - \bar{x}_1)^2 = s_1^2 / \text{var}(b_1) = 367.52/0.12504^2 = 23506$$

Men

```
  Source |       SS       df       MS              Number of obs =    2047
---------+------------------------------           F(  1,  2045) =  121.09
   Model |  44504.0296     1  44504.0296           Prob > F      =  0.0000
Residual |  751572.011  2045  367.516876           R-squared     =  0.0559
---------+------------------------------           Adj R-squared =  0.0554
   Total |  796076.041  2046  389.088974           Root MSE      =  19.171

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   1.375953   .1250382     11.004   0.000       1.130738    1.621168
   _cons |   96.43061   3.272571     29.466   0.000       90.01269    102.8485
------------------------------------------------------------------------------
```

$$\sum\nolimits_2 (x_{i2} - \bar{x}_2)^2 = s_2^2 / \text{var}(b_2) = 534.75/0.09884^2 = 54738$$

Women

```
  Source |       SS       df       MS              Number of obs =    2643
---------+------------------------------           F(  1,  2641) =  428.48
   Model |  229129.452     1  229129.452           Prob > F      =  0.0000
Residual |  1412279.85  2641  534.751932           R-squared     =  0.1396
---------+------------------------------           Adj R-squared =  0.1393
   Total |  1641409.31  2642  621.275286           Root MSE      =  23.125

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   2.045966   .0988403     20.700   0.000       1.852154    2.239779
   _cons |   81.30435   2.548909     31.898   0.000       76.30629    86.30241
------------------------------------------------------------------------------
```

$$\text{var}(b_1 - b_2) = s^2 \left\{ \frac{1}{\sum_1 (x_{i1} - \bar{x}_1)^2} + \frac{1}{\sum_2 (x_{i2} - \bar{x}_2)^2} \right\}$$

$$= 461.77 \times (1/23506 + 1/54738) = 0.02808$$

$t = (1.3760 - 2.0460) / \sqrt{0.02808}$

$= -4.00$ with 4686 degrees of freedom. $P = 0.00006$

$= 2 \times \text{ttail}(4686,4)$

A 95% CI for $b_1 - b_2$ is

$$1.3760 - 2.0460 \pm t_{4686,0.025} \sqrt{0.02808}$$

$$= -0.67 \pm 1.96 \times 0.1676$$

$$= (-1.0, -0.34)$$

24. Analyzing Subsets in Stata

The previous example illustrated how to restrict analyses to a subgroup such as men or women. This can be extended to more complex selections. Suppose that *sex* = 1 for males, 2 for females and that *age* = 1 for people < 10 years old,
= 2 for people 10 to 19 years old, and
= 3 for people ≥ 20 years old. Then

`sex==2 & age != 2` selects females who are not 10 to 19 years old. If there are no missing values this is equivalent to

`sex==2 & (age==1 | age == 3)`

`sex==1 | age==3` selects all men plus all women ≥ 20.

Logical expressions may be used to define new variables (*generate* command) , to drop records from the data set (*keep* or *drop* command) or to restrict the data used by analysis commands such as *regress*.

Logical expressions evaluate to 1 if true, 0 if false. Stata considers any non-zero value to be true.

Logical expressions may be used to keep or drop observations from the data. For example

```
. keep sex == 2 & age == 1
```

will keep young females and drop all other observations from memory

```
. drop sex == 2 | age == 1
```

will drop women and young people keeping males ≥ 10 years old.

```
. regress sbp bmi if  age == 1 | age == 3
```

will regress *sbp* against *bmi* for people who are less than 10 or at least 20 years of age.

24. What we have covered.

- Distinction between a parameter and a statistic
- The normal distribution
- Inference from a known sample about an unknown target population
- Simple linear regression: Assessing simple relationships between two continuous variables
- Interpreting the output from a linear regression program. Analyzing data with Stata
- Plotting linear regression lines with confidence bands
- Making inferences from simple linear regression models
- Lowess regression and residual plots. How do you know you have the right model?
- Transforming data to improve model fit
- Comparing slopes from two independent linear regressions

Cited References

Greene JW, Jr., Touchstone JC. Urinary estriol as an index of placental function. A study of 279 cases. *Am J Obstet Gynecol* 1963;85:1-9.

Gross, C. P., G. F. Anderson, et al. (1999). "The relation between funding by the National Institutes of Health and the burden of disease." *N Engl J Med* 340(24): 1881-7.

Levy D, National Heart Lung and Blood Institute., Center for Bio-Medical Communication. *50 Years of Discovery : Medical Milestones from the National Heart, Lung, and Blood Institute's Framingham Heart Study*. Hackensack, N.J.: Center for Bio-Medical Communication Inc.; 1999.

Rosner B. *Fundamentals of Biostatistics. 6th ed.* Belmont CA: Danbury 2006

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.

II. MULTIPLE LINEAR REGRESSION

- Extend simple linear regression to models with multiple covariates
- Meaning of parameters in a multiple linear regression model
- Exploratory data analysis
  - Density distribution sunflower plots for displaying high density bivariate data
  - Matrix scatterplots
- Additive models and models with interaction terms
- Building and interpreting complex linear models
- Stepwise methods of building regression models
- Model validation: Evaluating residuals, leverage and influence
- Goodness of model fit vs. model complexity: Using AIC and BIC to choose a good model.
- Restricted cubic splines: Using multiple linear regression to model non-linear relationships between continuous variables.
- Calculating 95% confidence bands for regression curves from restricted cubic spline models.

1. The Model

$$y_i = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

where

$\alpha, \beta_1, \beta_2, \ldots, \beta_k$ are unknown parameters,

$x_{i1}, x_{i2}, \ldots, x_{ik}$ are known variables,

ε_i are independently distributed and has a normal distribution with mean 0 and standard deviation σ, and

y_i is the value of the response variable for the i^{th} patient.

We usually assume that the patient's response y is causally related to the variables $x_{i1}, x_{i2}, \ldots, x_{ik}$ through the model. These latter variables are called **covariates** or **explanatory variables**; y is called the **dependent** or **response variable**.

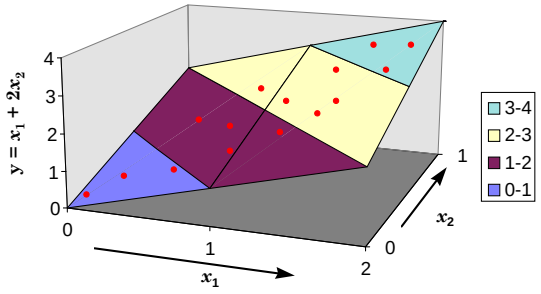
2. Reasons for Multiple Linear Regression

a) Adjusting for confounding variables

To investigate the effect of a variable on an outcome measure adjusted for the effects of other confounding variables.

i) β_1 estimates the rate of change of y_i with x_{i1} among patients with the same values of x_{i2}, x_{i3}, …, x_{ik}.

ii) If y_i increases rapidly with and x_{i1}, and x_{i1} and x_{i2} are highly correlated then the rate of increase of y_i with increasing x_{i1} when x_{i2} is held constant may be very different from this rate of increase when x_{i2} is not restrained.



NOTE: The model assumes that the rate of change of y_i with x_{i1} adjusted for x_{i1}, x_{i2}, …, x_{ik} is the same regardless of the values of these latter variables.

b) Prediction

To predict the value of y given x_1, x_2, …, x_k

3. Estimating Parameters

Let $\hat{y}_i = a + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}$ be the estimate of y_i given x_{i1}, x_{i2}, …, x_{ik}.

We estimate a, b_1, …, b_k by minimizing $\sum (y - \hat{y})^2$

4. Expected Response in the Multiple Model

The expected value of both y_i and $\hat{y}_i$ given her covariates is

$$\mathrm{E}[y_i \,|\, \mathbf{x}_i] = \mathrm{E}[\hat{y}_i \,|\, \mathbf{x}_i] = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_k x_{ik}.$$

We estimate the expected value of y_i among subjects whose covariate values are identical to those of the i^{th} patient by $\hat{y}_i$. The equation

$$\hat{y}_i = a + b_1 x_{i1} + b_2 x_{i2} + ... + b_k x_{ik}.$$

may be rewritten

$$\hat{y}_i = \overline{y}_i + b_1 (x_{i1} - \overline{x}_1) + b_2 (x_{i2} - \overline{x}_2) + ... + b_k (x_{ik} - \overline{x}_k). \qquad \{2.1\}$$

Thus, $\hat{y}_i = \overline{y}$ when $x_{i1} = \overline{x}_1$, $x_{i2} = \overline{x}_2$, ..., and $x_{ik} = \overline{x}_k$.

5. Framingham Example: SBP, Age, BMI, Sex and Serum Cholesterol

a) Preliminary univariate analysis

The Framingham data set contains data on 4,699 patients. On each patient we have the baseline values of the following variables:

sbp	Systolic blood pressure in mm Hg.
age	Age in years
scl	Serum cholesterol in mg/100ml
bmi	Body mass index in kg/m^2
sex	1 = Men 2 = Women

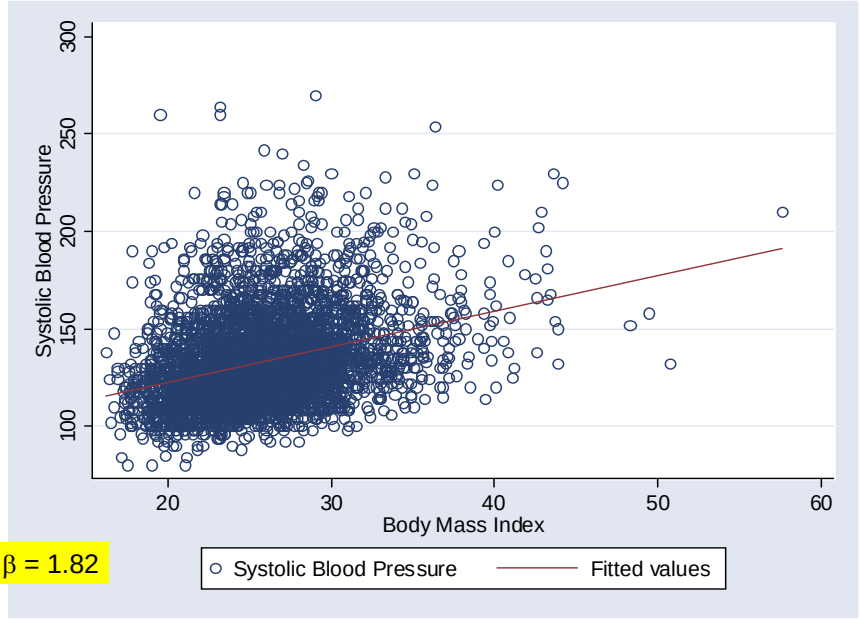
Follow-up information on coronary heart disease is also provided.

This data set is a subset of the 40 year data from the Framingham Heart Study that was conducted by the National Heart Lung and Blood Institute. Recruitment of patients started in 1948. At that time of the baseline exams there were no effective treatment for hypertension.

We first perform simple linear regressions of SBP on age, BMI, serum cholesterol.

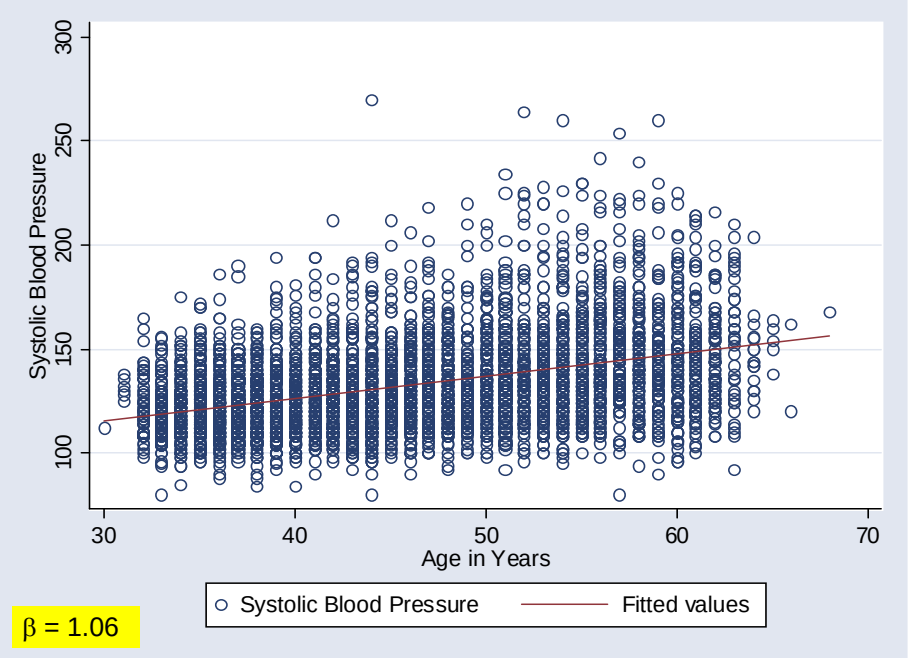
```
. * FramSBPbmiMulti.log
. *
. *  Framingham data set: Multiple regression analysis of the effect of bmi on
. *  sbp (Levy 1999).
. *
. use  "c:\WDDtext\2.20.Framingham.dta", clear
. regress sbp bmi
  Source |       SS       df       MS                  Number of obs =    4690
---------+------------------------------                F(  1,  4688) =  565.07
   Model |  262347.407     1  262347.407                Prob > F      =  0.0000
Residual |  2176529.37  4688  464.276742                R-squared     =  0.1076
---------+------------------------------                Adj R-squared =  0.1074
   Total |  2438876.78  4689  520.127271                Root MSE      =  21.547

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |    1.82675   .0768474     23.771   0.000       1.676093    1.977407
   _cons |   85.93592     1.9947     43.082   0.000       82.02537    89.84647
------------------------------------------------------------------------------
```

```
. scatter sbp  bmi, symbol(Oh)                                          ///
>    || lfit sbp bmi, ytitle(Systolic Blood Pressure)
```

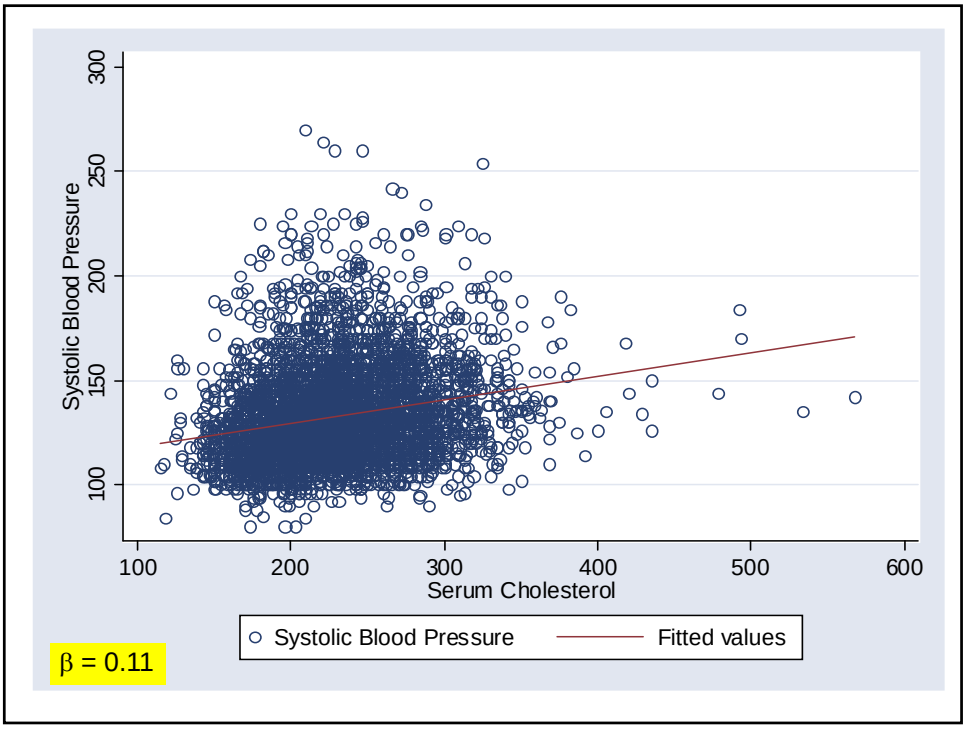


$\beta = 1.82$

```
. regress sbp age

      Source |       SS       df       MS              Number of obs =    4699
-------------+------------------------------           F(  1,  4697) =  865.99
       Model |  380213.315     1  380213.315            Prob > F      =  0.0000
    Residual |  2062231.59  4697  439.052924           R-squared     =  0.1557
-------------+------------------------------           Adj R-squared =  0.1555
       Total |  2442444.90  4698  519.890358           Root MSE      =  20.954

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.057829   .0359468     29.428   0.000       .9873561    1.128301
       _cons |   84.06298    1.68302     49.948   0.000       80.76347    87.36249
------------------------------------------------------------------------------


. scatter sbp  age, symbol(Oh)                                                ///
>     || lfit sbp age, ytitle(Systolic Blood Pressure)
```



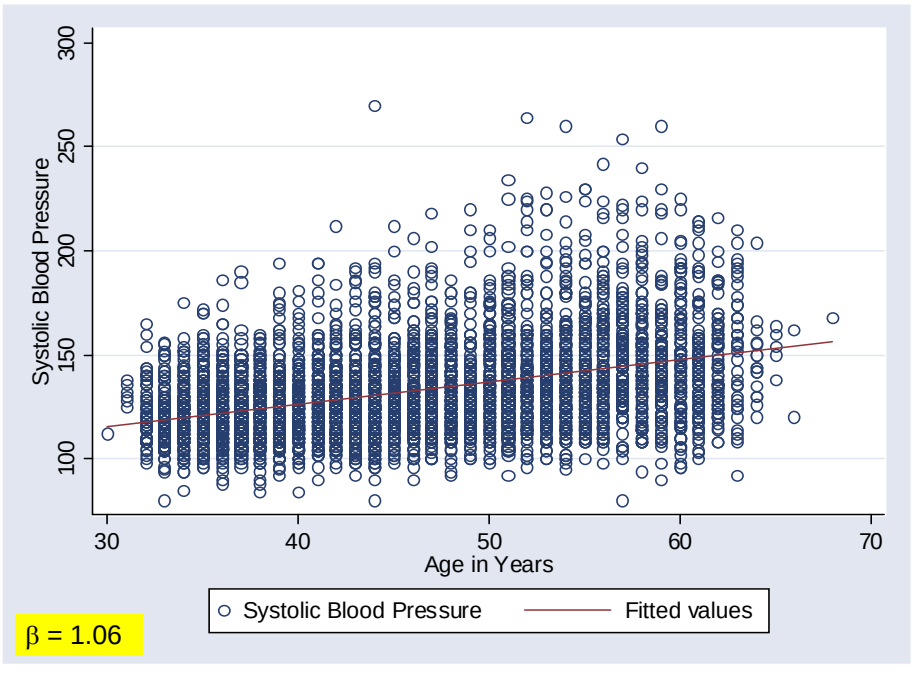
```
. regress sbp scl

      Source |       SS       df       MS              Number of obs =    4666
-------------+------------------------------           F(  1,  4664) =  231.52
       Model |  114616.314     1  114616.314            Prob > F      =  0.0000
    Residual |  2308993.33  4664   495.06718            R-squared     =  0.0473
-------------+------------------------------           Adj R-squared =  0.0471
       Total |  2423609.64  4665   519.53047            Root MSE      =   22.25

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
         scl |   .1112811   .0073136     15.216   0.000       .0969431    .1256192
       _cons |    107.378   1.701114     63.122   0.000        104.043    110.713
------------------------------------------------------------------------------

. scatter sbp  scl, symbol(Oh)                                              ///
>     || lfit sbp scl, ytitle(Systolic Blood Pressure)
```

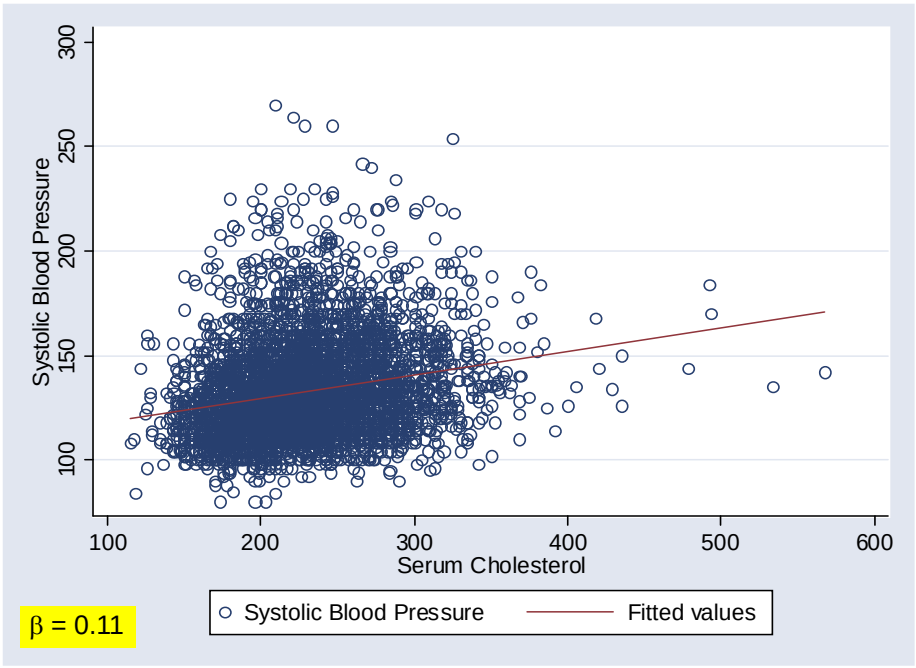


The univariate regressions show that *sbp* is related to *age* and *scl* as well as *bmi*. Although the statistical significance of the slope coefficients is overwhelming, the R-squared statistics are low. Hence, each of these risk factors individually only explain a modest proportion of the total variability in systolic blood pressure.

We would like better understanding of these relationships.

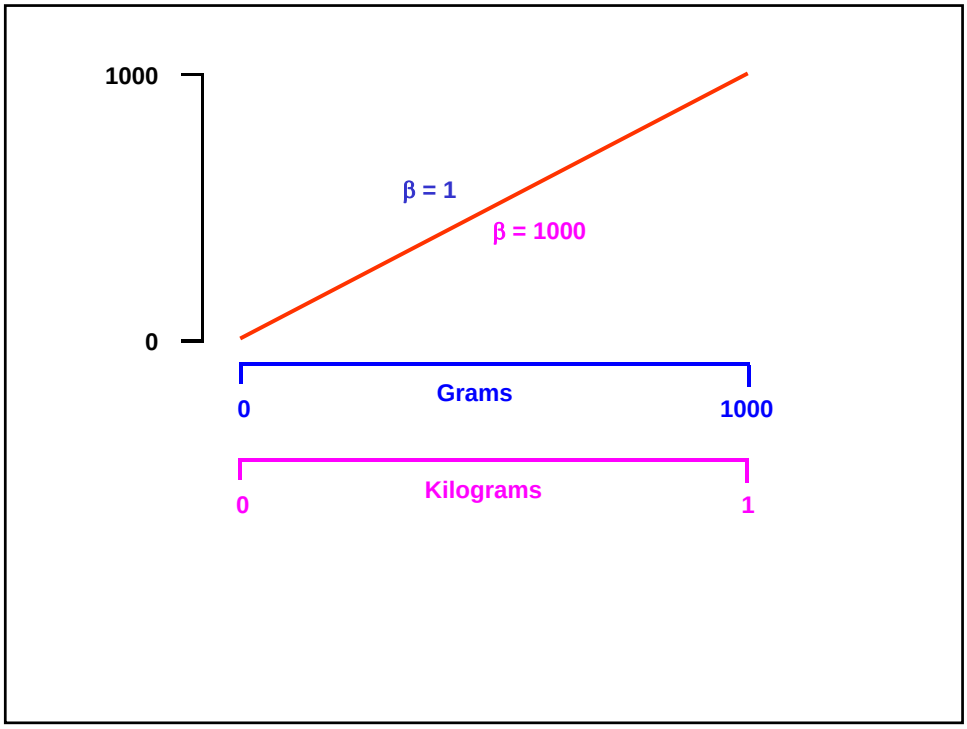
Note that the importance of a parameter depends not only on its magnitude but also on the range of the corresponding covariate. For example, the *scl* coefficient is only 0.11 as compared to 1.83 and 1.06 for *bmi* and *age*. However, the range of *scl* values is from 115 to 568 as compared to 16.2 - 57.6 for *bmi* and 30 - 68 for *age*. The large *scl* range increases the variation in *sbp* that is associated with *scl*.





Changing the units of measurement of a covariate can have a dramatic effect on the size of the slope estimate, but no effect on its biologic meaning.

For example, suppose we regressed blood pressure against weight in grams. If we converted weight from grams to kilograms we would increase the magnitude of the slope parameter by 1,000 but would have no effect on the true relationship between blood pressure and weight.



6. Density Distribution Sunflower Plots

Scatterplots are a simple but informative tool for displaying the relationship between two variables. Their utility decreases when the density of observations makes it difficult to see individual observations.

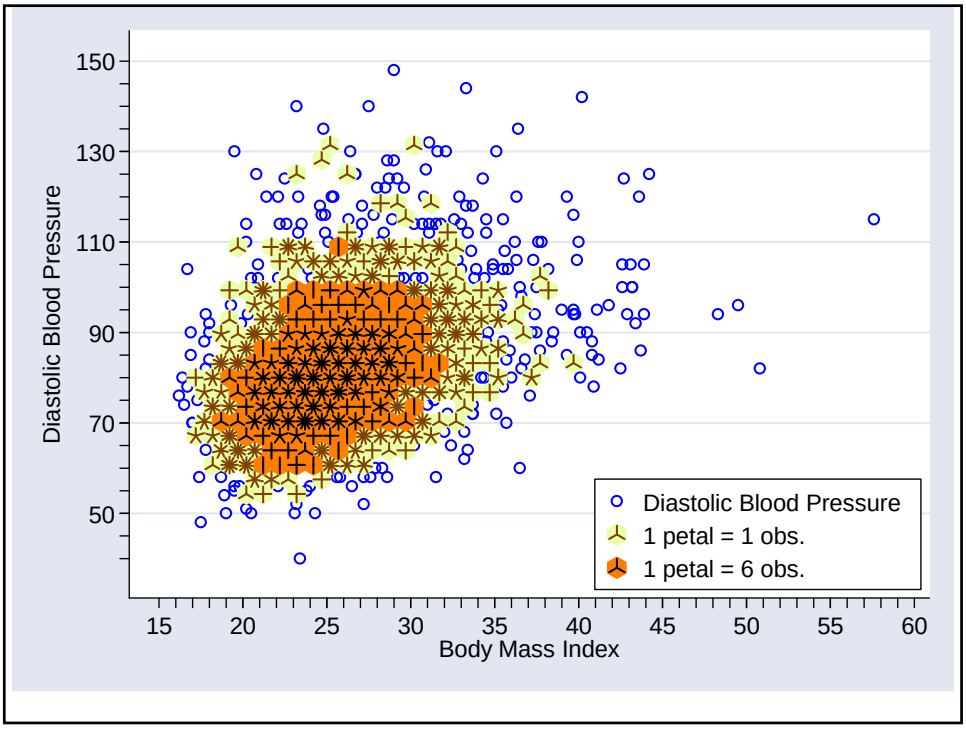
A **density distribution sunflower plot** is an attempt to provide a better sense of a bivariated distribution when observations are densely packed.

Data points are represented in one of three ways depending on the density of observations.

1) **Low Density:**
   Small circles representing individual data points as in a conventional scatterplot.

2) **Medium Density:**
   light sunflowers.

3) **High Density:**
   dark sunflowers.



A sunflower is a number of short line segments radiating from a central point.

In a light sunflower each petal represents one observation.

In a dark sunflower, each petal represents k observations, where k is specified by the user.

The x-y plane is divided into a lattice of hexagonal bins.

The user can control the bin width in the units of the x-axis and thresholds l and d that determine when light and dark sunflowers are drawn.

Whenever there are less than l data points in a bin the individual data points are depicted at their exact location.

When there are at least l but fewer than d data points in a bin they are depicted by a light sunflower.

When there are at least d observations in a bin they are depicted by a dark sunflower.

For more details see the Stata v8.2 online documentation on the sunflower command.

7. Creating Density Distribution Plots with Stata

```
. *  FramSunflower.log
. *
. *  Framingham data set: Exploratory analysis of sbp and bmi
. *
. set more on

. use "c:\WDDtext\2.20.Framingham.dta", clear

. * Graphics > Smoothing ... > Density-distribution sunflower plot
. sunflower sbp bmi                                          {1}
Bin width          =      1.15                               {2}
Bin height         =   11.8892                               {3}
Bin aspect ratio   =   8.95333
Max obs in a bin   =       115
Light              =         3                               {4}
Dark               =        13                               {5}
X-center           =      25.2
Y-center           =       130
Petal weight       =         9                               {6}
```

{1} Create a sunflower plot of *sbp* by *bmi*. Let the program choose all default values. The resulting graph is given in the next slide.

{2} The default bin width is given in units of x. It is chosen to provide 40 bins across the graph.

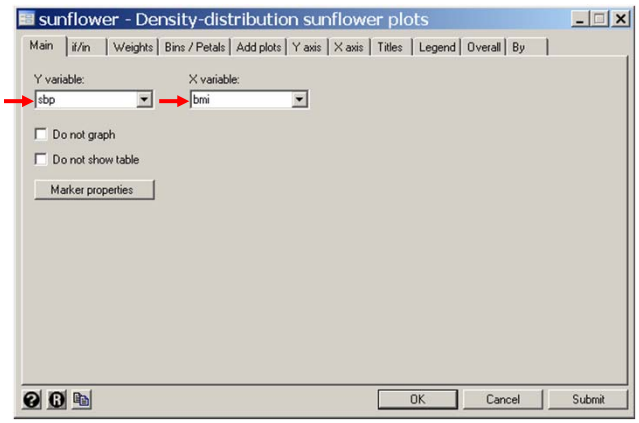
{3} The default bin height is given in units of y. It is chosen to make the bins regular hexagons on the graph.

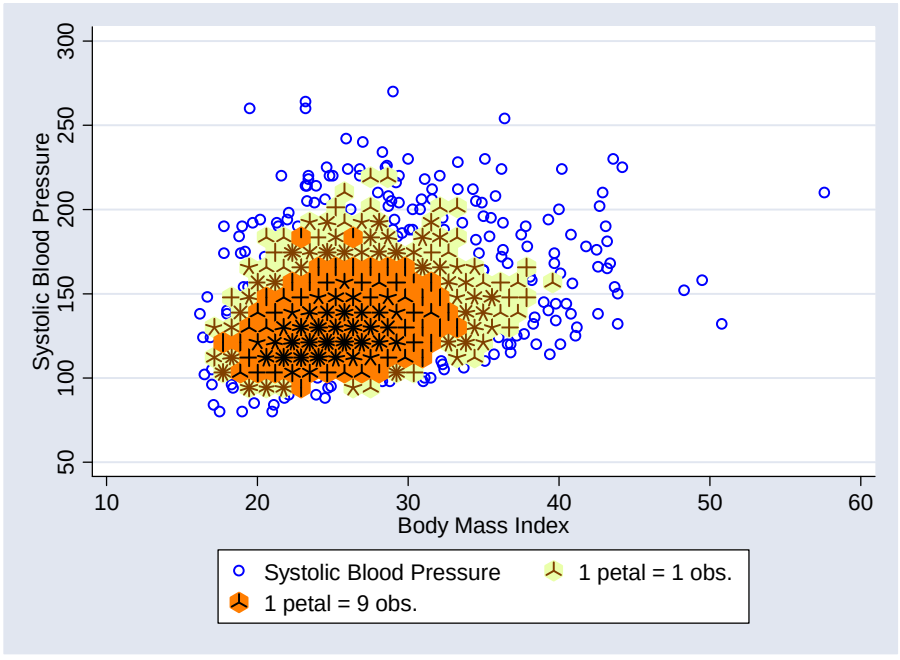
{4} The default minimum number of observations in a light sunflower bin is 3

{5} The default minimum number of observations in a dark sunflower bin is 13

{6} The default petal weight for dark sunflowers is chosen so that the maximum number of petals in a dark sunflower is 14.

flower type	petal weight	No. of petals	No. of flowers	estimated obs.	actual obs.
none				171	171
light	1	3	20	60	60
light	1	4	11	44	44
light	1	5	11	55	55
light	1	6	8	48	48
light	1	7	9	63	63
light	1	8	5	40	40
light	1	9	7	63	63
light	1	10	4	40	40
light	1	11	3	33	33
light	1	12	4	48	48
dark	9	1	4	36	52
dark	9	2	21	378	381
dark	9	3	11	297	285
dark	9	4	14	504	497
dark	9	5	7	315	322
dark	9	6	4	216	214
dark	9	7	5	315	314
dark	9	8	4	288	296
dark	9	9	5	405	410
dark	9	10	3	270	269
dark	9	11	2	198	197
dark	9	12	4	432	445
dark	9	13	3	351	343
				4670	4690


sunflower - Density-distribution sunflower plots
Main
if/in
Weights
Bins / Petals
Add plots
Y axis
X axis
Titles
Legend
Overall
By
Y variable:
sbp
X variable:
bmi
Do not graph
Do not show table
Marker properties
OK
Cancel
Submit

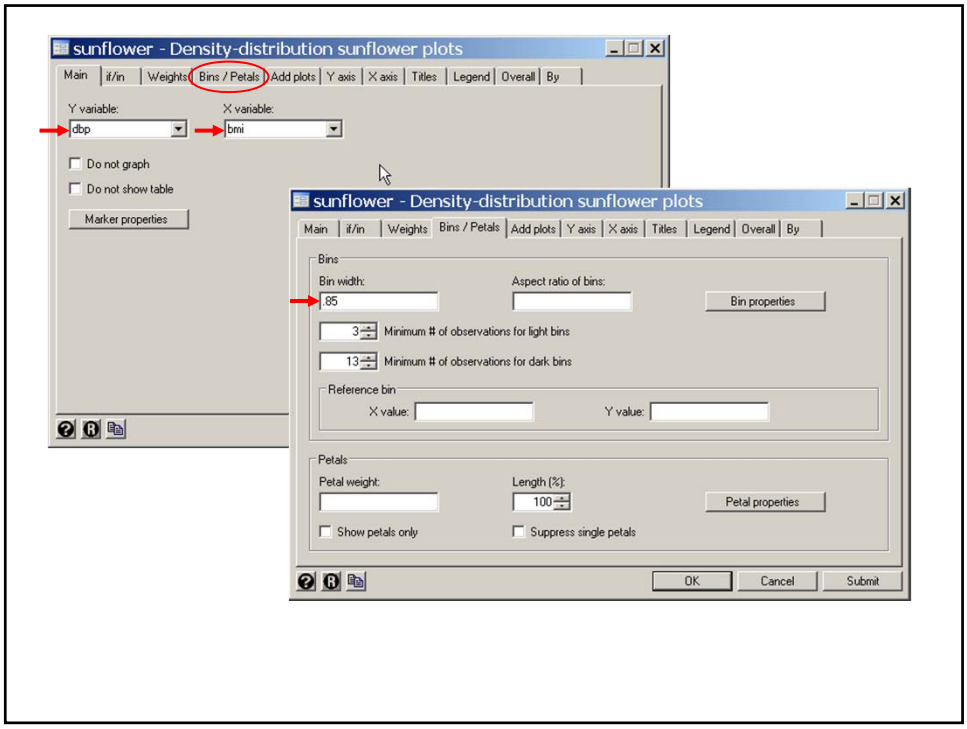

Systolic Blood Pressure
50
100
150
200
250
300
10
20
30
40
50
60
Body Mass Index
Systolic Blood Pressure
1 petal = 1 obs.
1 petal = 9 obs.

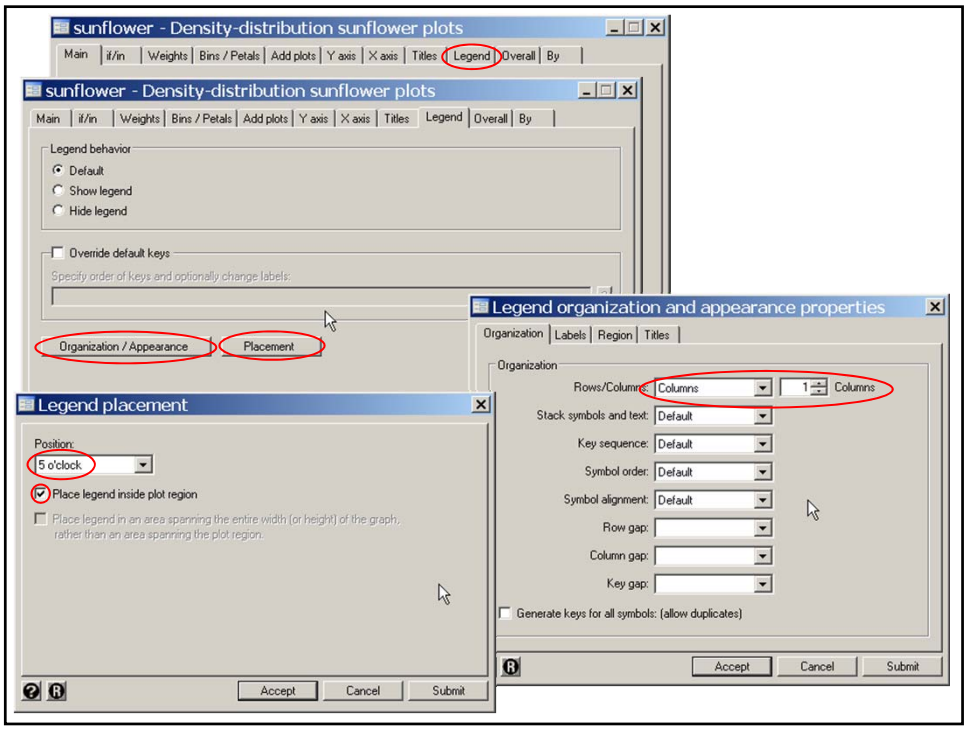
```
. more

. * Graphics > Smoothing ... > Density-distribution sunflower plot
. sunflower dbp bmi, binwidth(0.85)                                  /// {1}
>         ylabel(50 (20) 150, angle(0)) ytick(40 (5) 145)      ///
>         xlabel(20 (5) 55) xtick(16 (1) 58)                   ///
>         legend(position(5) ring(0) cols(1))                  /// {2}
>         addplot(lfit dbp bmi,  color(green)                  /// {3}
>           || lowess dbp bmi , bwidth(.2) color(cyan) )
Bin width          =       .85
Bin height         =   3.66924
Bin aspect ratio   =   3.73842
Max obs in a bin   =        59
Light              =         3
Dark               =        13
X-center           =      25.2
Y-center           =        80
Petal weight       =         5
```

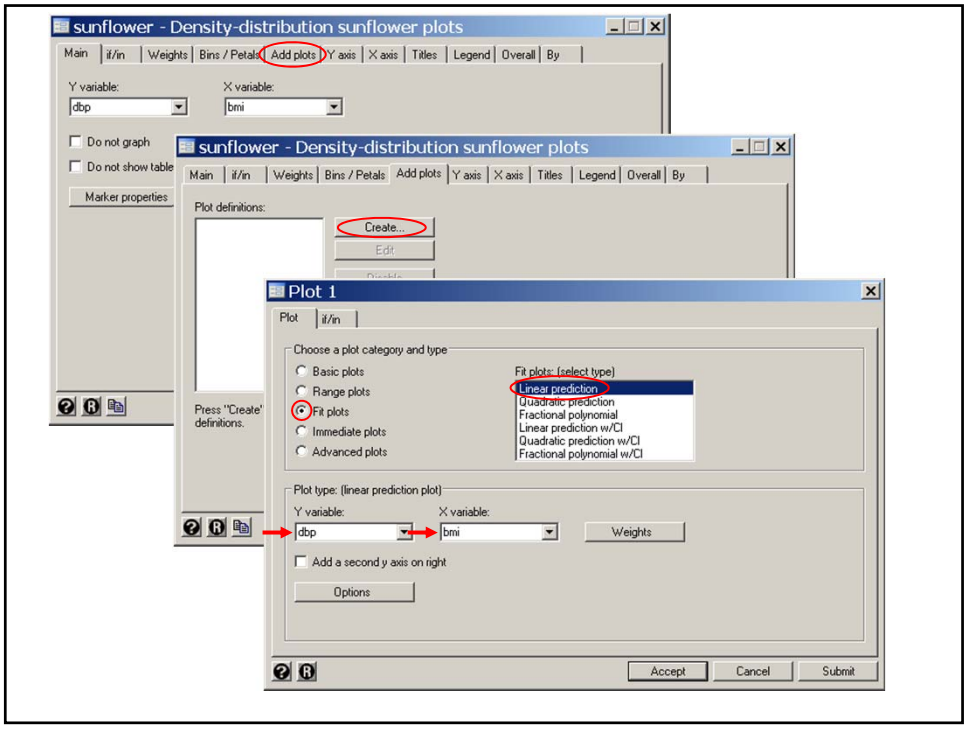
{1} *sunflower* accepts most standard graph options as well as special options that can control almost all aspects of the plot. Here *binwidth* specifies the bin width to be 0.85 kg/m^2.

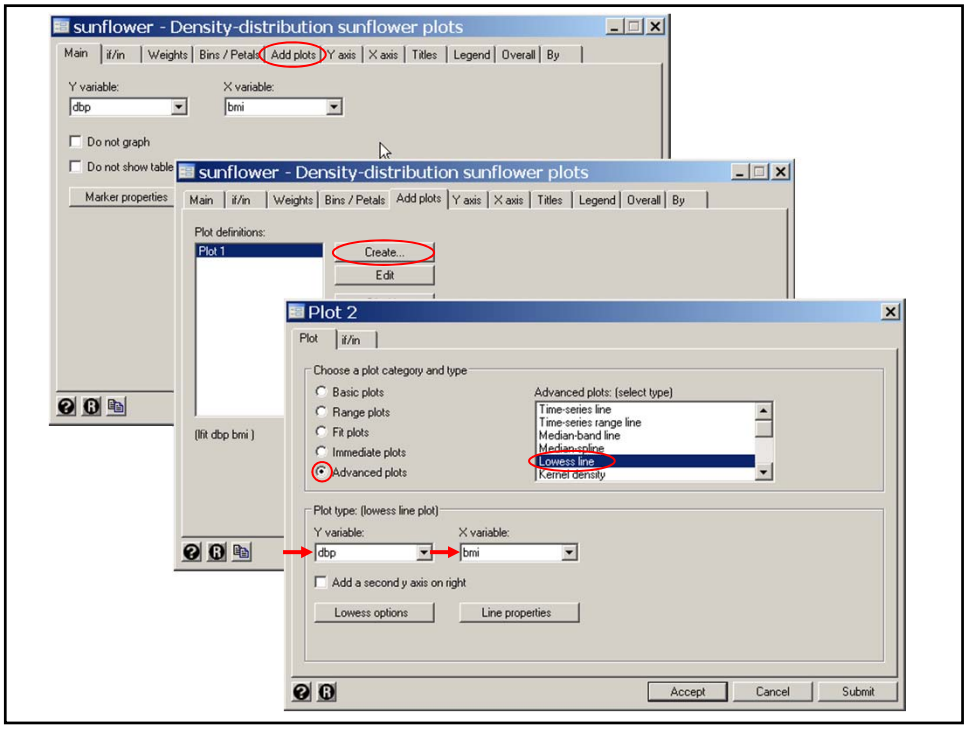
{2} The *position* sub-option of the legend option specifies that the legend will be located at 5 o'clock. *ring(0)* causes the legend to be drawn within the graph region. *cols(1)* requires that the legend keys be in a single column.

{3} The *addplot* option allows us to overlay other graphs on top of the sunflower plot. Here we draw the linear regression and lowess regression curves.


sunflower - Density-distribution sunflower plots
Main | if/in | Weights | Bins / Petals | Add plots | Y axis | X axis | Titles | Legend | Overall | By
Y variable:
dbp
X variable:
bmi
Do not graph
Do not show table
Marker properties
sunflower - Density-distribution sunflower plots
Bins
Bin width:
.85
Aspect ratio of bins:
Bin properties
3
Minimum # of observations for light bins
13
Minimum # of observations for dark bins
Reference bin
X value:
Y value:
Petals
Petal weight:
Length (%):
100
Petal properties
Show petals only
Suppress single petals
OK
Cancel
Submit


sunflower - Density-distribution sunflower plots
Legend behavior
Default
Show legend
Hide legend
Override default keys
Specify order of keys and optionally change labels:
Organization / Appearance
Placement
Legend placement
Position:
5 o'clock
Place legend inside plot region
Place legend in an area spanning the entire width (or height) of the graph, rather than an area spanning the plot region.
Accept
Cancel
Submit
Legend organization and appearance properties
Organization | Labels | Region | Titles
Organization
Rows/Columns:
Columns
1
Columns
Stack symbols and text:
Default
Key sequence:
Symbol order:
Symbol alignment:
Row gap:
Column gap:
Key gap:
Generate keys for all symbols: (allow duplicates)
Accept
Cancel
Submit


sunflower - Density-distribution sunflower plots
Main | if/in | Weights | Bins / Petals | Add plots | Y axis | X axis | Titles | Legend | Overall | By
Y variable: dbp
X variable: bmi
Do not graph
Do not show table
Marker properties
Plot definitions:
Create...
Edit
Press "Create" definitions.
Plot 1
Plot | if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Fit plots: (select type)
Linear prediction
Quadratic prediction
Fractional polynomial
Linear prediction w/CI
Quadratic prediction w/CI
Fractional polynomial w/CI
Plot type: (linear prediction plot)
Y variable: dbp
X variable: bmi
Weights
Add a second y axis on right
Options
Accept
Cancel
Submit


sunflower - Density-distribution sunflower plots
Main | if/in | Weights | Bins / Petals | Add plots | Y axis | X axis | Titles | Legend | Overall | By
Y variable: dbp
X variable: bmi
Do not graph
Do not show table
Marker properties
Plot definitions:
Plot 1
Create...
Edit
(fit dbp bmi)
Plot 2
Plot | if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Advanced plots: (select type)
Time-series line
Time-series range line
Median-band line
Median-spline
Lowess line
Kernel density
Plot type: (lowess line plot)
Y variable: dbp
X variable: bmi
Add a second y axis on right
Lowess options
Line properties
Accept
Cancel
Submit

8. Scatterplot matrix graphs

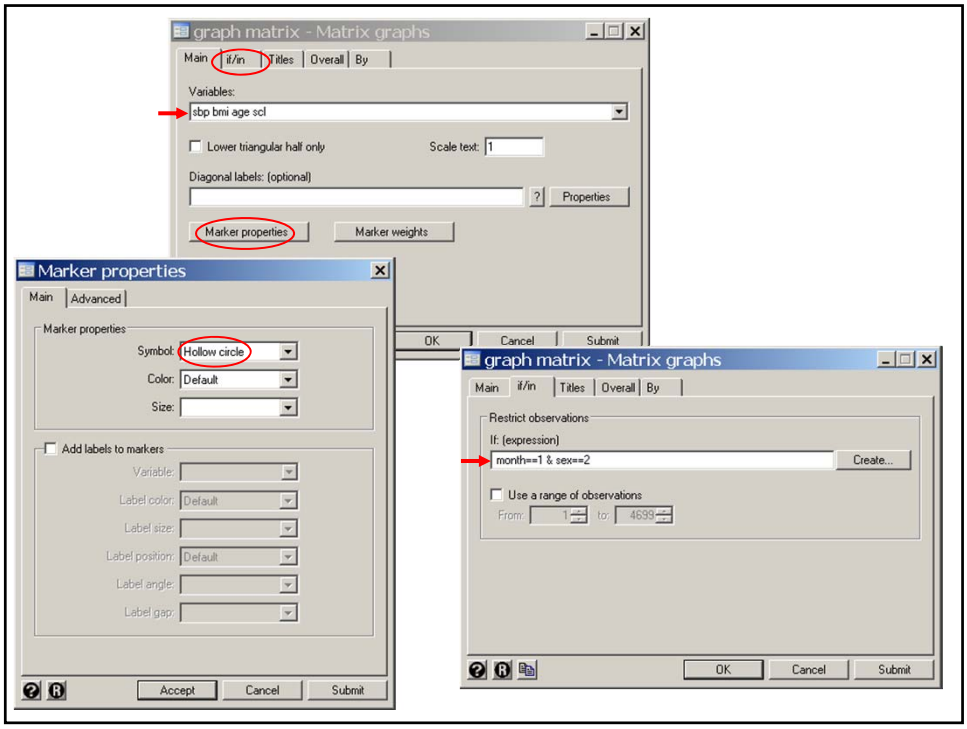
Another useful exploratory graphic is the scatter plot matrix. Here we look at the combined marginal effects of *sbp age bmi* and *scl*. The graph is restricted to women recruited in January to reduce the number of data points.

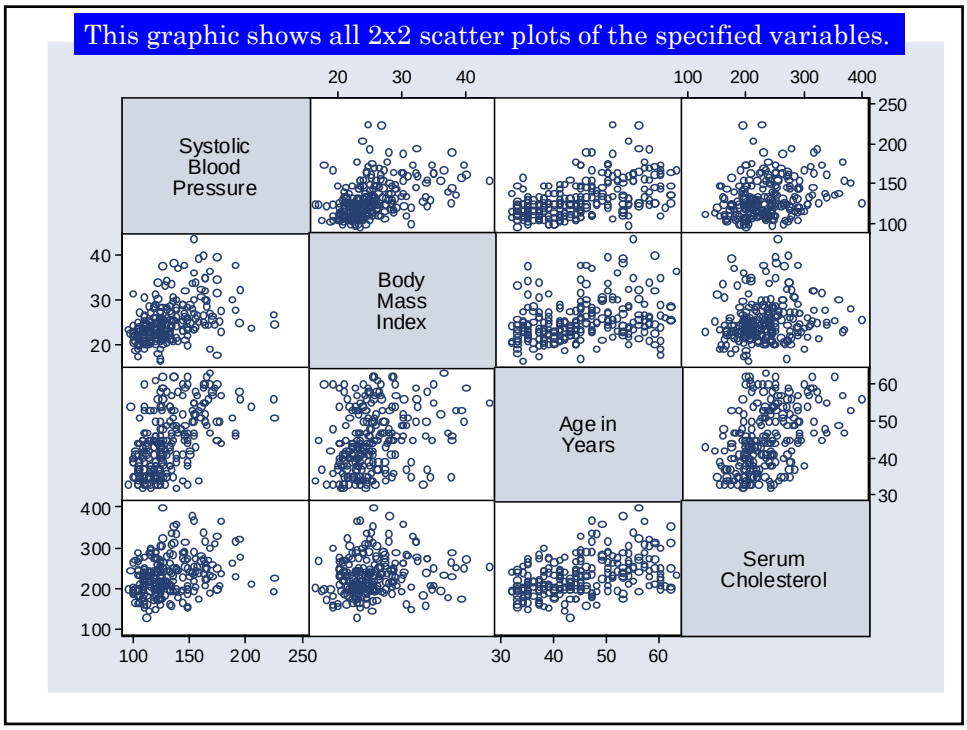
FramSBPbmiMulti.log continues as follows

```
. * Graphics > Scatterplot matrix
. graph matrix sbp bmi age scl if month==1 & sex==2 ,msymbol(oh)     {1}
```

{1} The **matrix** option generates a matrix scatter plot for *sbp bmi age* and *scl*. The *if* clause restricts the graph to women (sex==2) who entered the study in January (month==1).

oh specifies a small hollow circle as a plot symbol


graph matrix - Matrix graphs
Main
if/in
Titles
Overall
By
Variables:
sbp bmi age scl
Lower triangular half only
Scale text: 1
Diagonal labels: (optional)
Properties
Marker properties
Marker weights
Marker properties
Main
Advanced
Marker properties
Symbol: Hollow circle
Color: Default
Size:
Add labels to markers
Variable:
Label color: Default
Label size:
Label position: Default
Label angle:
Label gap:
Accept
Cancel
Submit
OK
Cancel
Submit
graph matrix - Matrix graphs
Main
if/in
Titles
Overall
By
Restrict observations
If: (expression)
month==1 & sex==2
Create...
Use a range of observations
From: 1
to: 4699
OK
Cancel
Submit


This graphic shows all 2x2 scatter plots of the specified variables.
Systolic Blood Pressure
Body Mass Index
Age in Years
Serum Cholesterol
20
30
40
100
200
300
400
250
150
60
50

9. Modeling interaction in the Framingham baseline data

The first model that comes to mind is

$$E[sbp_i \,|\, \mathbf{x}_i] = \alpha + \beta_1 \times bmi_i + \beta_2 \times age_i + \beta_3 \times scl_i + \beta_4 \times sex_i.$$

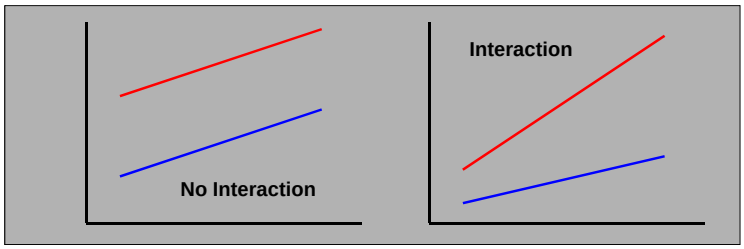
A potential weakness of this model is that it implies that the effects of the covariates on sbp_i are additive. To understand what this means, suppose we hold *age* and *scl* constant and look at *bmi* and *sex*. Then the model becomes

sbp = constant + bmi x β_1 + β_4 for men, and

sbp = constant + bmi x β_1 + $2\beta_4$ for women.

The β_4 parameter allows men and women with the same *bmi* to have different expected *sbps*.

However, the slope of the *sbp*-*bmi* relationship for both men and women is β_1.



We know, however, that this slope is higher for women than for men. This is an example of what we call interaction in which the effect of one variate on the dependent variable is influenced by the value of a second covariate.

We need a more complex model to deal with interaction.

Let *women* = *sex* - 1.

Then *women* = { 1: if subject is female
0: if subject is male

Consider the model

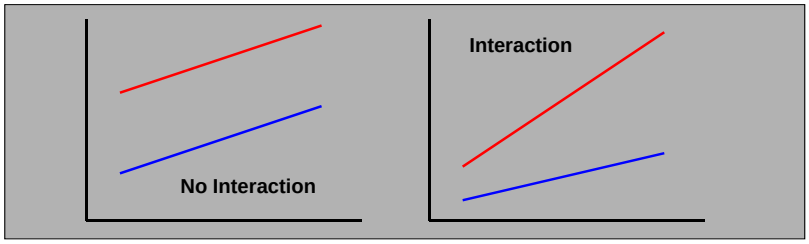
$sbp = \beta_1 + bmi \times \beta_2 + women \times \beta_3 + bmi \times women \times \beta_4$

This model reduces to

$sbp = \beta_1 + bmi \times \beta_2$ for men and

$sbp = \beta_1 + bmi \times (\beta_2 + \beta_4) + \beta_3$ for women.

Hence β_4 estimates the difference in slopes between men and women.



We use this approach to build an appropriate multivariate model for the Framingham data.

FramSBPbmiMulti.log continues as follows.

```
. *
. *  Use multiple regression models with interaction terms to analyze
. *  the effects of sbp, bmi, age and scl on sbp.
. *
. generate woman = sex - 1
. label define truth 0 "False" 1 "True"
. label values woman truth
. generate bmiwoman = bmi*woman
(9 missing values generated)
. generate agewoman = age*woman
. generate sclwoman = woman * scl
(33 missing values generated)
```

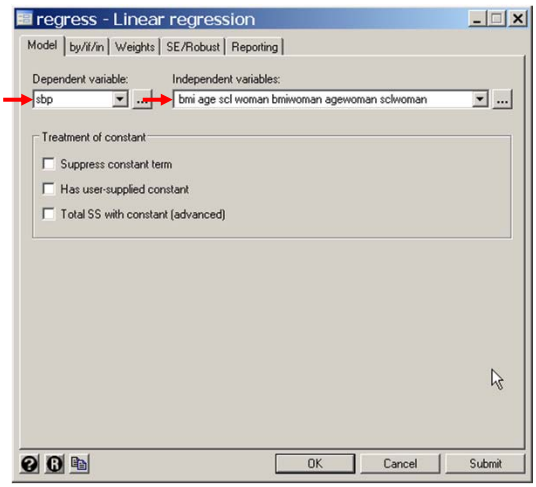
```
. regress sbp bmi age scl woman bmiwoman agewoman sclwoman

      Source |       SS       df       MS              Number of obs =    4658
-------------+------------------------------           F(  7,  4650) =  217.41
       Model |  596743.008     7  85249.0011            Prob > F      =  0.0000
    Residual |  1823322.50  4650  392.112365            R-squared     =  0.2466 {1}
-------------+------------------------------           Adj R-squared =  0.2454
       Total |  2420065.50  4657  519.661908            Root MSE      =  19.802

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.       t     P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |    1.260872    .130925     9.630   0.000      1.004197    1.517547
         age |    .5170311   .0518617     9.969   0.000      .4153576    .6187047
         scl |    .0376262   .0105242     3.575   0.000      .0169938    .0582586
       woman |   -31.06614    5.29534    -5.867   0.000     -41.44751   -20.68476
   bmiwoman |     .141898   .1582655     0.897   0.370     -.1683775    .4521735
   agewoman |    .6658219   .0734669     9.063   0.000      .5217919    .8098519
   sclwoman |   -.0078668    .014045    -0.560   0.575     -.0354017    .0196682 {2}
      _cons |    67.22324   4.427304    15.184   0.000      58.54362    75.90285
------------------------------------------------------------------------------
```

{1} **R-squared** equals the square of the correlation coefficient between $\hat{y}_i$ and y_i. It still equals $\sum(\hat{y}_i - \bar{y})^2 / \sum(y_i - \bar{y})^2$

and hence can be interpreted as the proportion of the **variation** in y **explained** by the **model**.

In the simple regression of *sbp* and *bmi* we had **R-squared = 0.11**. Thus, this multiple regression model explains more than twice the variation in *sbp* than did the simple model.

{2} The serum cholesterol-woman **interaction** coefficient, -0.0079, is five times **smaller** than the ***scl*** coefficient, and is not statistically significant. Lets drop it from the model and see what happens.




```
. regress sbp bmi age scl woman bmiwoman agewoman

  Source |       SS       df       MS              Number of obs =    4658
---------+------------------------------           F(  6,  4651) =  253.63
   Model |  596619.993     6  99436.6655           Prob > F      =  0.0000
Residual |  1823445.51  4651  392.054507           R-squared     =  0.2465 {3}
---------+------------------------------           Adj R-squared =  0.2456
   Total |  2420065.50  4657  519.661908           Root MSE      =   19.80

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   1.269339   .1300398      9.761   0.000       1.014399    1.524278
     age |   .5182974   .0518086     10.004   0.000        .416728    .6198668
     scl |   .0332092   .0069687      4.765   0.000       .0195472    .0468712
   woman |  -32.18538   4.903474     -6.564   0.000      -41.79851   -22.57224
bmiwoman |   .1323904    .157341      0.841   0.400      -.1760726    .4408534 {4}
agewoman |    .656538   .0715675      9.174   0.000       .5162319    .7968442
   _cons |   67.94892   4.233177     16.052   0.000       59.64988    76.24795
------------------------------------------------------------------------------
```

{3} Dropping the ***sclwoman*** term has a **trivial effect** on the R-squared statistic and little effect on the model coefficients.

{4} The *bmiwoman* **interaction** term is also not significant and is an order of magnitude **smaller** than the ***bmi*** term. Lets drop it.

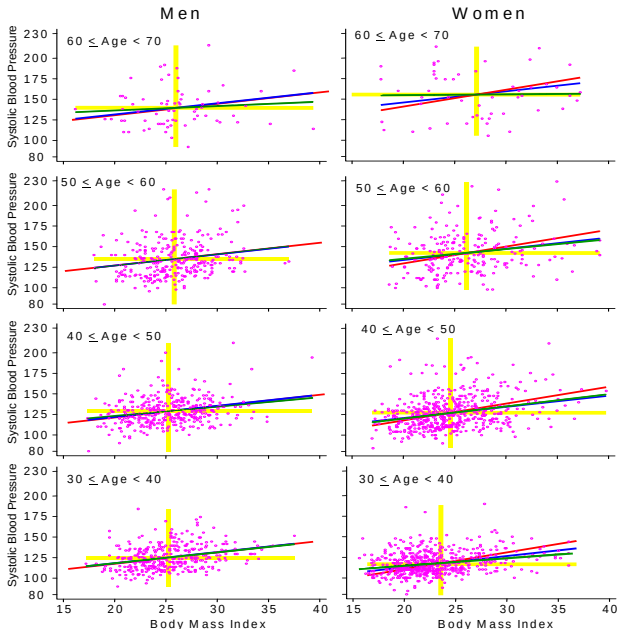
```
. regress sbp bmi age scl woman agewoman

  Source |       SS       df       MS              Number of obs =    4658
---------+------------------------------           F(  5,  4652) =  304.23
   Model |  596342.421     5  119268.484           Prob > F      =  0.0000
Residual |  1823723.08  4652  392.029897           R-squared     =  0.2464 {5}
---------+------------------------------           Adj R-squared =  0.2456
   Total |  2420065.50  4657  519.661908           Root MSE      =   19.80

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   1.359621   .0734663     18.507   0.000       1.215592     1.50365
     age |   .5173521   .0517948      9.988   0.000       .4158098    .6188944
     scl |   .0327898   .0069506      4.718   0.000       .0191632    .0464163
   woman |  -29.14655   3.316662     -8.788   0.000      -35.64878   -22.64432
agewoman |   .6646316   .0709159      9.372   0.000       .5256029    .8036603
   _cons |   65.74423   3.324712     19.774   0.000       59.22622    72.26224
------------------------------------------------------------------------------
```

{5} Dropping the preceding term reduces the R^2 value by **0.04%**. The remaining terms are highly significant.

When we did simple linear regression of *sbp* against *bmi* for men and women we obtained slope estimates of 1.38 and 2.05 for men and women, respectively.

Our multivariate model gives a single slope estimate of 1.36 for both sexes, but finds that the effect of increasing age on *sbp* is twice as large in women than men. I.e. For women this slope is 0.52 + 0.66 = 1.18 while for men it is 0.52.

How reasonable is our model? One way to increase our intuitive understanding of the model is to plot separate simple linear regressions of *sbp* against *bmi* in groups of patients who are homogeneous with respect to the other variables in the model. The following graphic is restricted to patients with a serum cholesterol of $\leq$ 225 and subdivides patients by age and sex. In these graphs, two versions of the graph are given drawn to different scales. The second only shows the regression lines.

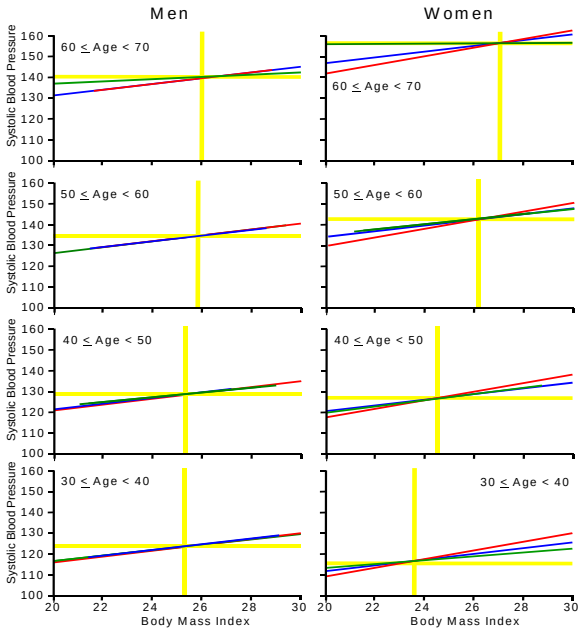


The **blue** lines have the slope from our **multiple regression model** of 1.36

The **red** lines have slopes 1.38 for men and 2.05 for women (the slopes of the **simple regressions** in men and women respectively.

The **green** lines have the slope of the **simple regression** for patients with the indicated **age** and gender.

The **yellow** lines mark the **mean *sbp*** and ***bmi*** for the indicated age-gender group.



For **men** the adjusted and unadjusted **slopes** are almost **identical** and are very close to the age restricted slope for all ages except 60 - 70.

However, for **women** the adjusted and unadjusted **slopes differ** appreciably. The adjusted slope is very close to the age restricted slopes in every case except age 60 - 70, where the adjusted slope is closed the age restricted slope than is the unadjusted slope.

Thus, our model is a marked improvement over the simple model. The **single** *sbp-bmi* adjusted **slope** estimate appears **reasonable** except, for the oldest subjects.

Note that the mean *sbp* increases with age for both sexes, but increases more **rapidly** in **women** than in **men**.

The mean *bmi* does not vary appreciably with age in men but does increase with increasing age in women.

Thus **age** and **gender confound** the effect of *bmi* on *sbp*. Do you think that the age-gender interaction of *sbp* is real or is this driven by some other unknown confounding variable?

10. Automatic Methods of Model Selection

Analyses loose power when we include variables in the model that are neither confounders nor variables of interest. When a large number of potential confounders are available it can be useful to use an automatic model selection program.

a) Forward Selection

i) Fit all simple linear models of *y* against each separate *x* variable. Select the variable with the greatest significance.

ii) Fit all possible models with the variable(s) selected in the preceding step(s) and one other. Select as the next variable the one with the greatest significance among these models.

iii) repeat step ii) to add additional variables, one variable at a time. Continue this process until none of the remaining variables have a significance level less than some threshold.

We next illustrate how this is done in Stata.

FramSBPbmiMulti.log continues as follows.

```
. *
. *  Fit a model of sbp against bmi age scl and sex with
. *  interaction terms. The variables woman, bmiwoman,
. *  agewoman, and sclwoman have been previously defined.
. *
. * statistics > other > stepwise estimation
. stepwise, pe(.1): regress sbp bmi age scl woman bmiwoman agewoman sclwoman
                                                                          {1}
                      begin with empty model
p = 0.0000 <  0.1000  adding   age                                        {2}
p = 0.0000 <  0.1000  adding   bmi                                        {3}
p = 0.0000 <  0.1000  adding   scl
p = 0.0001 <  0.1000  adding   agewoman
p = 0.0000 <  0.1000  adding   woman
      Source |       SS       df       MS              Number of obs =    4658
-------------+------------------------------           F(  5,  4652) =  304.23
       Model |  596342.421     5  119268.484           Prob > F      =  0.0000
    Residual |  1823723.08  4652  392.029897           R-squared     =  0.2464
-------------+------------------------------           Adj R-squared =  0.2456
       Total |   2420065.5  4657  519.661908           Root MSE      =    19.8

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .5173521   .0517948     9.99   0.000     .4158098    .6188944
         bmi |   1.359621   .0734663    18.51   0.000     1.215592     1.50365
         scl |   .0327898   .0069506     4.72   0.000     .0191632    .0464163
    agewoman |   .6646316   .0709159     9.37   0.000     .5256029    .8036603
       woman |  -29.14655   3.316662    -8.79   0.000    -35.64878   -22.64432
       _cons |   65.74423   3.324712    19.77   0.000     59.22622    72.26224
------------------------------------------------------------------------------
```

{1} Fit a model using forward selection; ***pe*(.1)** means that the ***P* value** for entry is **0.1**. At each step new variables will only be considered for entry into the model if their *P* value after adjustment for previously entered variables is <0.1.

{2} In the first step the program considers the following models.

$sbp = \beta_1 + bmi \times \beta_2$

$sbp = \beta_1 + age \times \beta_2$

$sbp = \beta_1 + scl \times \beta_2$

$sbp = \beta_1 + woman \times \beta_2$

$sbp = \beta_1 + bmiwoman \times \beta_2$

$sbp = \beta_1 + agewoman \times \beta_2$

$sbp = \beta_1 + sclwoman \times \beta_2$

Of these models the one with **age** has the most **significant** slope parameter. The *P* value associated with this parameter is **<0.1**. Therefore we select *age* and go on to step 2.

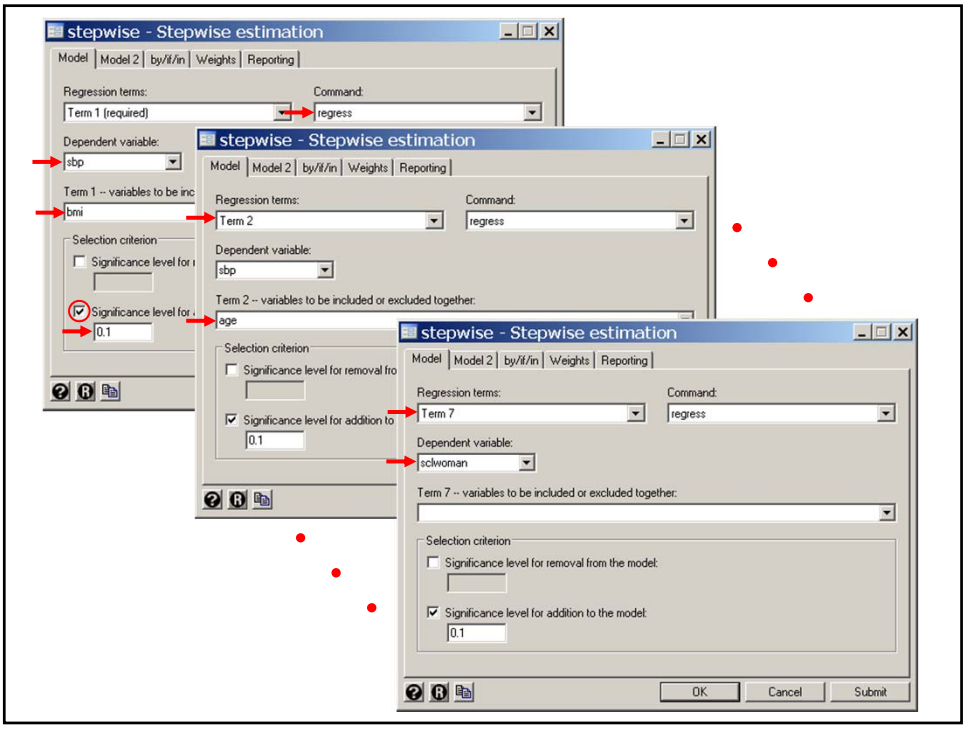
{3} In step 2 we consider the models

$sbp = \beta_1 + age \times \beta_2 + bmi \times \beta_3$

$sbp = \beta_1 + age \times \beta_2 + scl \times \beta_3$

•
•

$sbp = \beta_1 + age \times \beta_2 + sclwoman \times \beta_3$



The most significant new term in these models is *bmi*, which is selected. This process is continued until at the end of step 5 we have the model

$$sbp = \beta_1 + age \times \beta_2 + bmi \times \beta_3 + scl \times \beta_4 + agewoman \times \beta_5 + woman \times \beta_6$$

In step 6 we consider the models

$$sbp = \beta_1 + age \times \beta_2 + bmi \times \beta_3 + scl \times \beta_4 + agewoman \times \beta_5 + woman \times \beta_6 + bmiwoman \times \beta_7$$

and

$$sbp = \beta_1 + age \times \beta_2 + bmi \times \beta_3 + scl \times \beta_4 + agewoman \times \beta_5 + woman \times \beta_6 + sclwoman \times \beta_7$$

However, neither of the P values for the β_7 parameter estimates in these models are < 0.1. Therefore, neither of these terms are added to the model.

```
. *
. *  Fit a model of sbp against bmi age scl and sex with
. *  interaction terms. The variables woman, bmiwoman,
. *  agewoman, and sclwoman have been previously defined.
. *
. * statistics > other > stepwise estimation
. stepwise, pe(.1): regress sbp bmi age scl woman bmiwoman agewoman sclwoman

                      begin with empty model
p = 0.0000 <  0.1000  adding   age
p = 0.0000 <  0.1000  adding   bmi
p = 0.0000 <  0.1000  adding   scl
p = 0.0001 <  0.1000  adding   agewoman
p = 0.0000 <  0.1000  adding   woman
      Source |       SS       df       MS              Number of obs =    4658
-------------+------------------------------           F(  5,  4652) =  304.23
       Model |  596342.421     5  119268.484           Prob > F      =  0.0000
    Residual |  1823723.08  4652  392.029897           R-squared     =  0.2464
-------------+------------------------------           Adj R-squared =  0.2456
       Total |   2420065.5  4657  519.661908           Root MSE      =    19.8

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   .5173521   .0517948     9.99   0.000     .4158098    .6188944
         bmi |   1.359621   .0734663    18.51   0.000     1.215592     1.50365
         scl |   .0327898   .0069506     4.72   0.000     .0191632    .0464163
    agewoman |   .6646316   .0709159     9.37   0.000     .5256029    .8036603
       woman |  -29.14655   3.316662    -8.79   0.000    -35.64878   -22.64432
       _cons |   65.74423   3.324712    19.77   0.000     59.22622    72.26224
------------------------------------------------------------------------------
```

b) Backward Selection

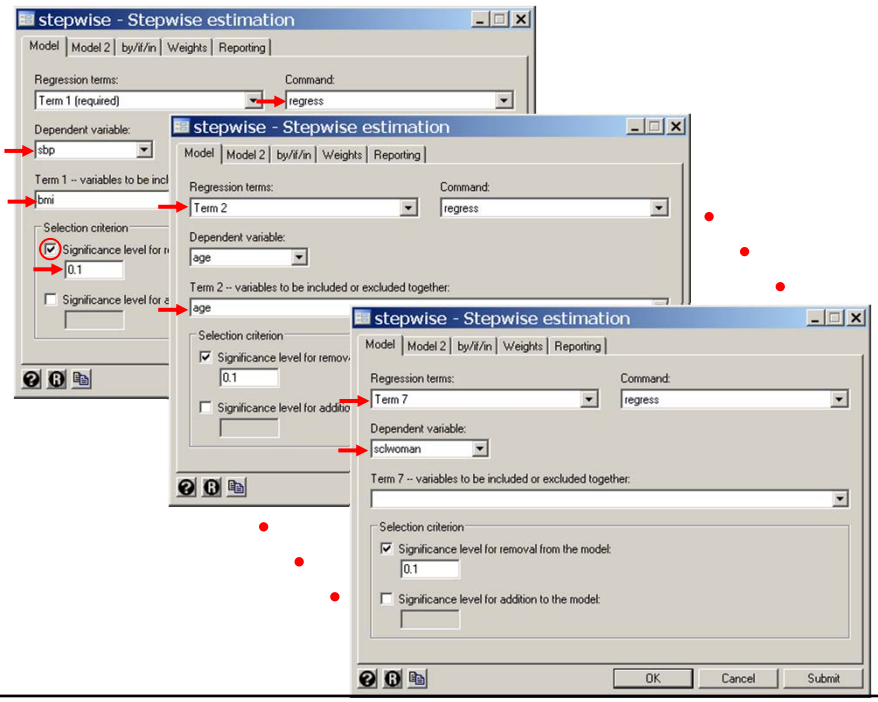
This method is similar to the forward method except that we start with all the variables and eliminate the variable with the least significance. The data is refit with the remaining variables and the process is repeated until all remaining variables have a significance level below some threshold.

The Stata command to use backward selection for our *sbp* example is

```
. * statistics > other > stepwise estimation
. stepwise, pr(.1):  regress sbp bmi age scl woman bmiwoman
>     agewoman sclwoman,
```

Here *pr*(.1) means that the program will consider variables for removal from the model if their associated P value is ≥ 0.1.

If you run this command in this example you will get the same answer as with the forward selection, which is reassuring. In general there is no guarantee that this will happen.



c) Stepwise Selection

This method is like the forward method except that at each step, previously selected variables whose significance has dropped below some threshold are dropped from the model.

Suppose:

x_1 is the best single predictor of y

x_2 and x_3 are chosen next and together predict y better than x_1

Then it makes sense to keep x_2 and x_3 and drop x_1 from the model.

In the Stata *stepwise* command this is done with the options -

```
,forward pe(.1) pr(.2)
```

which would consider new variables for selection with $P < 0.1$ and previously selected variables for removal with $P \geq 0.2$.

11. Pros and cons of automated model selection

i) Automatic selection methods are fast and easy to use.

ii) They are best used when we have a small number of variables of primary interest and wish to explore the effects of potential confounding variables on our models.

iii) They can be misleading when used for exploratory analyses in which the primary variables of interest are unknown and the number of potential covariates is large. In this case these methods can exaggerate the importance of a small number of variables due to multiple comparisons artifacts.

iv) It is a good idea to use more than one method to see if you come up with the same model.

v) Fitting models by hand may sometimes be worth the effort.

12. Residuals, Leverage, and Influence

a) Residuals

The residual for the i^{th} patient is $e_i = y_i - \hat{y}_i$

b) Estimating the variance σ^2

We estimate σ^2 by $s^2 = \Sigma(y_i - \hat{y}_i)^2/(n - k - 1)$ **{2.2}**

which is denoted Mean Square for Error in most computer programs. In Stata it is the term in the *Residual* row and the *MS* column. k is the number of covariates in the model.

```
. regress sbp bmi age scl woman agewoman


  Source |       SS       df       MS                 Number of obs =    4658
---------+------------------------------               F(  5,  4652) =  304.23
   Model |  596342.421     5  119268.484               Prob > F      =  0.0000
Residual |  1823723.08  4652  392.029897               R-squared     =  0.2464
---------+------------------------------               Adj R-squared =  0.2456
   Total |  2420065.50  4657  519.661908               Root MSE      =   19.80

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   1.359621   .0734663     18.507   0.000       1.215592     1.50365
     age |   .5173521   .0517948      9.988   0.000       .4158098    .6188944
     scl |   .0327898   .0069506      4.718   0.000       .0191632    .0464163
   woman |  -29.14655   3.316662     -8.788   0.000      -35.64878   -22.64432
agewoman |   .6646316   .0709159      9.372   0.000       .5256029    .8036603
   _cons |   65.74423   3.324712     19.774   0.000       59.22622    72.26224
------------------------------------------------------------------------------
```

c) Leverage

The leverage h_i of the i^{th} patient is a measure of her potential to influence the parameter estimates if the i^{th} residual is large.

h_i has a complex formula involving the covariates $x_1, x_2, \ldots, x_k$ (but not the dependent variable y).

In all cases $0 < h_i < 1$.

The larger h_i the greater the leverage.

The variance of $\hat{y}_i$ is
$\text{var}(\hat{y}_i) = h_i s^2$.

Note that $h_i = \text{var}(\hat{y}_i)/s^2$.

Hence h_i can be defined as the variance of $\hat{y}_i$ measured in units of s^2 .

d) Residual variance

The variance of e_i is $s^2(1-h_i)$

e) Standardized and Studentized residual

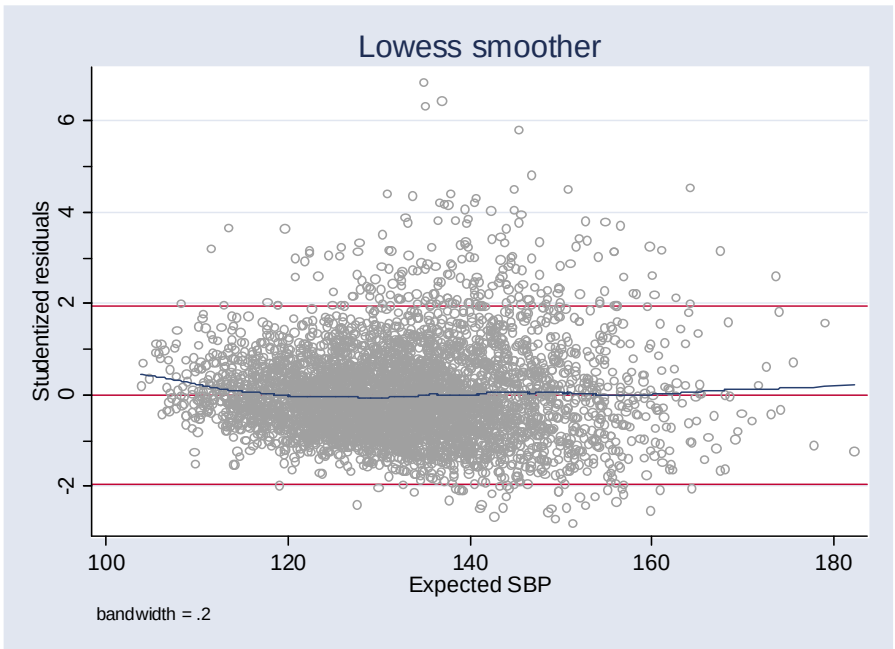
The standardized residual is $r_i = e_i / (s\sqrt{1-h_i})$ **{2.3}**

The studentized residual is $t_i = e_i/(s_{(i)}\sqrt{1-h_i})$ **{2.4}**

where $s_{(i)}$ is the estimate of σ obtained from equation (2.2) with the i^{th} case deleted (t_i is also called the jackknifed residual).

It is often helpful to plot the studentized residual against its expected value. We do this in Stata as we continue the session recorded in *FramSBPbmiMulti.log*.

```
. predict yhat, xb
(41 missing values generated)

. predict res, rstudent

. * Statistics > Nonparametric analysis > Lowess smoothing
. lowess res yhat,  bwidth(0.2) symbol(oh) color(gs10) lwidth(thick)      ///
>     yline(-1.96 0 1.96) ylabel(-2 (2) 6) ytick(-2 (1) 6)                ///
>     xlabel(100 (20) 180) xtitle(Expected SBP)
```



If our model fit perfectly, the lowess regression line would be flat and equal to zero, 95% of the studentized residuals would lie between ± 2 and should be symmetric about zero. In this example the residuals are skewed but the regression line keeps close to zero except for very low values of expected SBP.

Thus, this graph supports the validity of the model with respect to the expected SBP values but not with respect to the distribution of the residuals. The very large sample size, however, should keep the non-normally distributed residuals from adversely affecting our conclusions.

f) Influence

The influence of a patient is the extent to which he determines the value of the regression coefficients.

13. Cook's Distance: Detecting Multivariate Outliers

One measure of influence is Cook's distance, D_i, which is a function of r_i and h_i. The removal of a patient with a D_i value greater than 1 shifts the parameter estimates outside the 50% confidence region based on the entire data set.

Checking observations with a Cook's distance greater than 0.5 is worthwhile. Such observations should be double checked for errors. If they are valid you may need to discuss them explicitly in you paper.

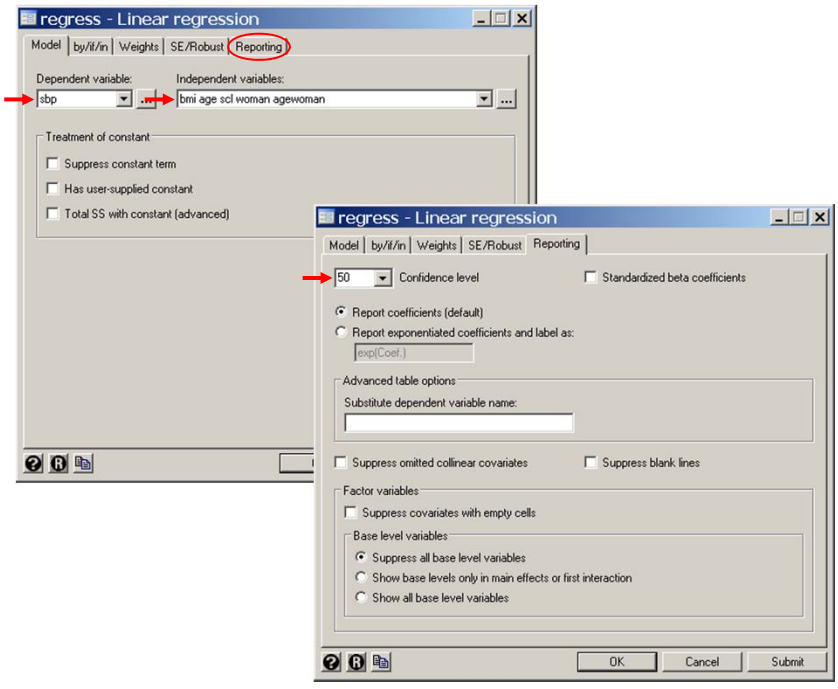
It is possible for a multivariate outlier to have a major effect on the parameter estimates but not be an obvious outlier on a 2×2 scatter plot.

14. Cook's Distance in the SBP Regression Example

The Framingham data set is so large that no individual observation has an appreciable effect on the parameter estimates (the maximum Cook's distance is 0.009). We illustrate the influence of individual patients in a subset analysis of subjects with IDs from 2001 to 2050.
FramSBPbmiMulti.log continues as follows.

```
. *
. *  Illustrate influence of individual data points on
. *  the parameter estimates of a linear regression.
. *
. *  Variables Manager (right click on variable to be dropped or kept)
. drop res
*  Data > Create or change data > Keep or drop observations
. keep if id > 2000 & id <= 2050
(4649 observations deleted)

. regress sbp bmi age scl woman agewoman, level(50)                    {1}
```

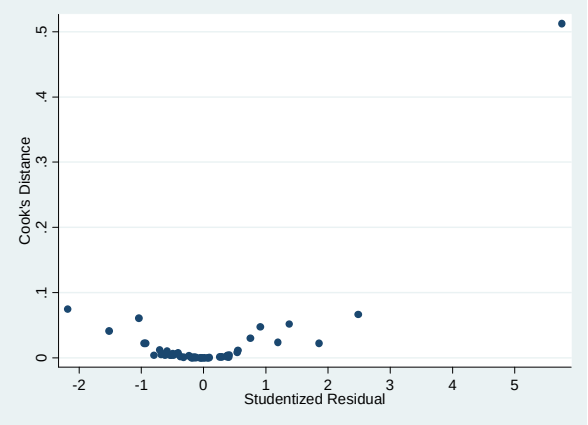
{1} The *level*(50) option specifies that 50% confidence intervals will be given for the parameter estimates.



```
  Source |       SS       df       MS                 Number of obs =      49
---------+------------------------------               F(  5,    43) =    2.13
   Model |  7953.14639     5  1590.62928               Prob > F      =  0.0796
Residual |  32056.6903    43  745.504427               R-squared     =  0.1988
---------+------------------------------               Adj R-squared =  0.1056
   Total |  40009.8367    48  833.538265               Root MSE      =  27.304

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [50% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   .5163516   1.004381      0.514   0.610      -.1668667     1.19957
     age |   .0232767   .7929254      0.029   0.977      -.5161014    .5626547
     scl |   .0618257   .0884284      0.699   0.488       .0016733    .1219781
   woman |  -72.75275    46.5895     -1.562   0.126      -104.4447   -41.06079
agewoman |   1.726515   1.018715      1.695   0.097       1.033546    2.419483
   _cons |   102.6837   46.23653      2.221   0.032       71.23184    134.1355
------------------------------------------------------------------------------
. predict res, rstudent
(1 missing value generated)

. predict cook, cooksd                                                    {2}
(1 missing value generated)
```

{2} Define *cook* to equal the Cook's distance for each data point.

```
. label variable res "Studentized Residual"
. label variable cook "Cook's Distance"
. scatter cook res, ylabel(0 (.1) .5) xlabel(-2 (1) 5)
```



The graph shows that we have one enormous residual with great influence. Note however that there are also large residuals with little influence.

The log file continues as follows:

```
. list cook res id bmi sbp if res > 2
```

	cook	res	id	bmi	sbp
46.	.	.	2046	25.6	118
48.	.06611	2.485642	2048	24.6	190
49.	.5121304	5.756579	2049	19.5	260

{1}

```
. regress sbp bmi age scl woman agewoman if id ~= 2049, level(50)
```

{2}

```
  Source |       SS       df       MS              Number of obs =      48
---------+------------------------------           F(  5,    42) =    2.83
   Model |  6036.25249     5   1207.2505           Prob > F      =  0.0273
Residual |  17918.7267    42  426.636349           R-squared     =  0.2520
---------+------------------------------           Adj R-squared =  0.1629
   Total |  23954.9792    47  509.680408           Root MSE      =  20.655

------------------------------------------------------------------------------
     sbp |      Coef.   Std. Err.       t     P>|t|       [50% Conf. Interval]
---------+--------------------------------------------------------------------
     bmi |   1.776421   .7907071      2.247   0.030       1.238443    2.314399
     age |  -.0069364    .599864     -0.012   0.991      -.4150694    .4011967
     scl |   .0568255    .066901      0.849   0.400       .0113077    .1023433
   woman |  -42.87799   35.62457     -1.204   0.235       -67.1161   -18.63989
agewoman |   .9782689   .7815332      1.252   0.218       .4465325    1.510005
   _cons |   73.63212   35.33972      2.084   0.043       49.58782    97.67642
------------------------------------------------------------------------------
```

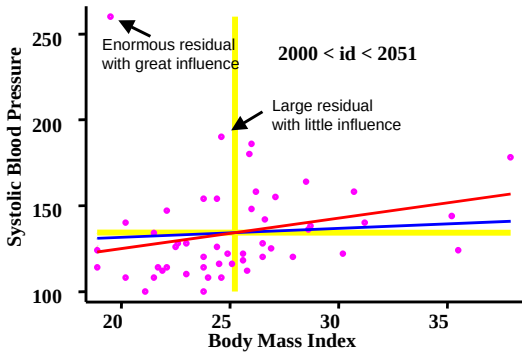
{1} The patient with the large Cook's D has ID 2049.

{2} We repeat the linear regression excluding this patient.

{3} **Excluding** this one patient increases the *bmi* coefficient from **0.516** to **1.78**, which exceeds the upper bound of the 50% confidence interval for *bmi* from the initial regression.

```
. regress sbp bmi age scl woman agewoman, level(50)

Source |       SS       df       MS                  Number of obs =      49
-------+------------------------------                F(  5,    43) =    2.13
 Model |  7953.14639     5  1590.62928                Prob > F      =  0.0796
Resid. |  32056.6903    43  745.504427                R-squared     =  0.1988
-------+------------------------------                Adj R-squared =  0.1056
 Total |  40009.8367    48  833.538265                Root MSE      =  27.304

------------------------------------------------------------------------------
         sbp |      Coef.   Std. Err.       t     P>|t|       [50% Conf. Interval]
-------------+----------------------------------------------------------------
         bmi |   .5163516   1.004381      0.514   0.610      -.1668667     1.19957
         age |   .0232767   .7929254      0.029   0.977      -.5161014    .5626547
         scl |   .0618257   .0884284      0.699   0.488       .0016733    .1219781
       woman |  -72.75275    46.5895     -1.562   0.126      -104.4447   -41.06079
    agewoman |   1.726515   1.018715      1.695   0.097       1.033546    2.419483
       _cons |   102.6837   46.23653      2.221   0.032       71.23184    134.1355
------------------------------------------------------------------------------
```

The following graph shows a scatter plot of *sbp* by *bmi* for these 50 patients. The red and blue lines have slopes of 1.78 and 0.516, respectively (the lines are drawn through the mean *sbp* and *bmi* values). Patients 2048 and 2049 are indicated by arrows. The influence of patient 2048 is greatly reduced by the fact that his *bmi* of 24.6 is near the mean *bmi*. The influence of patient 2049 is not only affected by her large residual but also by her low *bmi* that exerts leverage on the regression slope.



15. Least Squares Estimation

In simple linear regression we have introduced the concept of estimating parameters by the method of least squares.

- We chose a model of the form $E(y_i) = \alpha + \beta x_i$.
- We estimated α by a and β by b letting

  $\hat{y} = a + bx$ and then choosing a and b so as to minimize the sum of squared residuals $\Sigma(y - \hat{y})^2$

This approach works well for linear regression. It is ineffective for some other regression methods

Another approach which can be very useful is
maximum likelihood estimation

16. Maximum Likelihood Estimation

In simple linear regression we observed pairs of observations

$\{(y_i, x_i) : i = 1, 2, \cdots, n\}$ and fit the model $E(y_i) = \alpha + \beta x_i$

We calculate the likelihood function

$$L(\alpha, \beta \mid \{(y_i, x_i) : i = 1, 2, \cdots, n\}) \qquad \{1\}$$

which is the probability of obtaining
the observed data given the specified value of α and β.

The maximum likelihood estimates of α and β are those values of these parameters that maximize equation {1}

In linear regression the maximum likelihood and least squares estimates of α and β are identical.

17. Information Criteria for Assessing Statistical Models

We seek models that

- fit the data well
- are simple
- will be useful for future data

Increasing the number of parameters will

- improve the fit to the current data
- increase model complexity
- may exaggerate findings

We often must choose between a number of competing models. We seek measures of model fit that take into account both how well the data fit the model and the complexity of the model.

Suppose we have a model with k parameters and n observations. Let L be the maximum value of the likelihood function for this model. Then

Akaike's Information Criteria

$$\text{AIC} = -2 \log_e L + 2k$$

Schwarz's **Bayesian Information Criteria**

$$\text{BIC} = -2 \log_e L + k \log_e n$$

Models with lower values of AIC or BIC are usually preferred over models with higher values of these statistics.

Models that fit well will have higher values of L and hence lower values of $-2 \log_e L$.

Smaller models have smaller values of k and hence give lower AIC and BIC values. For studies with more than 8 patients, BIC gives a higher penalty per parameter than AIC.

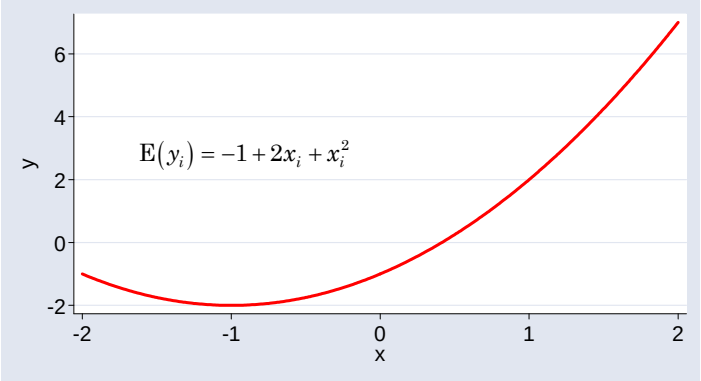
There are theoretical justifications for both methods. Neither is clearly better than the other.

18. Using Multiple Linear Regression for Non-linear Models

Multiple linear regression can be used to build simple non-linear models.

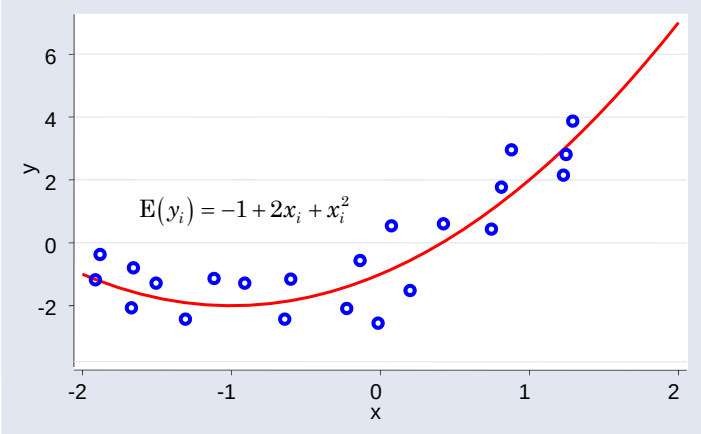
For example, suppose that there was a quadratic relationship between an independent variable x and the expected value of y. Then we could use the model

$$y_i = \alpha + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i \qquad \{2.5\}$$



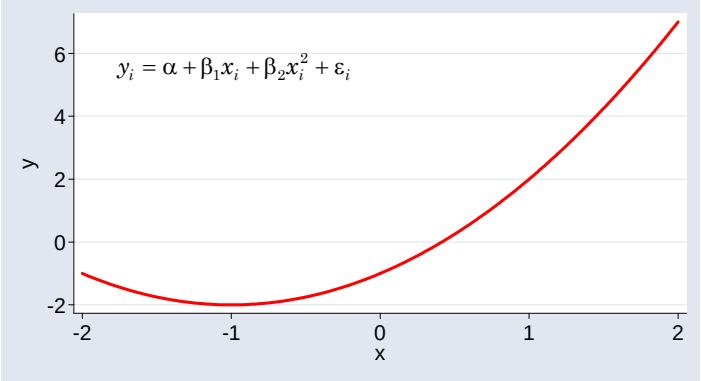
The preceding models $E(y_i)$ as a non-linear function of x_i. It is fine when correct but performs poorly for many non-linear models where the x-y relationship is not quadratic.

Extrapolating from this model is particularly problematic.



Note that {2.5} is a linear function of the parameters. Hence, it is a multiple linear regression model even though it is non-linear in x_i

We seek a more flexible approach to building non-linear regression models using multiple linear regression models.



19. Restricted Cubic Splines

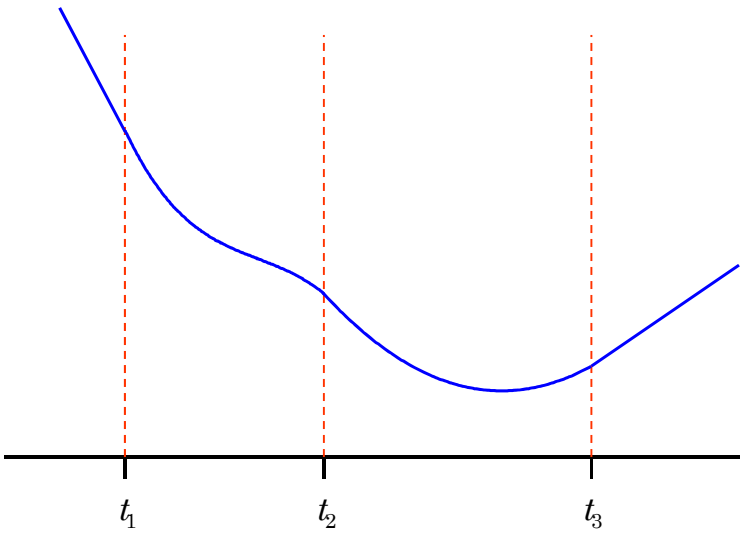
We wish to model y_i as a function of x_i using a flexible non-linear model. In a **restricted cubic spline model** we introduce k knots on the x-axis located at $t_1, t_2, \cdots, t_k$. We select a model of the expected value of y that

is linear before t_1 and after t_k.

consists of piecewise cubic polynomials between adjacent knots (i.e. of the form $ax^3 + bx^2 + cx + d$)

is continuous and smooth at each knot. (More technically, its first and second derivatives are continuous at each knot.)

An example of a restricted cubic spline with three knots is given on the next slide.



Example of a restricted cubic spline with three knots

Given x and k knots, a restricted cubic spline can be defined by

$$y = \alpha + x_1\beta_1 + x_2\beta_2 + \cdots + x_{k-1}\beta_{k-1}$$

for suitably defined values of x_i

These covariates are functions of x and the knots but are independent of y.

$x_1 = x$ and hence the hypothesis $\beta_2 = \beta_3 = \cdots = \beta_{k-1} = 0$ tests the linear hypothesis.

If x is less than the first knot then $x_2 = x_3 = \cdots = x_{k-1} = 0$
This fact will prove useful in survival analyses when calculating relative risks.

Programs to calculate $x_1, \cdots, x_{k-1}$ are available in Stata, R and other statistical software packages. The functional definitions of these terms are not pretty (see Harrell 2001), but this is of little concern given programs that will calculate them for you.

Users can specify the knot values. However, it is often reasonable to let you program choose them for you.

Harrell (2001) recommends placing knots at the quantiles of the x variable given in the following table

Number of knots k	Knot locations expressed in quantiles of the x variable						
3	0.1	0.5	0.9				
4	0.05	0.35	0.65	0.95			
5	0.05	0.275	0.5	0.725	0.95		
6	0.05	0.23	0.41	0.59	0.77	0.95	
7	0.025	0.1833	0.3417	0.5	0.6583	0.817	0.975

The basic idea of this table is to place t_1 and t_k near the extreme values of x and to space the remaining knots so that the proportion of observations between knots remains constant.

When there are fewer than 100 data points Harrell recommends replacing the smallest and largest knots by the fifth smallest and fifth largest observation, respectively.

The choice of number of knots involves a trade-off between model flexibility and number of parameters. Stone (1986) has found that more than 5 knots are rarely needed to obtain a good fit.

Five knots is a good choice when there are at least 100 data points.

Using fewer knots makes sense when there are fewer data points

It is important to always do a residual plot or, at a minimum, plot the observed and expected values to ensure that you have obtained a good fit.

The linear fits beyond the largest and smallest knots usually tracks the data well, but is not guaranteed to do so.

20. Example: the SUPPORT Study

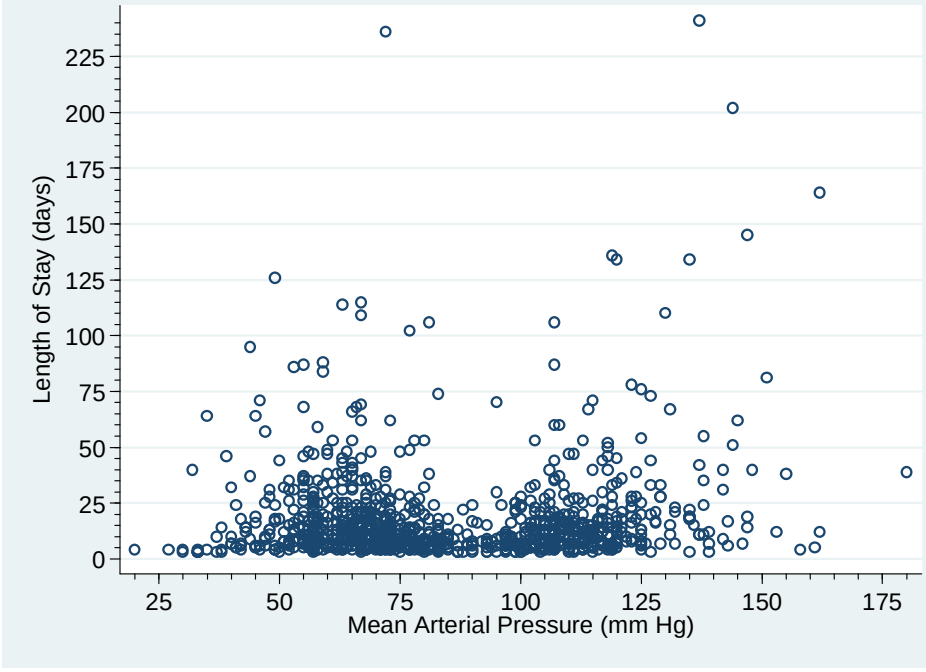
A prospective observational study of hospitalized patients

Lynn & Knauss: "Background for SUPPORT."
J Clin Epidemiol 1990; 43: 1S - 4S.

A random sample of data from 996 subjects in this study is available. See

3.25.2.SUPPORT.dta

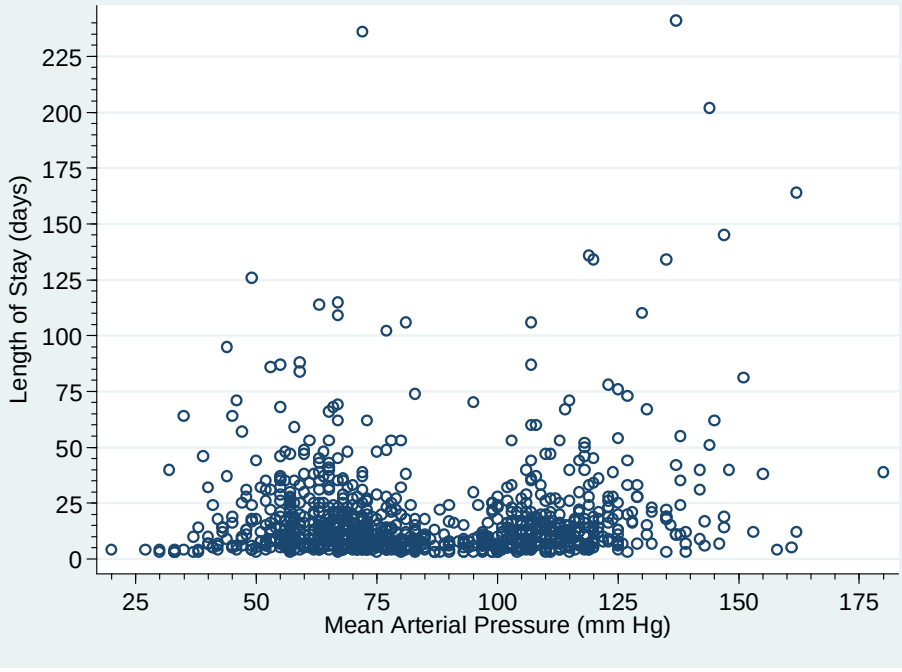
`los`	=	length of stay in days.
`map`	=	baseline mean arterial pressure
`fate`	=	1: Patient died in hospital 0: Patient discharged alive



21. Fitting a Restricted Cubic Spline with Stata

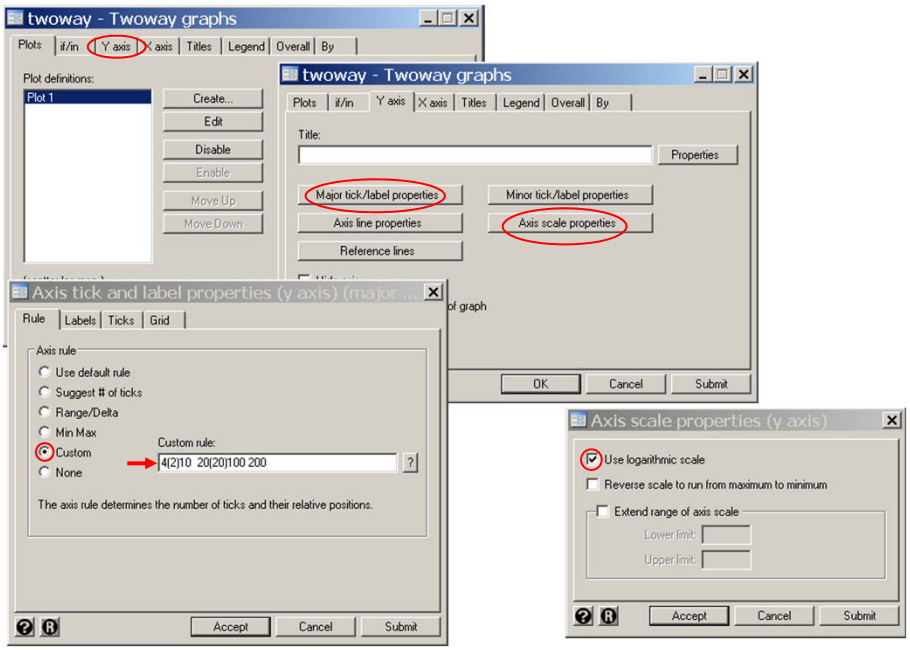
```
. *  SupportLinearRCS.log
. *
. *  Draw scatter plots of length-of-stay (LOS) by mean arterial
. *  pressure (MAP) and log LOS by MAP for the SUPPORT Study data
. *  (Lynn & Knauss, 1990).
. *
. use "C:\WDDtext\3.25.2.SUPPORT.dta" , replace

. scatter los map, symbol(Oh) xlabel(25 (25) 175) xmtick(20 (5) 180) /// {1}
>     ylabel(0(25)225, angle(0)) ymtick(5(5)240)
```

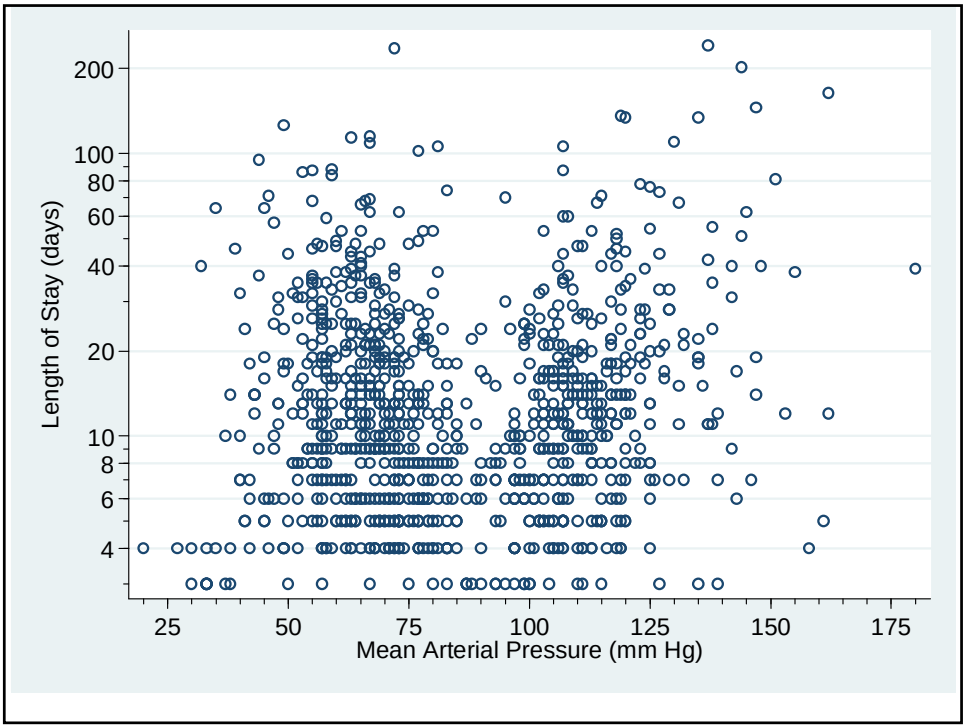
{1} Length of stay is highly skewed.



```
. scatter los map, symbol(Oh) xlabel(25 (25) 175) xmtick(20 (5) 180) ///
>     yscale(log) ylabel(4(2)10  20(20)100 200, angle(0))            /// {2}
>     ymtick(3(1)9 30(10)90)
```

{2} Plotting log LOS makes the distribution of this variable more normal. The *yscale(log)* option does this tranformation.





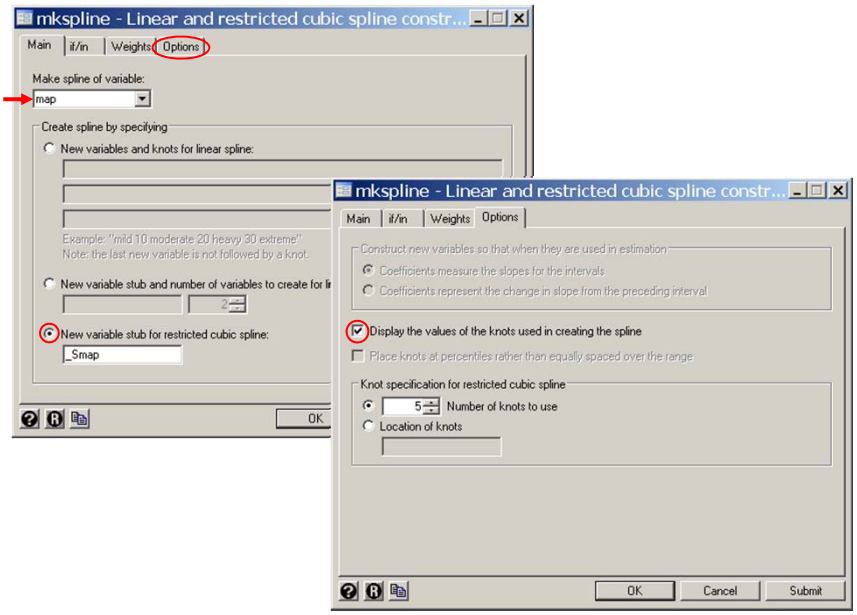

```
. *
. *  Regress log LOS against MAP using RCS models with
. *  5 knots at their default locations.  Overlay the expected
. *  log LOS from these models on a scatter plot of log LOS by MAP.
. *
. * Data > Create... > Other variable-creation... > linear and cubic...
mkspline _Smap = map, cubic displayknots                                {1}


             |     knot1      knot2      knot3      knot4      knot5
-------------+--------------------------------------------------------
         map |        47         66         78        106        129
```

{1} The **mkspline** command generates either linear or restricted cubic spline covariates. The **cubic** option specifies that restricted cubic spline covariates are to be created. This command generates these covariates for the variable **map**. By default, 5 knots are used at their default locations. Following Harrell's recommendation the computer places them at the 5th, 27.5th, 50th, 72.5th and 95th percentiles of **map**. The values of these knots are listed.

The 4 spline covariates associated with these 5 knots are named

_Smap1
_Smap2
_Smap3
_Smap4

These names are obtained by concatenating the name **_Smap** given before the equal sign with the numbers 1, 2, 3 and 4.




```
. summarize _Smap1 _Smap2 _Smap3 _Smap4                              {2}

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      _Smap1 |       996    85.31727    26.83566        20        180
      _Smap2 |       996    20.06288    27.34701         0    185.6341
      _Smap3 |       996    7.197497    11.96808         0    89.57169
      _Smap4 |       996    3.121013     5.96452         0    48.20881
```

{2} **_Smap1** is identical to **map**. The other spline covariates take non-negative values.

```
. generate log_los = log(los)

. regress log_los _S*                                                    {3}

      Source |       SS       df       MS              Number of obs =     996
-------------+------------------------------           F(  4,   991) =   24.70
       Model |  60.9019393     4  15.2254848            Prob > F      =  0.0000
    Residual |  610.872879   991  .616420665           R-squared     =  0.0907
-------------+------------------------------           Adj R-squared =  0.0870
       Total |  671.774818   995  .675150571           Root MSE      =  .78512

------------------------------------------------------------------------------
     log_los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Smap1 |   .0296009   .0059566     4.97   0.000      .017912    .0412899
      _Smap2 |  -.3317922   .0496932    -6.68   0.000    -.4293081   -.2342762
      _Smap3 |   1.263893   .1942993     6.50   0.000     .8826076    1.645178
      _Smap4 |  -1.124065   .1890722    -5.95   0.000    -1.495092   -.7530367
       _cons |    1.03603   .3250107     3.19   0.001     .3982422    1.673819
------------------------------------------------------------------------------
```

{3} This command regresses **log_los** against all variables that start with the characters **_S**. The only variables with these names are the spline covariates. An equivalent way of running this regression would be

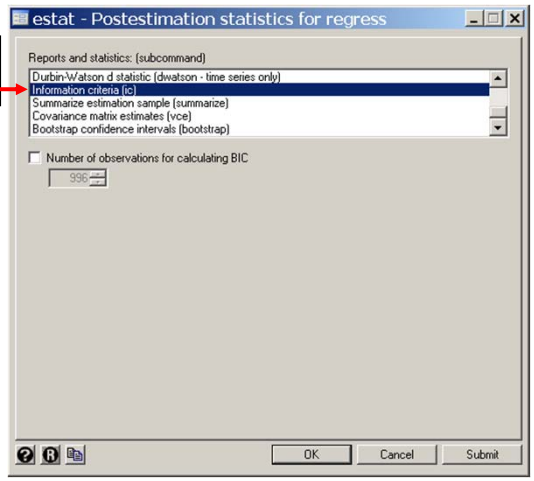
regress log_los _Smap1 _Smap2 _Smap3 _Smap4

```
. generate log_los = log(los)

. regress log_los _S*

      Source |       SS       df       MS              Number of obs =     996
-------------+------------------------------           F(  4,   991) =   24.70 {4}
       Model |  60.9019393     4  15.2254848            Prob > F      =  0.0000
    Residual |  610.872879   991  .616420665           R-squared     =  0.0907
-------------+------------------------------           Adj R-squared =  0.0870
       Total |  671.774818   995  .675150571           Root MSE      =  .78512

------------------------------------------------------------------------------
     log_los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Smap1 |   .0296009   .0059566     4.97   0.000      .017912    .0412899
      _Smap2 |  -.3317922   .0496932    -6.68   0.000    -.4293081   -.2342762
      _Smap3 |   1.263893   .1942993     6.50   0.000     .8826076    1.645178
      _Smap4 |  -1.124065   .1890722    -5.95   0.000    -1.495092   -.7530367
       _cons |    1.03603   .3250107     3.19   0.001     .3982422    1.673819
------------------------------------------------------------------------------
```

{4} This F statistic tests the null hypothesis that the coefficients associated with the parameters of the spline covariates are simultaneously zero. In other words, it tests the hypothesis that length of stay is unaffected by MAP. It is significant with P < 0.00005.

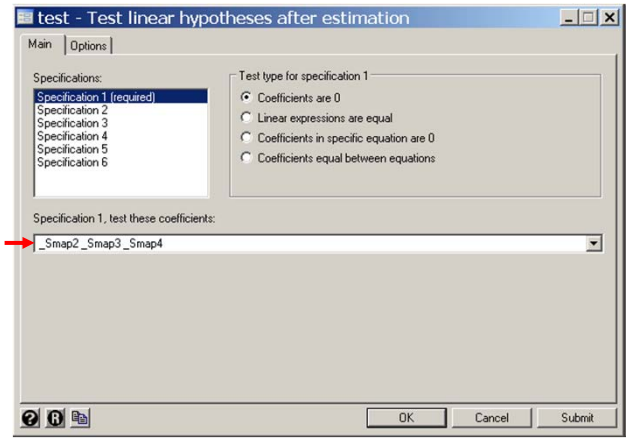
```
* Statistics > Postestimation > Reports and statistics
. estat ic                                                              {5}

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    996   -1217.138   -1169.811      5     2349.623    2374.141
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

{5} Calculate the AIC and BIC for this model.




```
. * Statistics > Postestimation > Tests > Test linear hypotheses
. test _Smap2  _Smap3 _Smap4                                            {6}

 ( 1)  _Smap2 = 0
 ( 2)  _Smap3 = 0
 ( 3)  _Smap4 = 0

       F(  3,   991) =   30.09
            Prob > F =    0.0000
```

{6} Test the null hypothesis that there is a linear relationship between **map** and **log_los**. Since **_Smap1** = **map**, this is done by testing the null hypothesis that the coefficients associated with **_Smap2**, **_Smap3** and **_Smap4** are all simultaneously zero. This test is significant with $P < 0.00005$.



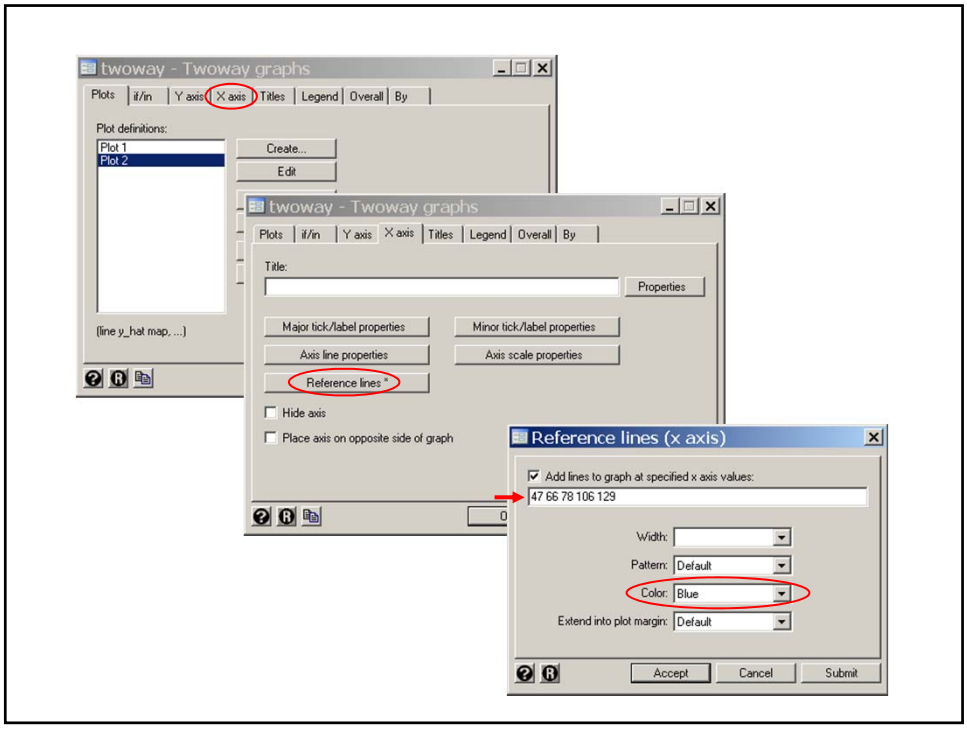
{7} **y_hat** is the estimated expected value of **log_los** under this model.

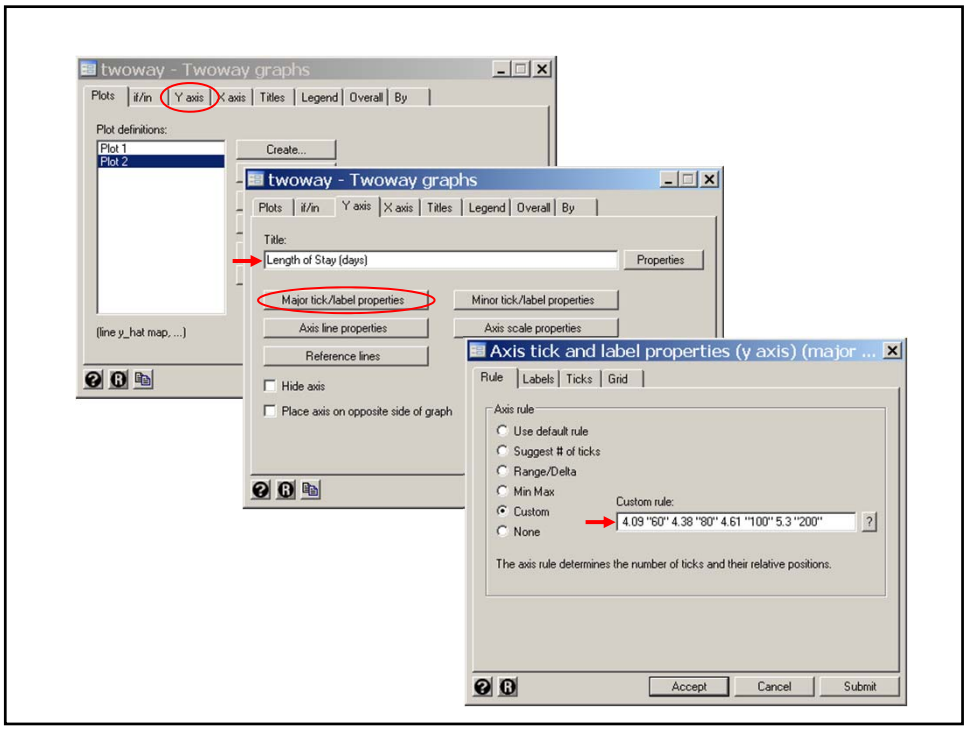
```
. predict y_hat5, xb                                                    {7}

. scatter log_los map, symbol(Oh)                                       /// {8}
>     || line y_hat5 map, color(red) lwidth(medthick)                   ///
>     , xlabel(25 (25) 175) xmtick(20 (5) 180)                          ///
>       xline(47 66 78 106 129, lcolor(blue))                           /// {9}
>       ylabel(1.39 "4" 1.79 "6" 2.08 "8" 2.3 "10" 3 "20"              /// {10}
>           3.69 "40" 4.09 "60" 4.38 "80" 4.61 "100" 5.3 "200", angle(0)) ///
>       ymtick(1.1 1.39 1.61 1.79 1.95 2.08 2.2 3.4 3.91 4.25 4.5)      ///
>       ytitle(Length of Stay (days))                                   ///
>        legend(order(1 "Observed" 2 "Expected"))
```

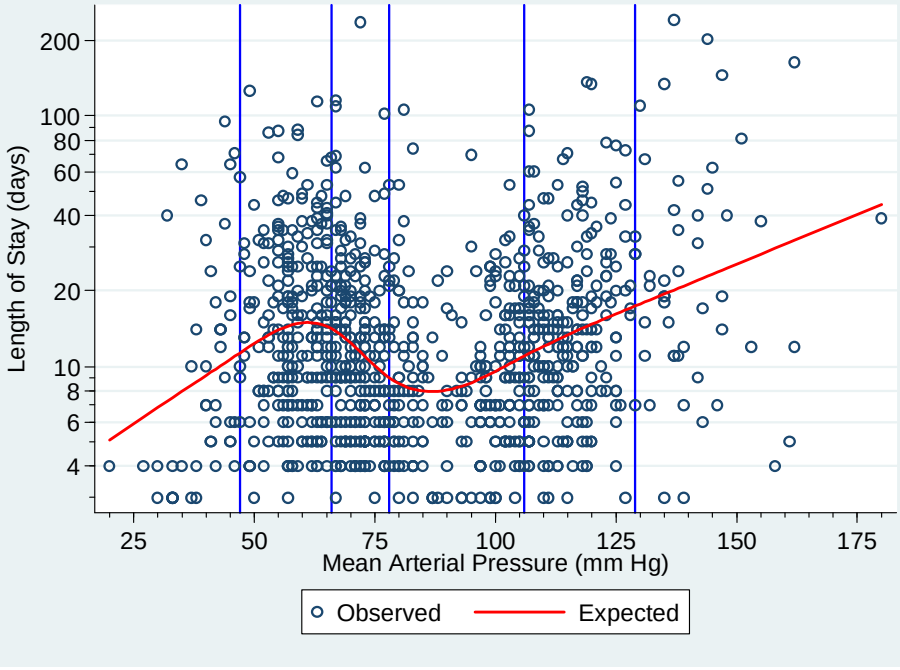
{8} Graph a scatterplot of **log_los** vs. **map** together with a line plot of the expected **log_los** vs. **map**.

{9} This **xline** option draws vertical lines at each of the five knots. The **lcolor** suboption colors these lines blue.

{10} The units of the *y*-axis is length of stay. This *ylabel* option places the label 4 at the *y*-axis value 1.39 = log(4), 6 at the value 1.79 = log(6), etc.

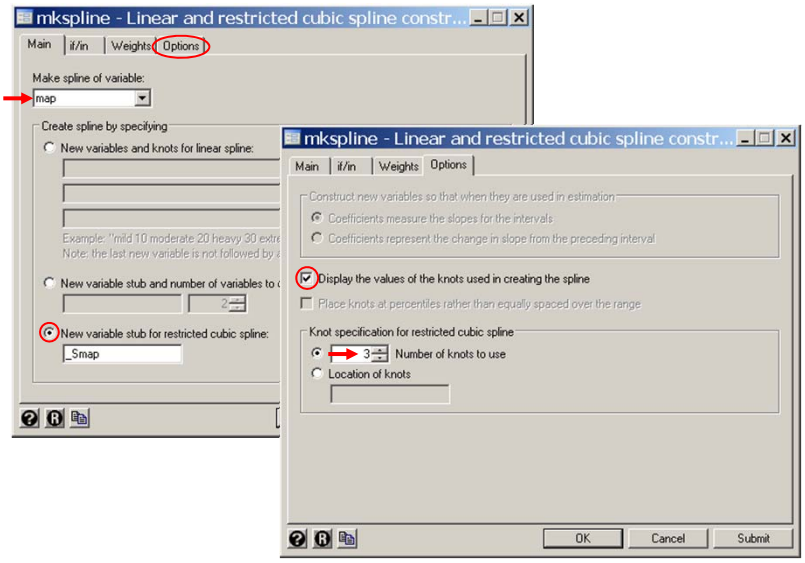

twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
Edit
(line y_hat map, ...)
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines *
Hide axis
Place axis on opposite side of graph
Reference lines (x axis)
Add lines to graph at specified x axis values:
47 66 78 106 129
Width:
Pattern:
Default
Color:
Blue
Extend into plot margin:
Default
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
(line y_hat map, ...)
Title:
Length of Stay (days)
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis tick and label properties (y axis) (major ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
Custom rule:
4.09 "60" 4.38 "80" 4.61 "100" 5.3 "200"
The axis rule determines the number of ticks and their relative positions.
Accept
Cancel
Submit




```
. *
. *  Plot expected LOS for models with 3, 4, 6 and 7 knots.
. *  Use the default knot locations. Calculate AIC and BIC for each model.
. *
. *  Variables Manager
. drop _S*

. * Data > Create... > Other variable-creation... > linear and cubic...
. mkspline _Smap = map, nknots(3) cubic displayknots                {11}

             |     knot1      knot2      knot3
-------------+---------------------------------
         map |        55         78        120
```

{11} Define **2** spline covariates associated with **3** knots at their default locations. The **`nknots`** option specifies the number of knots.




```
. regress log_los _S*

      Source |       SS       df       MS              Number of obs =     996
-------------+------------------------------           F(  2,   993) =   18.24
       Model |  23.8065057     2  11.9032528            Prob > F      =  0.0000
    Residual |  647.968313   993  .652536065           R-squared     =  0.0354
-------------+------------------------------           Adj R-squared =  0.0335
       Total |  671.774818   995  .675150571           Root MSE      =   .8078

------------------------------------------------------------------------------
     log_los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Smap1 |  -.0110138   .0027449    -4.01   0.000    -.0164002   -.0056274
      _Smap2 |   .0226496    .004248     5.33   0.000     .0143135    .0309858
       _cons |   3.124095   .1827706    17.09   0.000     2.765435    3.482756
------------------------------------------------------------------------------

. predict y_hat3, xb

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    996   -1217.138    -1199.17      3      2404.34    2419.051
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

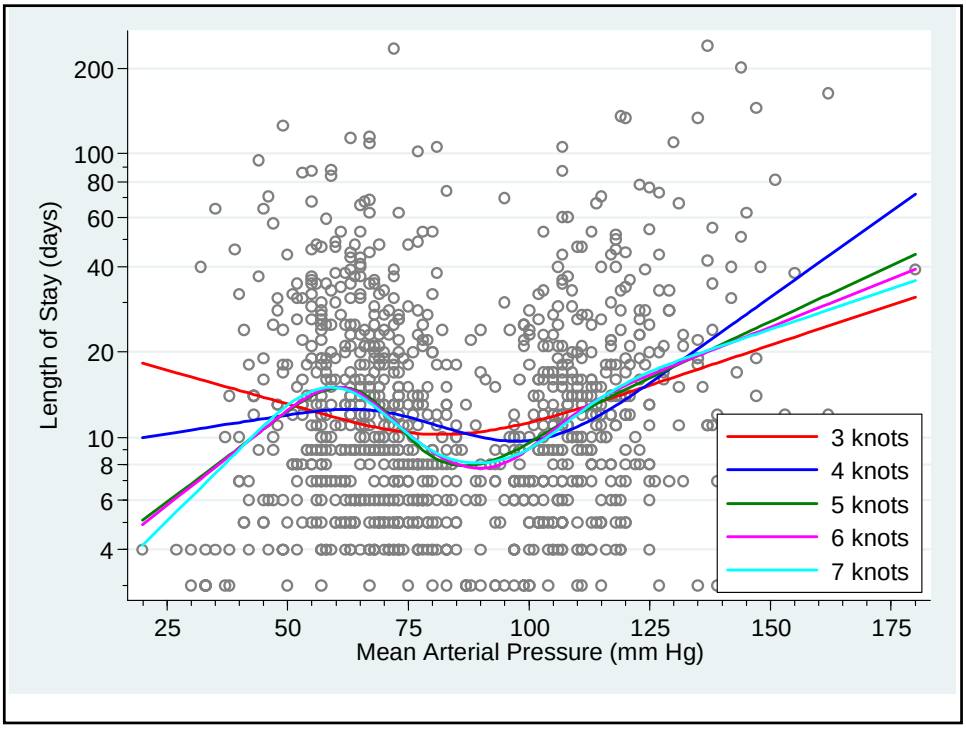
```
. drop _S*

. mkspline _Smap = map, nknots(4) cubic displayknots

             |     knot1      knot2      knot3      knot4
-------------+--------------------------------------------
         map |        47         69        100        129

. regress log_los _S*

      Source |       SS       df       MS              Number of obs =     996
-------------+------------------------------           F(  3,   992) =   21.40
       Model |  40.8276008     3  13.6092003           Prob > F      =  0.0000
    Residual |  630.947217   992  .636035501           R-squared     =  0.0608
-------------+------------------------------           Adj R-squared =  0.0579
       Total |  671.774818   995  .675150571           Root MSE      =  .79752

------------------------------------------------------------------------------
     log_los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Smap1 |   .0060744    .004387     1.38   0.166    -.0025343    .0146832
      _Smap2 |  -.0533119   .0155968    -3.42   0.001    -.0839184   -.0227054
      _Smap3 |   .1509453   .0342118     4.41   0.000     .0838095    .2180812
       _cons |   2.180462   .2600792     8.38   0.000     1.670093     2.69083
------------------------------------------------------------------------------

. predict y_hat4, xb
```

```
. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    996   -1217.138   -1185.913      4     2379.827    2399.442
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

. drop _S*

. mkspline _Smap = map, nknots(6) cubic displayknots

             |     knot1      knot2      knot3      knot4      knot5      knot6
-------------+------------------------------------------------------------------
         map |        47         63         73         93     108.69        129
```

```
. regress log_los _S*

      Source |       SS       df       MS              Number of obs =     996
-------------+------------------------------           F(  5,   990) =   20.18
       Model |  62.1303583     5  12.4260717            Prob > F      =  0.0000
    Residual |   609.64446   990  .615802485            R-squared     =  0.0925
-------------+------------------------------           Adj R-squared =  0.0879
       Total |  671.774818   995  .675150571            Root MSE      =  .78473

------------------------------------------------------------------------------
     log_los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Smap1 |     .03099    .006904     4.49   0.000     .0174418    .0445382
      _Smap2 |  -.3837563   .0874071    -4.39   0.000    -.5552809   -.2122318
      _Smap3 |   1.111961   .3834093     2.90   0.004     .3595729    1.864349
      _Smap4 |  -.5873248   .4457995    -1.32   0.188    -1.462145    .2874957
      _Smap5 |  -.4824613   .2991149    -1.61   0.107    -1.069433    .1045108
       _cons |   .9745223   .3623654     2.69   0.007     .2634297    1.685615
------------------------------------------------------------------------------
. predict y_hat6, xb

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    996   -1217.138   -1168.809      6     2349.618     2379.04
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note
```

```
. drop _S*

. mkspline _Smap = map, nknots(7) cubic displayknots

             |  knot1    knot2     knot3    knot4   -  knot5     knot6      knot7
-------------+-------------------------------------------------------------------
         map |     41       60        69       78   101.3251       113    138.075

. regress log_los _S*

      Source |       SS       df       MS              Number of obs =     996
-------------+------------------------------           F(  6,   989) =   16.92
       Model |  62.5237582     6  10.4206264            Prob > F      =  0.0000
    Residual |   609.25106   989  .616027361            R-squared     =  0.0931
-------------+------------------------------           Adj R-squared =  0.0876
       Total |  671.774818   995  .675150571            Root MSE      =  .78487

------------------------------------------------------------------------------
     log_los |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Smap1 |   .0389453   .0092924     4.19   0.000     .0207101    .0571804
      _Smap2 |  -.3778786     .12678    -2.98   0.003    -.6266673     -.12909
      _Smap3 |   .9316267   .8933099     1.04   0.297    -.8213739    2.684627
      _Smap4 |   .1269005    1.58931     0.08   0.936    -2.991907    3.245708
      _Smap5 |  -.7282771   1.034745    -0.70   0.482    -2.758824     1.30227
      _Smap6 |  -.3479716   .4841835    -0.72   0.473    -1.298117    .6021733
       _cons |   .6461153   .4496715     1.44   0.151    -.2363046    1.528535
------------------------------------------------------------------------------
```

```
. predict y_hat7, xb

. estat ic

-----------------------------------------------------------------------------
       Model |    Obs    ll(null)   ll(model)     df          AIC         BIC
-------------+---------------------------------------------------------------
           . |    996   -1217.138   -1168.487      7     2350.975    2385.301
-----------------------------------------------------------------------------
               Note:  N=Obs used in calculating BIC; see [R] BIC note

.
. twoway scatter log_los map, symbol(Oh) color(gray)                        ///
>     || line y_hat3 map, color(red) lwidth(medthick)                       ///
>     || line y_hat4 map, color(blue) lwidth(medthick)                      ///
>     || line y_hat5 map, color(green) lwidth(medthick)                     ///
>     || line y_hat6 map, color(magenta) lwidth(medthick)                   ///
>     || line y_hat7 map, color(cyan) lwidth(medthick)                      ///
>     , xlabel(25 (25) 175) xmtick(20 (5) 180)                              ///
>       ylabel(1.39 "4" 1.79 "6" 2.08 "8" 2.3 "10" 3 "20"                  ///
>           3.69 "40" 4.09 "60" 4.38 "80" 4.61 "100" 5.3 "200", angle(0))  ///
>       ymtick(1.1 1.39 1.61 1.79 1.95 2.08 2.2 3.4 3.91 4.25 4.5)          ///
>       ytitle(Length of Stay (days)) legend(ring(0) position(4) col(1)     ///
>         order(2 "3 knots" 3 "4 knots" 4 "5 knots"                         ///
>               5 "6 knots" 6 "7 knots"))
```



Restricted cubic spline models of log length-of-stay by mean arterial pressure

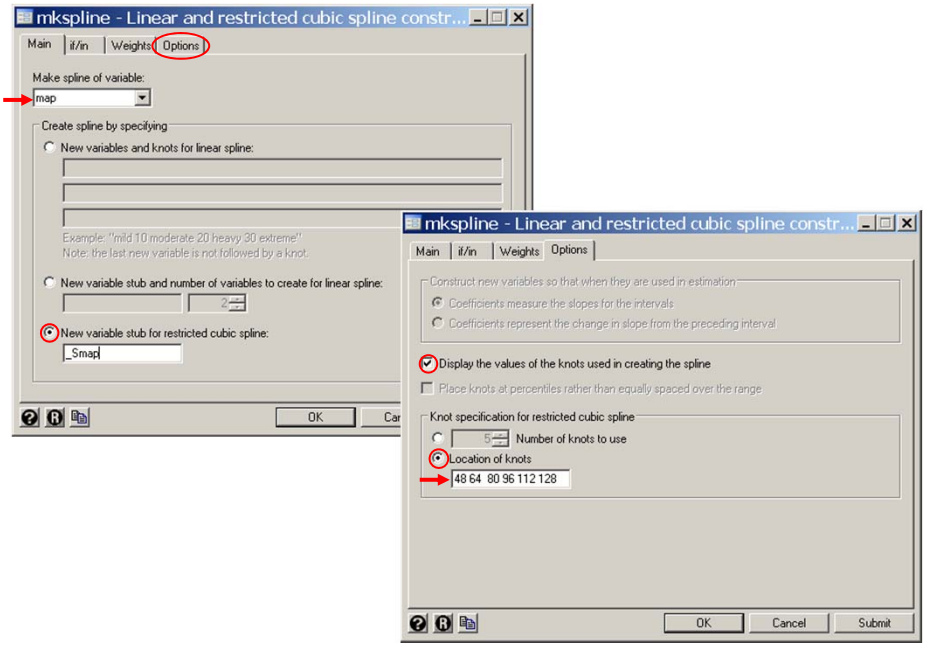
Knots	AIC	BIC
3	2,404.340	2,419.051
4	2,379.827	2,399.442
5	2,349.623	2,374.141
6	2,349.618	2,379.040
7	2,350.975	2,385.301

- Models with AIC values within 1 or 2 of the minimum disserve consideration.
- Models with AIC values > 10 above the minimum may be discarded.

- Clearly the 3 and 4 knot models provide a poor fit.
- I have decided to use the 6 knot model but 5 or 7 knots would also be fine. Note that the 6 knot model lies between the 5 and 7 knot model,
- We have lots of observation and few parameters so the number of knots is not too important.

```
. *
. *  Plot expected LOS for the 6 knot model together with 95%
. *  confidence bands.  Use evenly spaced knot locations.
. *
. drop _S*

. * Data > Create... > Other variable-creation... > linear and cubic...
. mkspline _Smap = map,  knots(48 64 80 96 112 128)                    /// {12}
>     cubic displayknots

             |     knot1      knot2      knot3      knot4      knot5      knot6
-------------+------------------------------------------------------------------
         map |        48         64         80         96        112        128
```

{12} Define **5** spline covariates associated with **6** knots at evenly spaced locations. The `knots` option specifies the knot locations




```
. regress log_los _S*

. predict y_hat, xb

. predict se, stdp

. generate lb = y_hat - invttail(_N-6, 0.025)*se

. generate ub = y_hat + invttail(_N-6, 0.025)*se

. twoway rarea lb ub map , color(yellow)
>     || scatter log_los map, symbol(Oh) color(blue)
>     || line y_hat map, color(red) lwidth(medthick)
>     , xlabel(25 (25) 175) xmtick(20 (5) 180)
>       xline(48(16)128, lcolor(gray))
>       ylabel(1.39 "4" 1.79 "6" 2.08 "8" 2.3 "10" 3 "20"
>           3.69 "40" 4.09 "60" 4.38 "80" 4.61 "100" 5.3 "200", angle(0))
>       ymtick(1.1 1.39 1.61 1.79 1.95 2.08 2.2 3.4 3.91 4.25 4.5)
>       subtitle("Evenly" "Spaced" "Knots", ring(0) position(10))
>       ytitle(Length of Stay (days)) legend(off)
```

{Output Omitted}

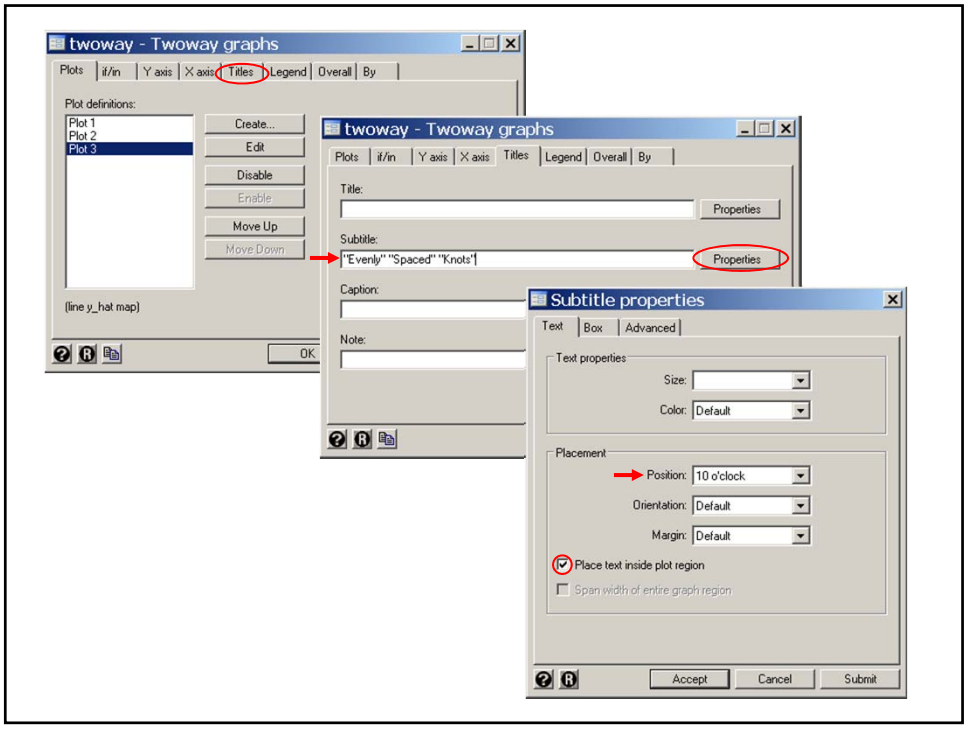
{15} Add a subtitle inside the graph at the 10 o'clock position. Placing the words in separate quotes causes them to be printed on separate lines.

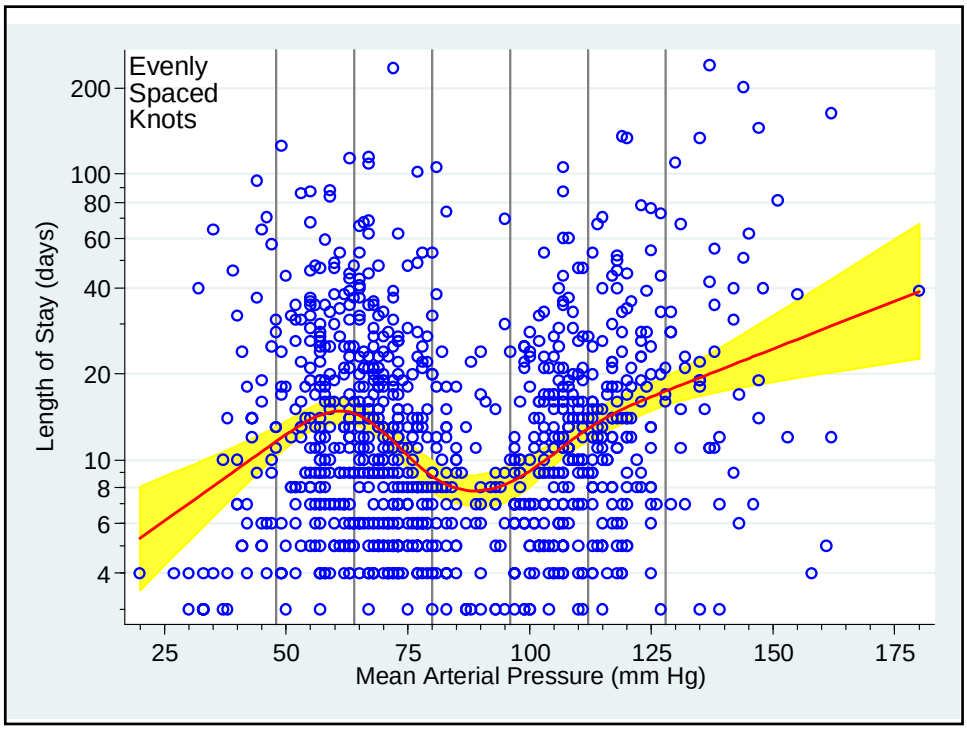
{13}

/// {14}
/// {15}

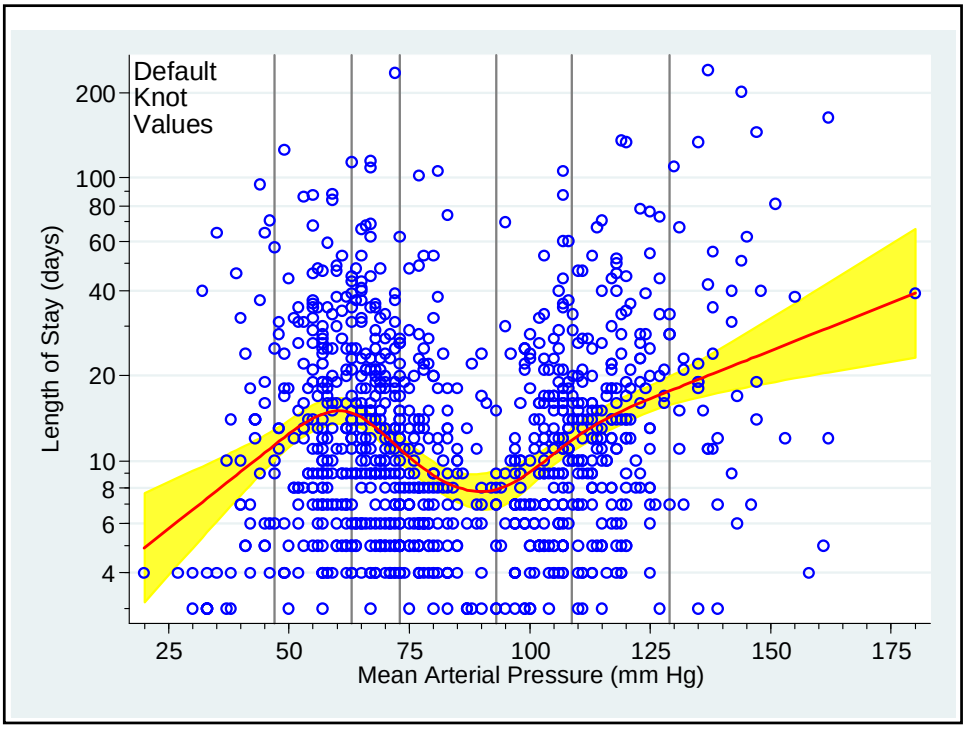
{13} _N-6 = the number of observations minus the number of parameters = 990 = the degrees of freedom of the MSE s^2. `lb` is the lower bound of the 95% confidence interval for `y_hat`.

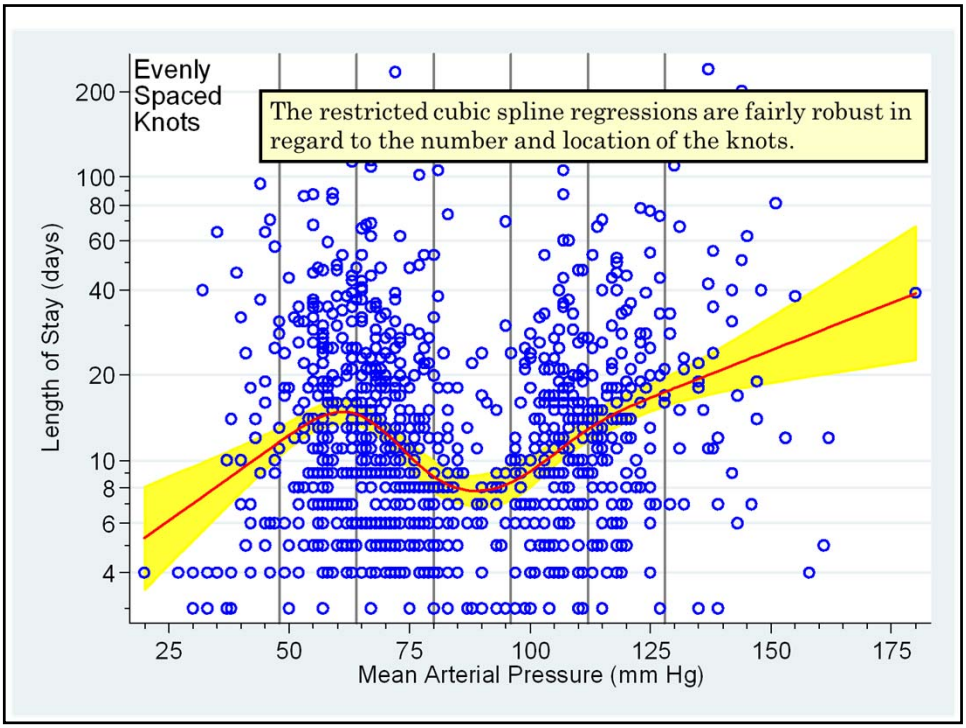
{14} This plot adds the 95% confidence region for the regression curve.


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Create...
Edit
Disable
Enable
Move Up
Move Down
(line y_hat map)
OK
twoway - Twoway graphs
Title:
Properties
Subtitle:
"Evenly" "Spaced" "Knots"
Properties
Caption:
Note:
Subtitle properties
Text
Box
Advanced
Text properties
Size:
Color:
Default
Placement
Position:
10 o'clock
Orientation:
Default
Margin:
Default
Place text inside plot region
Span width of entire graph region
Accept
Cancel
Submit


Evenly
Spaced
Knots
Length of Stay (days)
200
100
80
60
40
20
10
8
6
4
25
50
75
100
125
150
175
Mean Arterial Pressure (mm Hg)

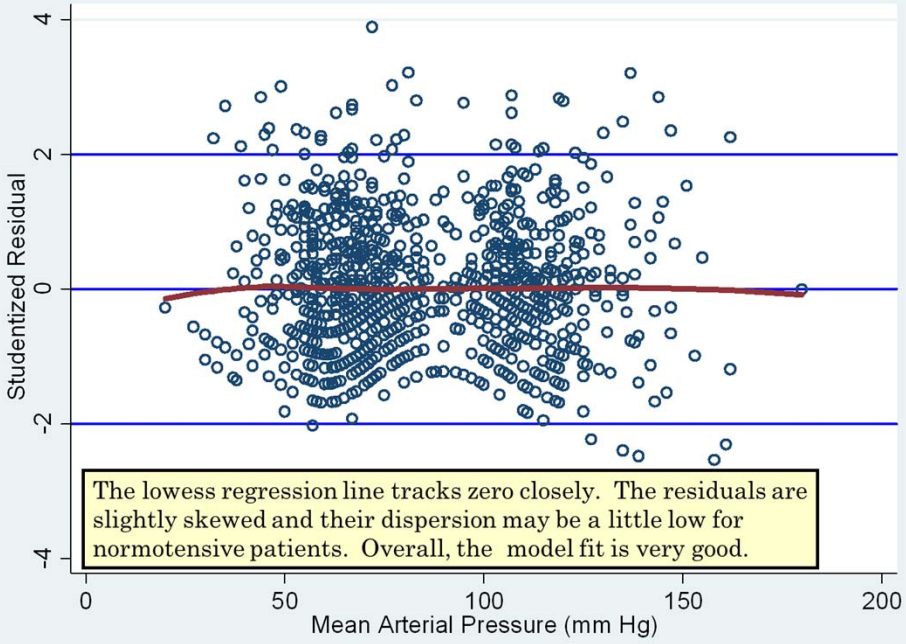
```
. *
. *  Replot 6 knot model with default knot spacing.
. *
. drop _S* y_hat se lb ub

. mkspline _Smap = map, nknots(6) cubic

. regress log_los _S*
                                                              {Output Omitted}

. predict y_hat, xb

. predict se, stdp

. generate lb = y_hat - invttail(_N-6, 0.025)*se

. generate ub = y_hat + invttail(_N-6, 0.025)*se

. twoway rarea lb ub map , color(yellow)                                  ///
    || scatter log_los map, symbol(Oh) color(blue)                        ///
    || line y_hat map, color(red) lwidth(medthick)                        ///
    , xlabel( 25 (25) 175) xmtick( 20 (5) 180)                            ///
      ylabel( 1.39 "4" 1.79 "6" 2.08 "8" 2.3 "10" 3 "20"                 ///
          3.69 "40" 4.09 "60" 4.38 "80" 4.61 "100" 5.3 "200", angle(0)) ///
      ymtick( 1.1 1.39 1.61 1.79 1.95 2.08 2.2 3.4 3.91 4.25 4.5)         ///
      xline( 47 63 73 93 108.69 129, lcolor(gray))                        ///
      ytitle( Length of Stay (days))                                      ///
      subtitle( "Default" "Knot" "Values"                                 ///
          , ring(0) position(10)) legend(off)
```






```
. predict rstudent, rstudent

. generate big = abs(rstudent)>2

. * Statistics > Summaries, tables and tests > Tables > One-way tables
. tabulate big

        big |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |        949       95.28       95.28
          1 |         47        4.72      100.00
------------+-----------------------------------
      Total |        996      100.00
. *
. *  Draw a scatter plot of the studentized residuals against MAP
. *  Overlay the associated lowess regression curve on this graph.
. *
. twoway scatter rstudent map, symbol(Oh)                          ///
>     || lowess rstudent map, lwidth(thick)                        ///
>     ,  ytitle(Studentized Residual) yline(-2 0 2, lcolor(blue)) legend(off)
```

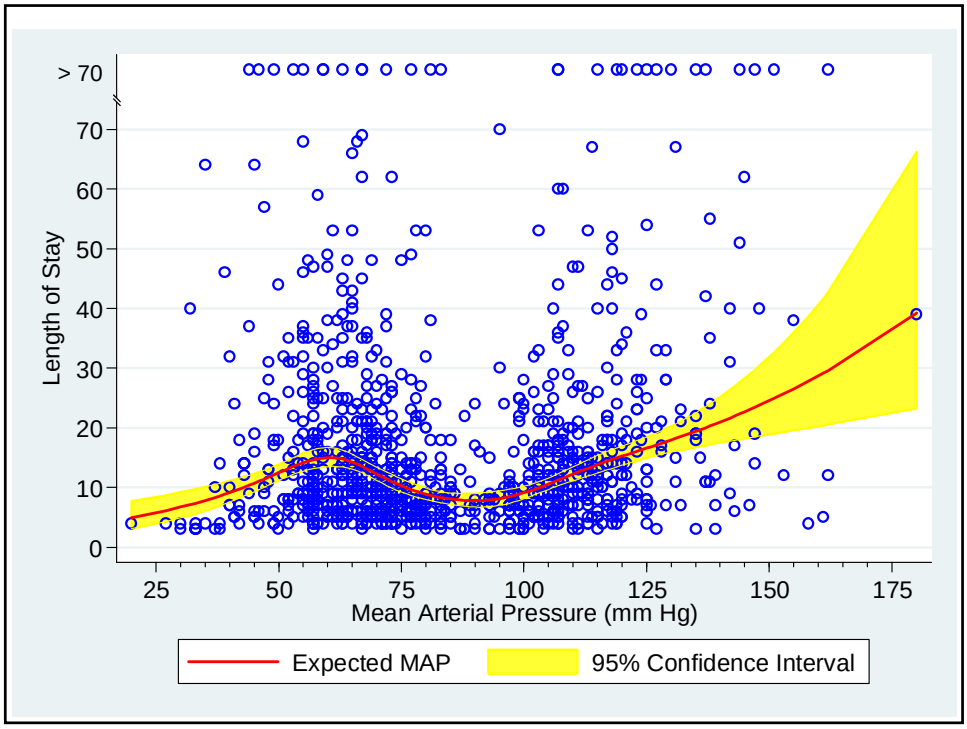
Note that 4.72% of the studentized residuals are greater than 2.




```
. *
. *  Plot expected LOS against MAP on a linear scale.
. *  Truncate LOS > 70.
. *
. generate e_los = exp(y_hat)

. generate lb_los = exp(lb)

. generate ub_los = exp(ub)

. generate truncated_los = los

. * Data > Create or change data > Create new variable
. replace truncated_los = 80 if los > 70
(29 real changes made)

. twoway rarea lb_los ub_los map , color(yellow)                  ///
>     ||  scatter truncated_los map , symbol(Oh)  color(blue)     ///
>     || line e_los map, color(red) lwidth(medthick)              ///
>     || rline lb_los ub_los map , color(yellow)                  /// {16}
>          lwidth(thin thin)                                      ///
>     , xlabel(25 (25) 175) xmtick(30 (5) 170)                    ///
>       ylabel(0 (10) 70) ytitle(Length of Stay)                  ///
>       legend(order(3 "Expected MAP"                             ///
>           1 "95% Confidence Interval") rows(1))
```

Length of stay is badly skewed. In order to keep the range of the y-axis reasonable lets truncate this distribution at 70. We will use a graphics editor to fix the resulting graph.

{15} The scatter plot is so dense that it often obscures the 95% confidence band. Plotting the outline of this band on top of the scatter plot makes it easier to see.



22. What we have covered.

- Extend simple linear regression to models with multiple covariates
- Meaning of parameters in a multiple linear regression model
- Exploratory data analysis
  - Density distribution sunflower plots for displaying high density bivariate data
  - Matrix scatterplots
- Additive models and models with interaction terms
- Building and interpreting complex linear models
- Stepwise methods of building regression models
- Model validation: Evaluating residuals, leverage and influence
- Goodness of model fit vs. model complexity: Using AIC and BIC to choose a good model.
- Restricted cubic splines: Using multiple linear regression to model non-linear relationships between continuous variables.
- Calculating 95% confidence bands for regression curves from restricted cubic spline models.

Cited References

Levy D, National Heart Lung and Blood Institute., Center for Bio-Medical Communication. *50 Years of Discovery : Medical Milestones from the National Heart, Lung, and Blood Institute's Framingham Heart Study*. Hackensack, N.J.: Center for Bio-Medical Communication Inc.; 1999.

Knaus,W.A., Harrell, F.E., Jr., Lynn, J., Goldman, L., Phillips, R.S., Connors, A.F., Jr. et al. The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Study to understand prognoses and preferences for outcomes and risks of treatments. *Ann Intern Med.* 1995; 122:191-203.

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.

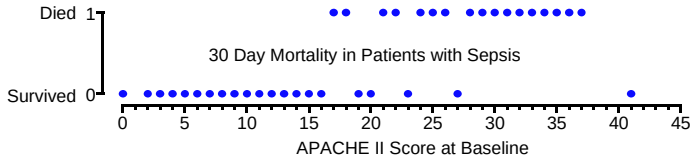
III. INTRODUCTION TO LOGISTIC REGRESSION

- Simple logistic regression: Assessing the effect of a continuous variable on a dichotomous outcome
- How logistic regression parameters affect the probability of an event
- Probability, odds and odds ratios
- Generalized linear models: The relationship between linear and logistic regression
- Confidence intervals for proportions
- Plotting probability of death with 95% confidence bands as a function of a continuous risk factor
- Review of classic 2x2 case-control studies
- Analyzing case-control studies with logistic regression

1. Simple Logistic Regression

a) Example: APACHE II Score and Mortality in Sepsis

The following figure shows 30 day mortality in a sample of septic patients as a function of their baseline APACHE II Score. Patients are coded as 1 or 0 depending on whether they are dead or alive in 30 days, respectively.



We wish to predict death from baseline APACHE II score in these patients.

Let $\pi(x)$ be the probability that a patient with score x will die.

Note that linear regression would not work well here since it could produce probabilities less than zero or greater than one.

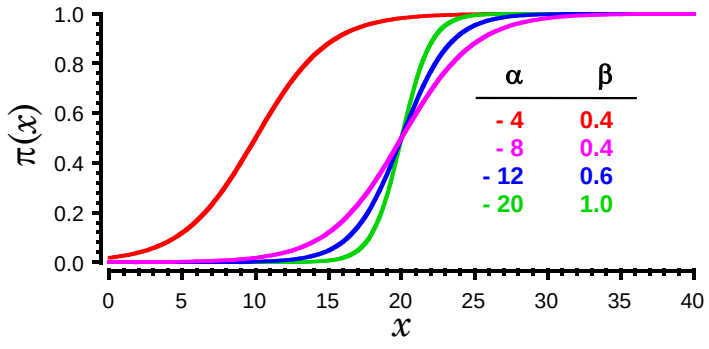
b) Sigmoidal family of logistic regression curves

Logistic regression fits probability functions of the following form:

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x)) \qquad \{3.1\}$$

This equation describes a family of sigmoidal curves, three examples of which are given below.

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$



c) Parameter values and the shape of the regression curve

For now assume that $\beta > 0$.

For negative values of x, $\exp(\alpha+\beta x) \to 0$ as $x \to -\infty$
and hence $\pi(x) \to 0/(1+0) = 0$

For very large values of x, $\exp(\alpha+\beta x) \to \infty$ and hence
$\pi(x) \to \infty/(1+\infty) = 1$

When $x = -\alpha/\beta$, $\alpha+\beta x = 0$ and hence $\pi(x) = 1/(1+1) = 0.5$

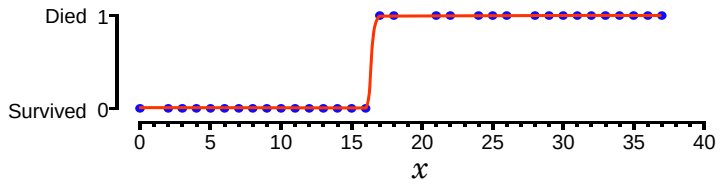
The slope of $\pi(x)$ when $\pi(x)$=.5 is $\beta/4$.

Thus β controls how fast $\pi(x)$ rises from 0 to 1.

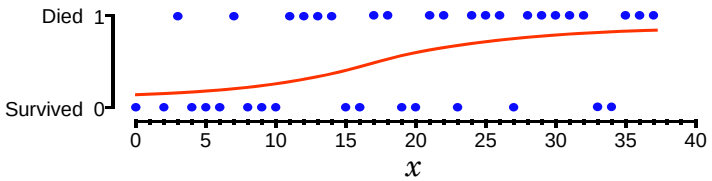
For given β, α controls were the 50% survival point is located.

We wish to choose the best curve to fit the data.

Data that has a sharp survival cut off point between patients who live or die should have a large value of β.



Data with a lengthy transition from survival to death should have a low value of β.



d) The probability of death under the logistic model

This probability is

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

Hence $1 - \pi(x) =$ probability of survival

$$= \frac{1 + \exp(\alpha + \beta x) - \exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

$$= 1/(1 + \exp(\alpha + \beta x))$$, and the odds of death is

$$\pi(x)/(1 - \pi(x)) = \exp(\alpha + \beta x)$$

The log odds of death equals

$$\log(\pi(x)/(1 - \pi(x)) = \alpha + \beta x \qquad \textbf{\{3.2\}}$$

e) The logit function

For any number π between 0 and 1 the logit function is defined by

$$\text{logit}(\pi) = \log(\pi / (1 - \pi))$$

Let $d_i = \begin{cases} 1\text{: } i^{\text{th}} \text{ patient dies} \\ 0\text{: } i^{\text{th}} \text{ patient lives} \end{cases}$

x_i be the APACHE II score of the i^{th} patient

Then the expected value of d_i is

$$E(d_i) = \pi(x_i)$$

Thus we can rewrite the **logistic regression equation** {3.1} as

$$\text{logit}(E(d_i)) = \alpha + \beta x_i \qquad \textbf{\{3.3\}}$$

2. The Binomial Distribution

Let

m be the number of people at risk of death

d be the number of deaths

π be the probability that any patient dies.

The death of one patient has no effect on any other.

Then d has a **binomial distribution** with

parameters m and π,

mean $m\pi$, and

variance $m\pi(1-\pi)$.

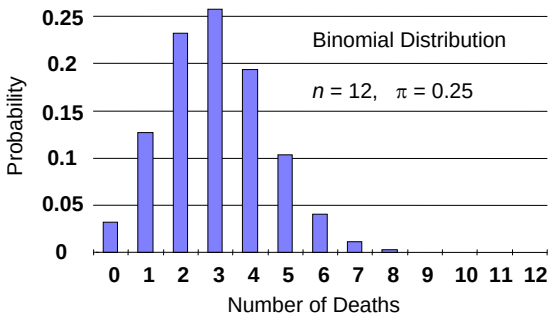
Pr[d deaths]

$$= \frac{m!}{(m-d)!d!} \quad \pi^d(1-\pi)^{(m-d)} \quad : d = 0,1,\cdots,m \qquad \{3.4\}$$

The population mean of any random variable x is also equal to its expected value and is written $E(x)$. Hence

$E(d) = \pi m$ and $E(d/m) = \pi$

For $m = 12$ and $\pi = 0.25$ this distribution is as follows.



A special case of the binomial distribution is when $m = 1$, which is called a Bernoulli distribution.

In this case we can have 0 or 1 deaths with probability 1-π and π, respectively.

The complete logistic regression model for the sepsis data is specified as follows

d_i has a binomial distribution with 0 or 1 failures and probability of failure $\pi(x_i) = E(d_i)$

$E(d_i)$ is determined by $\text{logit}(E(d_i)) = \alpha + \beta x_i$

3. Generalized Linear Models

Logistic regression is an example of a generalized linear model. These models are defined by three attributes: The distribution of the model's random component, its linear predictor, and its link function. For logistic regression these are defined as follows.

a) The random component

d_i *is* the **random component** of the model. In logistic regression, d_i has a binomial distribution obtained from m_i trials with mean $E(d_i)$. (In the sepsis example, $m_i = 1$ for all i.)

Stata refers to the distribution of the random component as the distributional family.

b) The linear predictor

$\alpha + x_i\beta$ is called the **linear predictor**

c) The link function

$E(d_i)$ is related to the linear predictor through a **link function**. Logistic regression uses a logit link function

$$\text{logit}(E(d_i)) = \alpha + x_i\beta$$

4. Contrast Between Logistic and Linear Regression

In linear regression, the expected value of y_i given x_i is

$$E(y_i) = \alpha + \beta x_i \text{ for } i = 1,2,\ldots,n$$

y_i	has a normal distribution with standard deviation σ.
y_i	is the random component of the model, which has a normal distribution.
$\alpha + \beta x_i$	is the linear predictor.

The link function is the identity function $E(y_i) = I(E(y_i))$.

5. Maximum Likelihood Estimation

In linear regression we used the method of least squares to estimate regression coefficients.

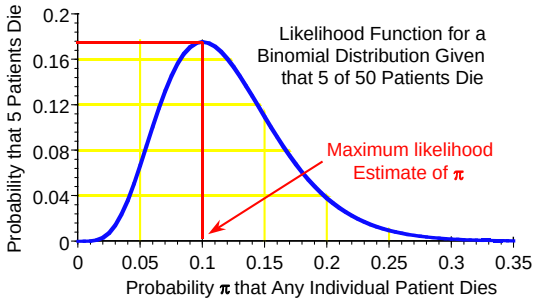
In generalized linear models we use another approach called maximum likelihood estimation.

Suppose that 5 of 50 AIDS patients die in one year. We wish to estimate π, the probability of death for these patients.

We assume that the number of deaths has a binomial distribution obtained from m= 50 patients with probability of death π for each patient.

Let $L(\pi \mid d = 5)$ be the probability of the observed outcome (5) given different values of π.

$L(\pi \mid d = 5)$ is called a likelihood function and looks like this.



The maximum likelihood estimate of π is the value of π that assigns the greatest probability to the observed outcome.

In this example, $\hat{\pi} = 0.1$

Note that if $\pi = \hat{\pi} = 0.1$ that $E(d) = 50 \times 0.1 = 5 = d$

Thus, in this example, the maximum likelihood estimate of π is that value that sets the expected number of deaths equal to the observed number of deaths.

In general, maximum likelihood estimates do not have simple closed solutions, but must be solved interactively using numerical methods. This, however, is not a serious drawback given ubiquitous and powerful desktop computers.

a) Variance of maximum likelihood parameter estimates

It can be shown that when a maximum likelihood estimate is based on large number of patients, its variance is approximately equal to

-1/C, where C is the expected curvature of the likelihood surface at $\hat{\pi}$

6. Logistic Regression with glm

a) Example: APACHE II score and mortal outcome

```
. *  4.11.Sepsis.log
. *
. *  Simple logistic regression of mortal status at 30 days (fate) against
. *  baseline APACHE II score (apache) in a random sample of septic patients
. *
. use C:\\WDDtext\4.11.Sepsis.dta, clear
. summarize fate apache

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
        fate |        38    .4473684    .5038966          0          1
      apache |        38    19.55263    11.30343          0         41
```

```
. codebook
apache ---------------------------------------------- APACHE II Score at Baseline
                  type:  numeric (byte)

                 range:  [0,41]                       units:  1
         unique values:  38                   coded missing:  0 / 38

                  mean:   19.5526
             std. dev:   11.3034

           percentiles:        10%       25%       50%       75%       90%
                                4        10      19.5        29        35



fate ------------------------------------------------- Mortal Status at 30 Days
                  type:  numeric (byte)
                 label:  fate

                 range:  [0,1]                        units:  1
         unique values:  2                    coded missing:  0 / 38

            tabulation:  Freq.   Numeric  Label
                           21         0  Alive
                           17         1  Dead
```

```
. glm fate apache, family(binomial) link(logit)                          {1}

Iteration 0:   log likelihood = -15.398485
Iteration 1:   log likelihood =   -14.9578
Iteration 2:   log likelihood = -14.956086
Iteration 3:   log likelihood = -14.956085

Generalized linear models                         No. of obs      =         38
Optimization     : ML: Newton-Raphson             Residual df     =         36
                                                  Scale param     =          1
Deviance         =  29.91217061                   (1/df) Deviance =  .8308936
Pearson          =  66.34190718                   (1/df) Pearson  =  1.842831

Variance function: V(u) = u*(1-u)                 [Bernoulli]              {2}
Link function    : g(u) = ln(u/(1-u))             [Logit]
Standard errors  : OIM

Log likelihood   = -14.95608531                   AIC             =  .8924255
BIC              = -101.0409311

------------------------------------------------------------------------------
        fate |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
-------------+----------------------------------------------------------------
      apache |   .2012365   .0608998      3.304   0.001       .0818752    .3205979
       _cons |  -4.347807   1.371609     -3.170   0.002      -7.036111   -1.659503
------------------------------------------------------------------------------
. predict logodds, xb                                                    {3}
. generate prob = exp(logodds)/(1 + exp(logodds))                        {4}
. list apache fate logodds prob in 1/3                                   {5}
```

{1} This *glm* command regresses ***fate*** against ***apache*** using a generalized linear model. The *family* and *link* options specify that the **random component** of the model is **binomial** and the **link function** is **logit**. In other words, a **logisitic** model is to be used.

{2} When there is only one patient per record Stata refers to the binomial distribution as a **Bernoulli** distribution. Along with the **logit** link function this implies a **logisitc** regression model.

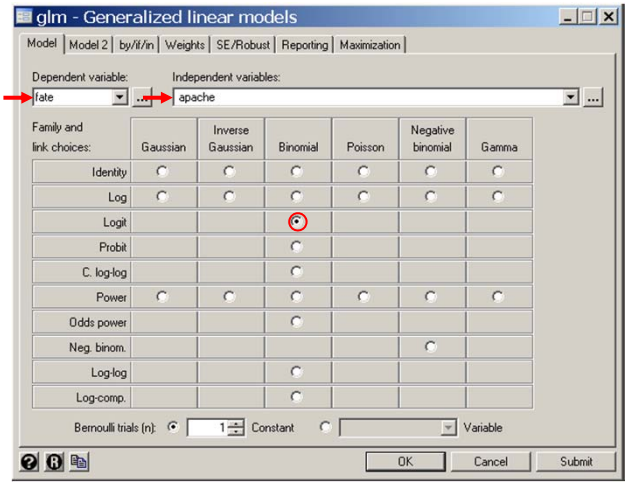
{3} The *xb* option of the ***predict*** command specifies that the **linear predictor** will be evaluated for each patient and stored in a variable named ***logodds.***

Recall that *predict* is a **post estimation** command whose meaning is determined by the latest estimation command, which in this example is *glm*.

{4} ***prob*** equals the estimated **probability** that a patient will **die**. It is calculated using the equation

$$\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

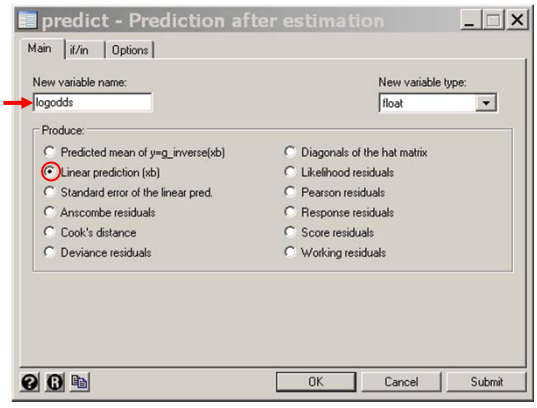
{5} The ***in*** modifier specifies that the first through third record are to be listed.




```
. predict logodds, xb                                                {3}
```

{3} The *xb* option of the ***predict*** command specifies that the **linear predictor** will be evaluated for each patient and stored in a variable named ***logodds.***

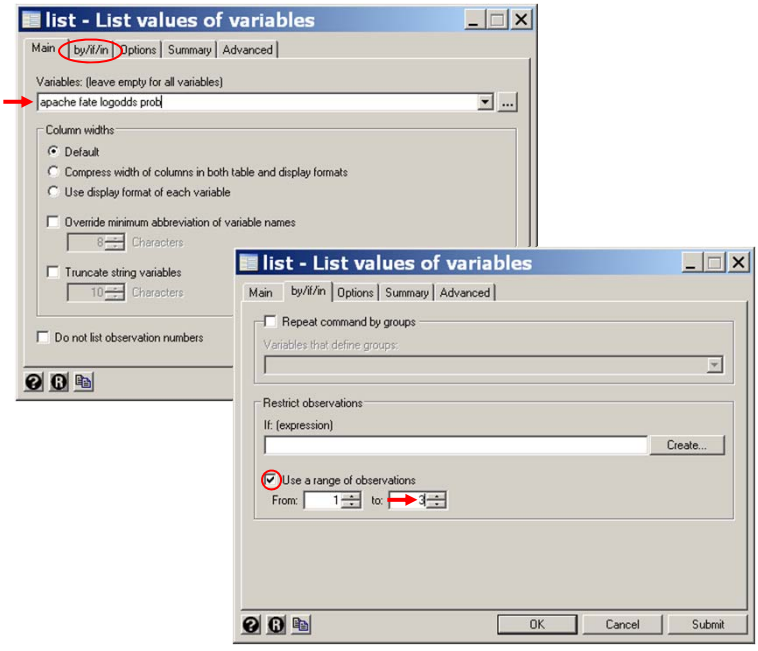
Recall that *predict* is a **post estimation** command whose meaning is determined by the latest estimation command, which in this example is *glm*.



```
. generate prob = exp(logodds)/(1 + exp(logodds))                    {4}
. * Data > Describe data > List data
. list apache fate logodds prob in 1/3                               {5}
```

{4} ***prob*** equals the estimated **probability** that a patient will **die**. It is calculated using equation 3.1.

{5} The ***in*** modifier specifies that the first through third record are to be listed.



	apache	fate	logodds	prob
1.	16	Alive	-1.128022	.2445263
2.	25	Dead	.6831065	.6644317
3.	19	Alive	-.5243126	.3718444

{6}

```
. sort apache
. * Variables Manager
. label variable prob "Probability of Death"
```

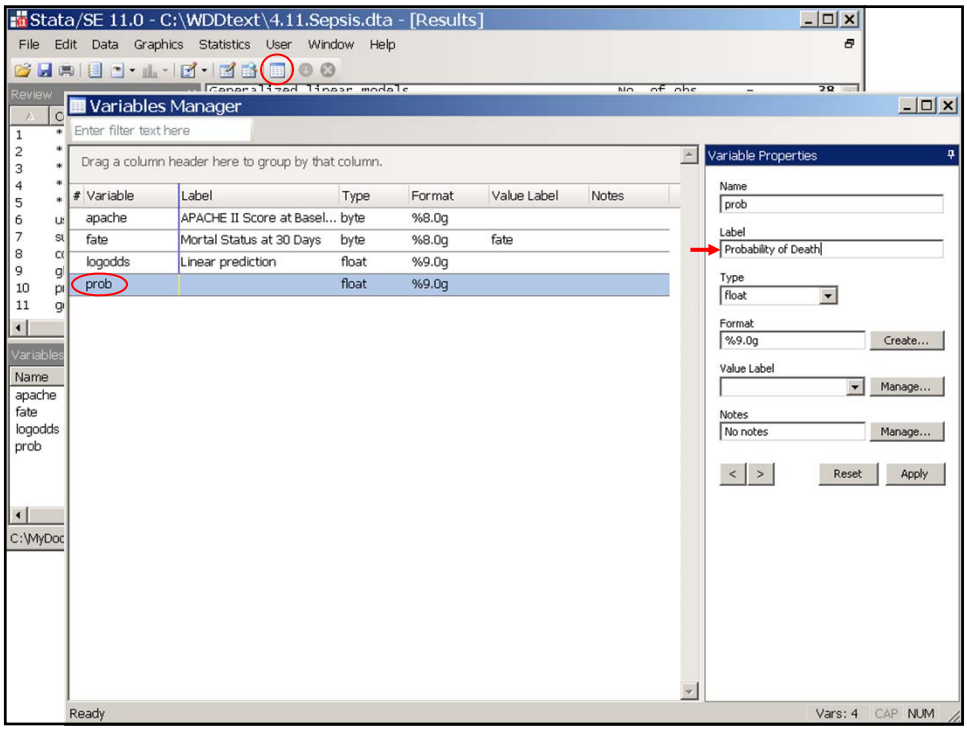
{7}

{7} Assign the label *Probability of Death* to the variable *prob*.

{6} The first patient has an APACHE II score of 16. Hence the estimated linear predictor for this patient is $logodds = \alpha + x_i\beta =$ *_cons* + 16×*apache* = -4.3478 + 16×0.2012 = -1.1286. The second patient has apache = 25 giving logodds = -4.3478 + 25×0.2012 = 0.6831.

For the first patient

$$\text{prob} = \exp(\alpha+\beta x)/(1+\exp(\alpha+\beta x))$$
$$= \exp(logodds)/(1+\exp(logodds))$$
$$= \exp(-1.128)/(1+\exp(-1.128)) = 0.2445$$



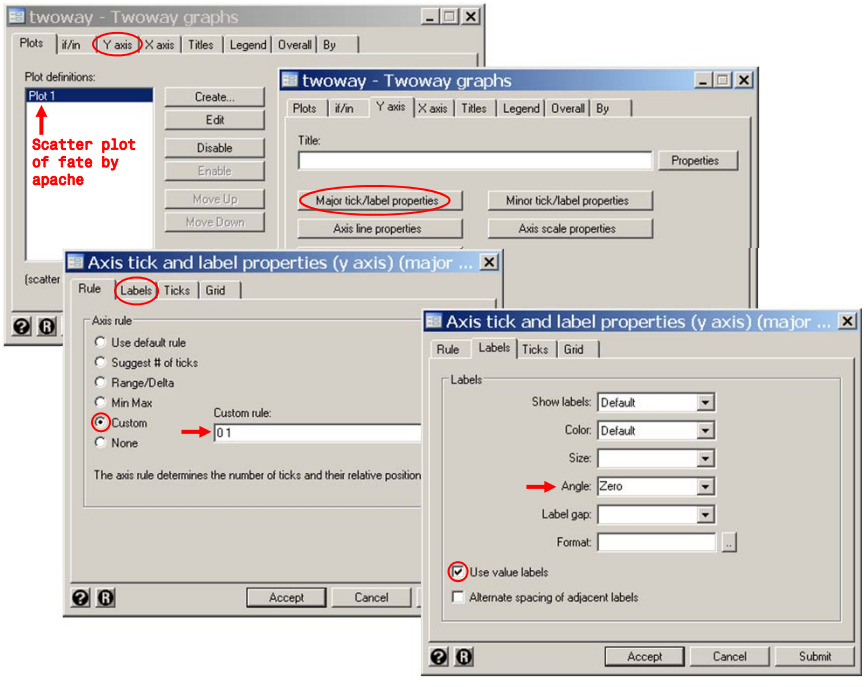
```
. scatter fate apache                                              ///
>     , ylabel(0 1, valuelabel angle(0)) yscale(titlegap(-8)) ///     {8,9}
>     || line prob apache, yaxis(2) xlabel(0(10)40)                    {10}
```

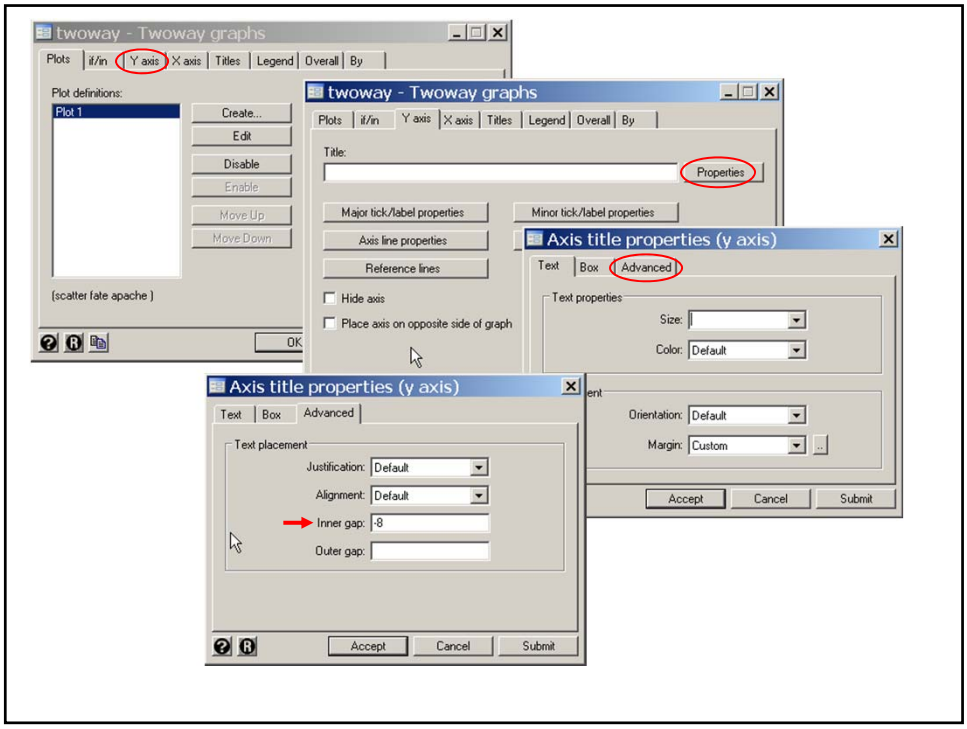
{8} **valuelabel** and **angle** are suboptions of the *ylabel* option. The labels for the *y*-axis are at 0 and 1. *valuelabel* indicates that the value labels of *fate* are to be used. That is, *Alive* and *Dead*.

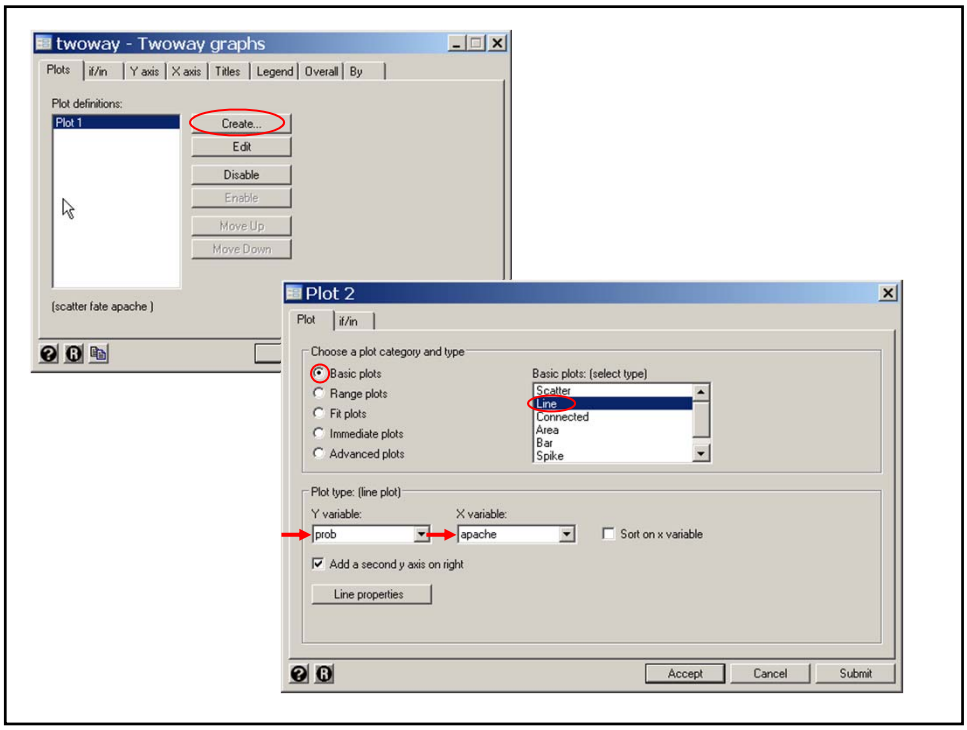
angle specifies the angle at which the labels are written; an angle of 0 means that the labels will be written parallel to the *x*-axis.

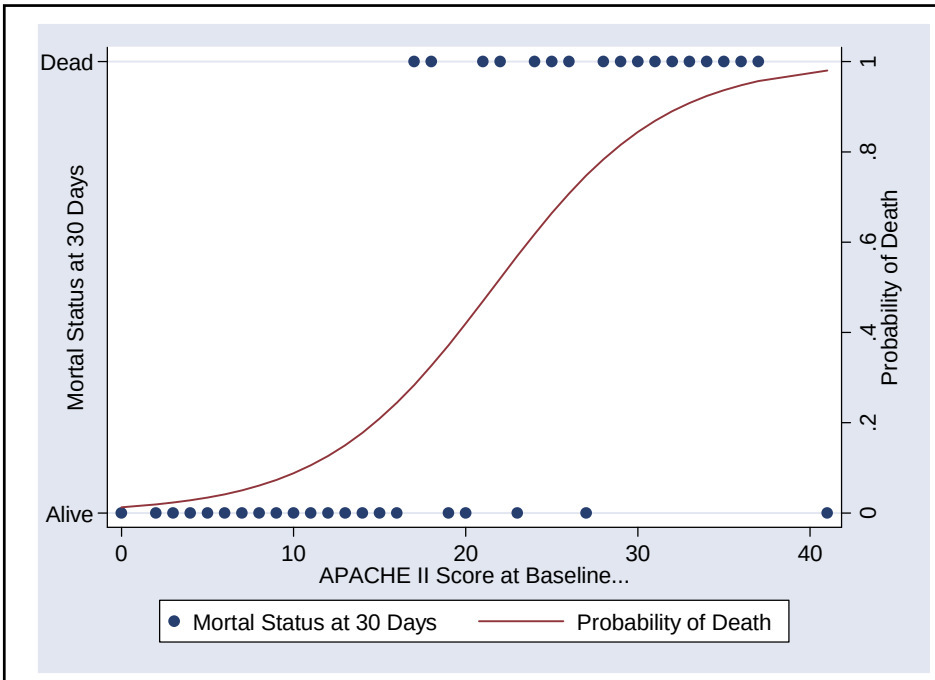
{9} *yscale(titlegap(-8))* specifies how close the title of the *y*-axis is to the axis itself. The default, *titlegap(0)* would place the title just to the left of the labels *Dead* and *Alive*.

{10} **yaxis(2)** indicates that the y-axis for the graph of *prob* vs. *apache* is to be drawn on the right.




twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
Move Up
Move Down
(scatter fate apache)
OK
twoway - Twoway graphs
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis title properties (y axis)
Text
Box
Advanced
Text properties
Size:
Color:
Default
Orientation:
Margin:
Custom
Accept
Cancel
Submit
Axis title properties (y axis)
Text
Box
Advanced
Text placement
Justification:
Default
Alignment:
Default
Inner gap:
-8
Outer gap:
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Edit
Disable
Enable
Move Up
Move Down
(scatter fate apache)
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Area
Bar
Spike
Plot type: (line plot)
Y variable:
X variable:
prob
apache
Sort on x variable
Add a second y axis on right
Line properties
Accept
Cancel
Submit



7. Odds Ratios and the Logistic Regression Model

a) Odds ratio associated with a unit increase in *x*

The log odds that patients with APACHE II scores of x and $x + 1$ will die are

$$\text{logit}(\pi(x)) = \alpha + \beta x \qquad \textbf{\{3.5\}}$$

and

$$\text{logit}(\pi(x+1)) = \alpha + \beta(x+1) = \alpha + \beta x + \beta \qquad \textbf{\{3.6\}}$$

respectively.

subtracting {3.5} from {3.6} gives $\beta = \text{logit}(\pi(x+1)) - \text{logit}(\pi(x))$

$$\beta = \text{logit}\,(\pi(x+1)) - \text{logit}(\pi(x))$$

$$= \log\left(\frac{\pi(x+1)}{1-\pi(x+1)}\right) - \log\left(\frac{\pi(x)}{1-\pi(x)}\right)$$

$$= \log\left(\frac{\pi(x+1)/(1-\pi(x+1))}{\pi(x)/(1-\pi(x))}\right)$$

and hence

exp(β) is the odds ratio for death associated with a unit increase in *x*.

A property of logistic regression is that this ratio remains constant for all values of *x*.

8. 95% Confidence Intervals for Odds Ratio Estimates

In our sepsis example the parameter estimate for *apache* (β) was 0.2012 with a standard error or 0.0609. Therefore, the odds ratio for death associated with a unit rise in APACHE II score is

exp(0.2012) = 1.223

with a 95% confidence interval of $\exp(0.2012 \pm 1.96 * 0.0609)$

(1.223exp(-1.96×0.0609), 1.223exp(1.96×0.0609))

= (1.09, 1.38).

9. 95% Confidence Interval for $\pi[x]$

Let $\sigma^2_{\hat{\alpha}}$ and $\sigma^2_{\hat{\beta}}$ denote the variance of $\hat{\alpha}$ and $\hat{\beta}$.
Let $\sigma_{\hat{\alpha}\hat{\beta}}$ denote the covariance between $\hat{\alpha}$ and $\hat{\beta}$.

Then it can be shown that the standard error of is

$$\text{se}\left[\hat{\alpha}+\hat{\beta}x\right] = \sqrt{\sigma^2_{\hat{\alpha}}+2x\sigma_{\hat{\alpha}\hat{\beta}}+x^2\sigma^2_{\hat{\beta}}}$$

A 95% confidence interval for $\alpha+\beta x$ is

$$\hat{\alpha}+\hat{\beta}x \pm 1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]$$

A 95% confidence interval for $\alpha+\beta x$ is

$$\hat{\alpha}+\hat{\beta}x \pm 1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]$$

Hence, a 95% confidence interval for $\pi[x]$ is $\left(\hat{\pi}_L[x],\hat{\pi}_U[x]\right)$, where

$$\hat{\pi}_L[x]=\frac{\exp\left[\hat{\alpha}+\hat{\beta}x-1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}{1+\exp\left[\hat{\alpha}+\hat{\beta}x-1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}$$

and

$$\hat{\pi}_U[x]=\frac{\exp\left[\hat{\alpha}+\hat{\beta}x+1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}{1+\exp\left[\hat{\alpha}+\hat{\beta}x+1.96\times\text{se}\left[\hat{\alpha}+\hat{\beta}x\right]\right]}$$

10. 95% Confidence Intervals for Proportions

It is useful to be able to estimate a 95% confidence interval for the proportion d_i/m_i in the sepsis study.

Let d be the number of deaths that occur in m patients,

π be the probability that an individual dies..

Then d/m has mean π and standard error $s(\pi) = \sqrt{\pi(1-\pi)/m}$

Estimating π by $\hat{\pi} = d/m$ gives $s(\hat{\pi}) = \sqrt{\hat{\pi}(1-\hat{\pi})/m}$

as the estimated standard error of $\hat{\pi}$

The distribution of $\hat{\pi}$ is approximately normal as long as $\hat{\pi}$ is not too close to 0 or 1 and m is fairly large. This approximation gives a Wald 95% confidence interval for π of

$$\hat{\pi} \pm 1.96 s(\hat{\pi})$$

This estimate works poorly when $\hat{\pi}$ is near 0 or 1. Note that it is possible for this confidence interval to contain values that are less than 0 or greater than 1.

The 100(1-α)% Wald Confidence interval is

$$\hat{\pi} \pm z_{\alpha/2} s(\hat{\pi}) \qquad \text{(recall that } z_{.025} = 1.96\text{)}$$

This interval consists of all π for which

$$-z_{\alpha/2} \le (\hat{\pi} - \pi)/s(\hat{\pi}) \le z_{\alpha/2}$$

Wilson Confidence Interval for a Proportion.

A better 100(1-α)% confidence interval due to Wilson is given by all values of π for which

$$-z_{\alpha/2} \le (\hat{\pi} - \pi)/s(\pi) \le z_{\alpha/2} \qquad \{3.7\}$$

This interval differs from the Wald Interval in that the denominator is $s(\pi)$ rather than $s(\hat{\pi})$. This makes a big difference when π is near 0 or 1.

Equation {3.7} can be rewritten as a complex but unedifying function of d, m and $z_{\alpha/2}$

```
. * proportions.log
. *
. * Illustrate Wald, Wilson and exact confidence intervals
. *
. use proportions.dta
. list

     +-----------------+
     | fate   patients |
     |-----------------|
  1. |    0         10 |
  2. |    1         10 |
     +-----------------+
```

Here is data on 20 patients grouped into two records with 10 patients per record.

Half of these patients live (fate = 0) and the other half die (fate = 1).

```
* Statistics > Summaries, tables ... > Summary ... > Confidence intervals
. ci fate [freq = patients], binomial wald                                  {1}

                                                      -- Binomial Wald ---
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
        fate |         20          .5    .1118034        .2808694    .7191306

. ci fate [freq = patients], binomial wilson                                {2}

                                                      ------ Wilson ------
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+------------------------------------------------------------------
        fate |         20          .5    .1118034         .299298     .700702
```

{1} This **ci** command calculated confidence intervals for the proportion of patients who die (fate = 1). **[freq=patients]** ensures that each record contributes the number of patients indicated by the *patients* variable. (Without this command modifier, each record would count as a single observation.)

binomial specifies that fate is a dichotomous variable. It must be specified whenever Wald or Wilson confidence intervals are required. **wald** indicates that Wald confidence intervals are to be calculated.

{2} **wilson** indicates that Wilson confidence intervals are to be calculated.

These confidence intervals are quite close to each other.

```
. replace patients = 18 in 1
(1 real change made)

. replace patients = 2 in 2
(1 real change made)

. list

     +-----------------+
     | fate   patients |
     |-----------------|
  1. |    0         18 |
  2. |    1          2 |
     +-----------------+
```

Suppose that the mortality rate is 10%

```
. ci fate [freq = patients], binomial wald

                                                      -- Binomial Wald ---
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        fate |         20          .1     .067082                0    .2314784*

(*) The Wald interval was clipped at the lower endpoint

. ci fate [freq = patients], binomial wilson

                                                      ------ Wilson ------
    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
        fate |         20          .1     .067082        .0278665    .3010336
```

The Wald interval is much narrower than the Wilson and would extend below zero if Stata did not clip it at zero.

```
. return list                                                    {3}

scalars:
                  r(ub) =  .3010336452284873
                  r(lb) =  .0278664812137682
                  r(se) =  .0670820393249937
                r(mean) =  .1
                   r(N) =  20

. display r(ub)                                                  {4}
.30103365
```

{3} Stata commands store most of their output were they can be used by other commands. This feature greatly extends the power and flexibility of this software. The **return list** command lists some of these values.

{4} This **display** command displays the upper bound of the confidence interval calculated by the last **ci** command.

Baseline APACHE II Score	Number of Patients	Number of Deaths	Baseline APACHE II Score	Number of Patients	Number of Deaths
0	1	0	20	13	6
2	1	0	21	17	9
3	4	1	22	14	12
4	11	0	23	13	7
5	9	3	24	11	8
6	14	3	25	12	8
7	12	4	26	6	2
8	22	5	27	7	5
9	33	3	28	3	1
10	19	6	29	7	4
11	31	5	30	5	4
12	17	5	31	3	3
13	32	13	32	3	3
14	25	7	33	1	1
15	18	7	34	1	1
16	24	8	35	1	1
17	27	8	36	1	1
18	19	13	37	1	1
19	15	7	41	1	0

Example: APACHE II Score & Mortality in Sepsis

The Ibuprofen and Sepsis Trial contained 454 patients with known baseline APACHE II scores (Bernard et al. 1997). The 30 day mortal outcome for these patients is summarized on the right.

11. Logistic Regression with Grouped Response Data

Suppose that there are m_i patients with covariate x_i.

Let d_i be the number of deaths in these m_i patients.

Then d_i has a binomial distribution with mean $m_i\pi(x_i)$ and hence $E(d_i/m_i) = \pi(x_i)$.

Thus the logistic model becomes

$\text{logit}(E(d_i/m_i)) = \alpha + \beta x_i$

```
. *  4.18.Sepsis.Wilson.log
. *
. *  Simple logistic regression of mortality against APACE score in the
. *  Ibuprofen in Sepsis study (Bernard et al. 1997).  There are two
. *  records in 4.18.Sepsis.Weighted.dta for each observed APACE score.
. *    apache = an APACHE II score at baseline
. *    fate   = 0 or 1 indicating patients who were alive or dead after
. *                    30 days, respectively
. *    n      = number of study subjects with a given fate and APACE score.
. *
. use 4.18.Sepsis.Weighted.dta, clear

. list if apache==6 | apache==7

     +--------------------+
     | apache   fate    n |
     |--------------------|
 11. |      6      0   11 |
 12. |      6      1    3 |
 13. |      7      0    8 |
 14. |      7      1    4 |
     +--------------------+
```

We need to calculate the observed mortality rate and its associated confidence interval for each APACHE score.

There were 37 distinct scores.

We could issue 47 distinct **ci** commands and transcribe the confidence intervals back into Stata.

This would be tedious. Fortunately it is unnecessary.

```
. *
. *  Collapse data to one record per APACHE score.
. *  Calculate observed mortality rate for each score and its
. *  Wilson 95% confidence interval.
. *
. * Statistics > Other > Collect statistics for a command across a by list
. statsby, by(apache): ci fate [freq=n], binomial wilson                  {1}
(running ci on estimation sample)

      command:  ci fate [fweight= n], binomial wilson
           ub:  r(ub)
           lb:  r(lb)
           se:  r(se)
         mean:  r(mean)
            N:  r(N)
           by:  apache

Statsby groups
----+--- 1 ---+--- 2 ---+--- 3 ---+--- 4 ---+--- 5
.....................................
```

{1} The **statsby** command can be used in combination with most other Stata commands. It executes the command to the right of the colon for each unique combination of values of the variable(s) specified by the **by** option. This command executes

```
ci fate [freq=n], binomial wilson
```

separately for each unique value of **apache**. The data in memory is replaced by new data with one record for each distinct value of apache. Output from each command is also stored with the indicated variable names.

```
. list if apache==6 | apache==7

     +-------------------------------------------------------+
     | apache         ub         lb         se       mean    N |
     |-------------------------------------------------------|
  6. |      6   .4758923   .0757139   .1096642   .2142857   14 |
  7. |      7   .6093779   .1381201   .1360828   .3333333   12 |          {2}
     +-------------------------------------------------------+
. generate patients = N

. generate p = mean

. generate deaths = p*patients                                          {3}
```

{2} There is now only one record for each value of **apache**. The variables **N** and **mean** store the number of patients with the specified value of **apache** and their associated mortality rate, respectively. **ub** and **lb** give the Wilson 95% confidence interval for this rate.

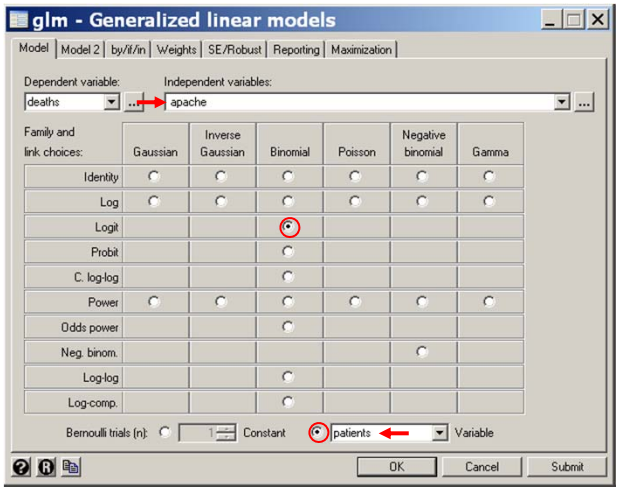
N.B. All other variables that are not specified by the **by** option are lost. Do not use this command with data that you value and have not saved!

{3} **deaths** give the number of patients with the indicated value of **apache** who die.

```
. * Statistics > Generalized linear models > Generalized linear models (GLM)
. glm deaths apache, family(binomial patients) link(logit)              {1}
    .
    .
    .
Generalized linear models                         No. of obs      =         38
Optimization     : ML: Newton-Raphson             Residual df     =         36
                                                  Scale param     =          1
Deviance         =  84.36705142                   (1/df) Deviance =  2.343529
Pearson          =  46.72842945                   (1/df) Pearson  =  1.298012

Variance function: V(u) = u*(1-u/patients)        [Binomial]
Link function    : g(u) = ln(u/(patients-u))      [Logit]
Standard errors  : OIM

Log likelihood   = -60.93390578                   AIC             =  3.312311
BIC              = -46.58605033
------------------------------------------------------------------------------
  deaths |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  apache |   .1156272   .0159997      7.23   0.000       .0842684     .146986
   _cons |  -2.290327   .2765283     -8.28   0.000      -2.832313   -1.748342
------------------------------------------------------------------------------
```

{1} Regress **deaths** against **apache** score. The **bionomial** random component and **logit** link function specify that **logistic** regression is to be used.

family(binomial patients) indicates that each observation describes the outcomes of multiple patients with the same apache score; **patients** records the number of subjects with each score; **deaths** records the number of deaths observed in these subjects.




```
. predict logodds, xb     {2}
. generate e_prob = exp(logodds)/(1+exp(logodds))
. label variable e_prob "Expected Mortality at 30 Days"
. *
. *  Calculate 95% confidence region for e_prob
. *
. predict stderr, stdp
. generate lodds_lb = logodds - 1.96*stderr
. generate lodds_ub = logodds + 1.96*stderr
. generate prob_lb = exp(lodds_lb)/(1+exp(lodds_lb))
. generate prob_ub = exp(lodds_ub)/(1+exp(lodds_ub))
. label variable p "Observed Mortality Rate"
. * Data > Describe data > List data
. list p e_prob prob_lb prob_ub ci95lb ci95ub apache if apache == 20

     +-------------------------------------------------------------------------+
     |        p    e_prob    prob_lb   prob_ub         lb        ub   apache |
     |-------------------------------------------------------------------------|
 20. | .4615385   .505554   .4462291   .564723   .2320607   .708562       20 |
     +-------------------------------------------------------------------------+
```

{2} The **linear predictor** is *logodds* = -2.2903 + .115627**apache*.

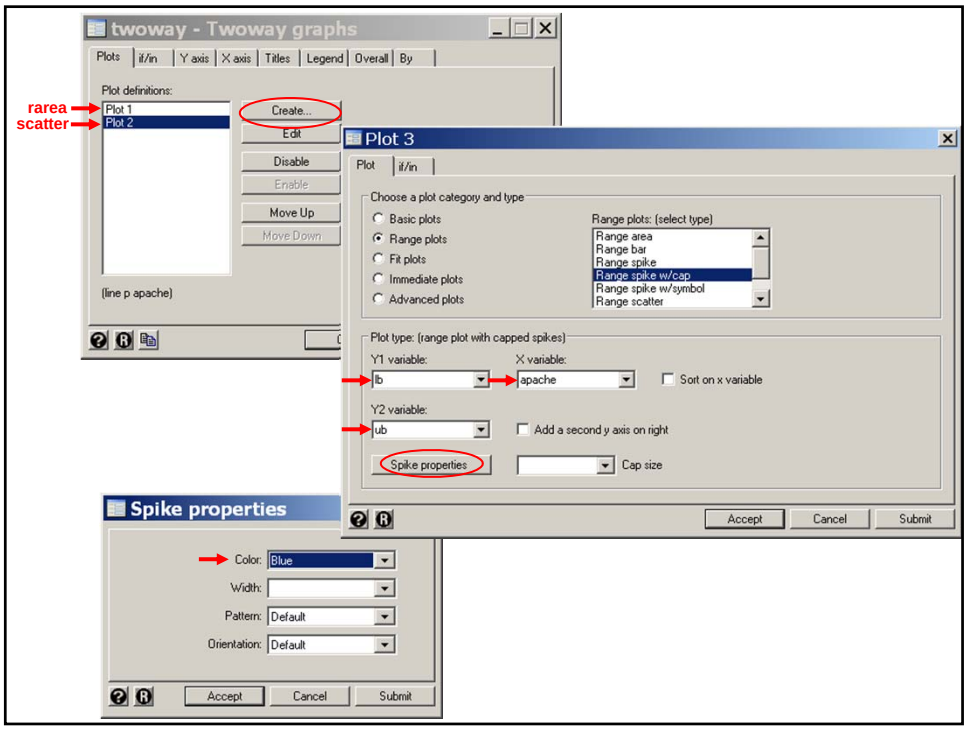
```
. twoway rarea prob_ub prob_lb apache, color(yellow)                    ///
>     || scatter p apache, color(blue)                                  ///
>     || rcap ub lb apache, color(blue)                                 /// {3}
>     || line e_prob apache, yaxis(2)  clwidth(medthick) color(red)     ///
>     ,  ylabel(0(.2)1,labcolor(blue) angle(0))                         /// {4}
>        ytick(0(.1)1, tlcolor(blue))                                   /// {5}
>        ylabel(0(.2)1, axis(2) labcolor(red) angle(0))                 /// {6}
>        ytick(0(.1)1, axis(2) tlcolor(red))                            ///
>        xlabel(0(5)40) xtick(0(1)40)                                   ///
>        ytitle(,axis(2) color(red))                                    ///
>        ytitle(Observed Mortality Rate, color(blue))                   ///
>        legend(order(1 "95% CI from model" 2 3 "95% CI from this obs." 4))
```

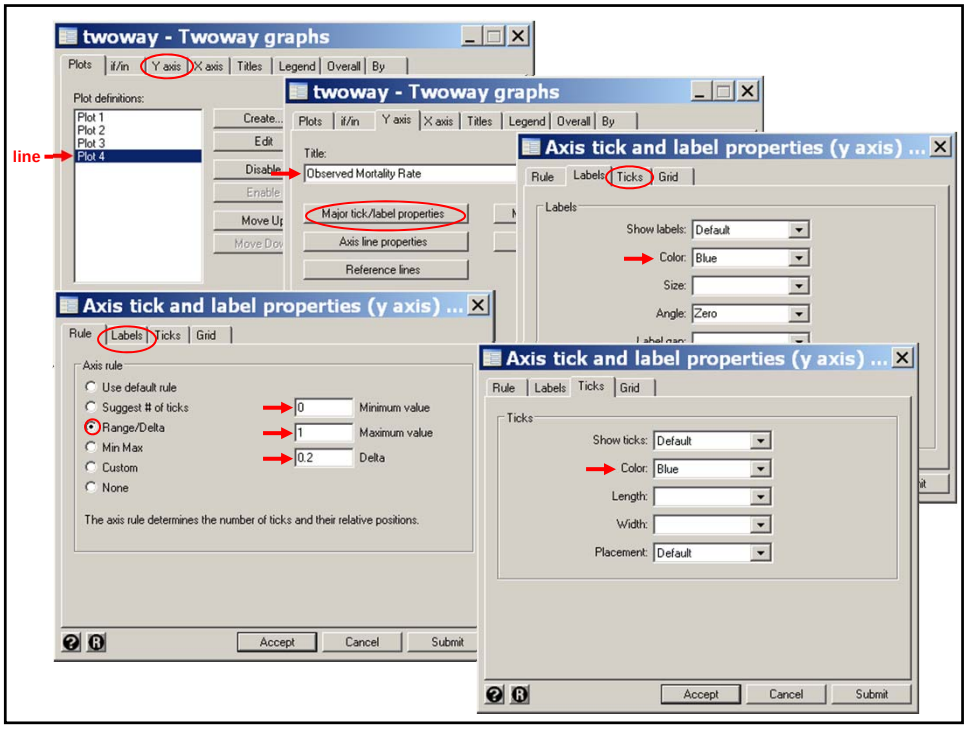
{3} **rcap** plots capped rods (error bars) joining the values of **ub** and **lb** for each value of **apache**.

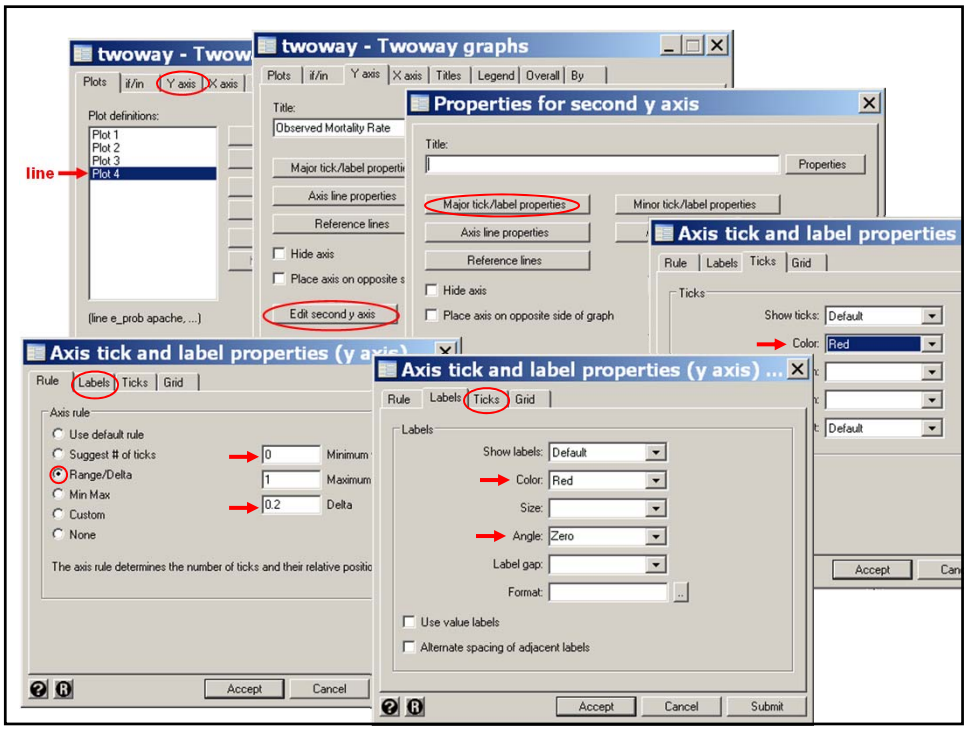
{4} This graph will have two *y*-axes: a left-axis for the observed mortality rate and a right-axis for the expected morbidity rate. Here we color the default (left) axis blue to match the blue scatterplot of observed mortality rates.

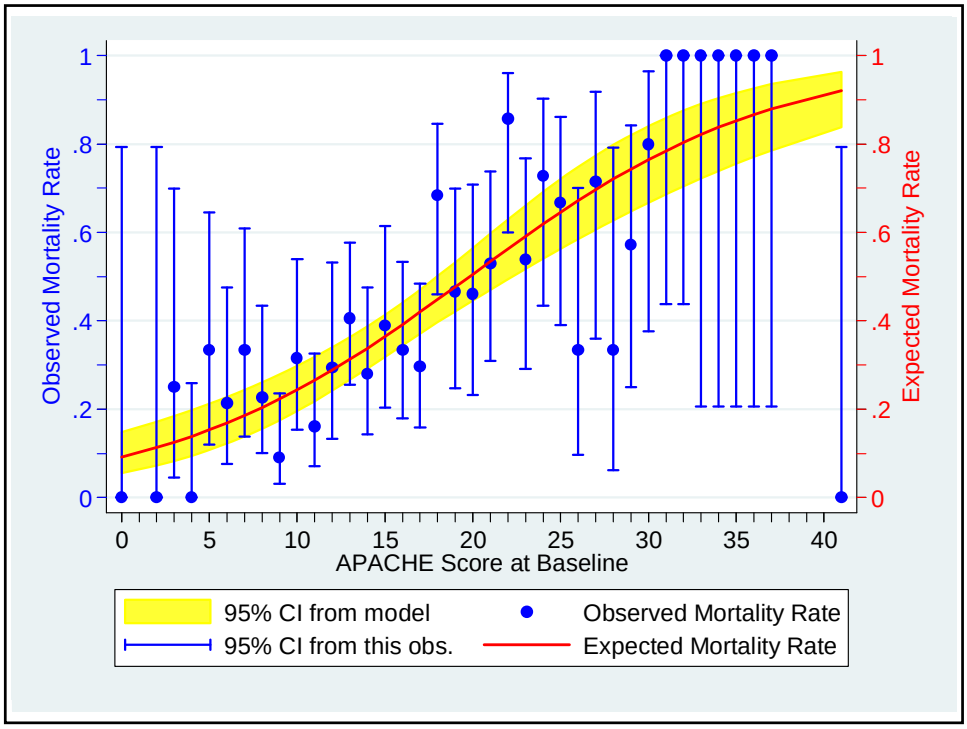
{5} Also, color the tick lines blue on the left axis.

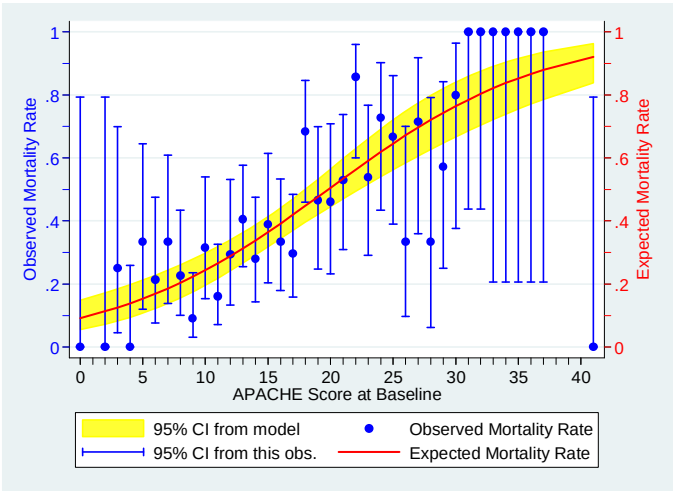
{6} The **axis(2)** suboption indicates that this **ylabel** option refers to the right axis. It is colored red to match the expected mortality curve.




twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Plot 4
line
Create...
Edit
Disable
Enable
Move Up
Move Down
Title:
Observed Mortality Rate
Major tick/label properties
Axis line properties
Reference lines
Axis tick and label properties (y axis) ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
0
1
0.2
Minimum value
Maximum value
Delta
The axis rule determines the number of ticks and their relative positions.
Labels
Show labels:
Default
Color:
Blue
Size:
Angle:
Zero
Ticks
Show ticks:
Length:
Width:
Placement:
Accept
Cancel
Submit


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Plot 3
Plot 4
line
(line e_prob apache, ...)
Title:
Observed Mortality Rate
Major tick/label properti
Axis line properties
Reference lines
Hide axis
Place axis on opposite s
Edit second y axis
Properties for second y axis
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis tick and label properties
Rule
Labels
Ticks
Grid
Show ticks:
Default
Color:
Red
Axis tick and label properties (y axis) ...
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
0
1
0.2
Minimum
Maximum
Delta
The axis rule determines the number of ticks and their relative positic
Labels
Show labels:
Default
Color:
Red
Size:
Angle:
Zero
Label gap:
Format:
Use value labels
Alternate spacing of adjacent labels
Accept
Cancel
Submit





Note that the width of these intervals depends on the number of patients with a given score and the observed mortality rate.

The **blue** error bars in the regression graph give 95% confidence intervals that are derived from the observed mortality rates at each separate APACHE II score. These confidence intervals are not given for scores with zero or 100% mortality. The **yellow shaded region** gives 95% confidence intervals for the expected mortality that are derived from the entire logistic regression.

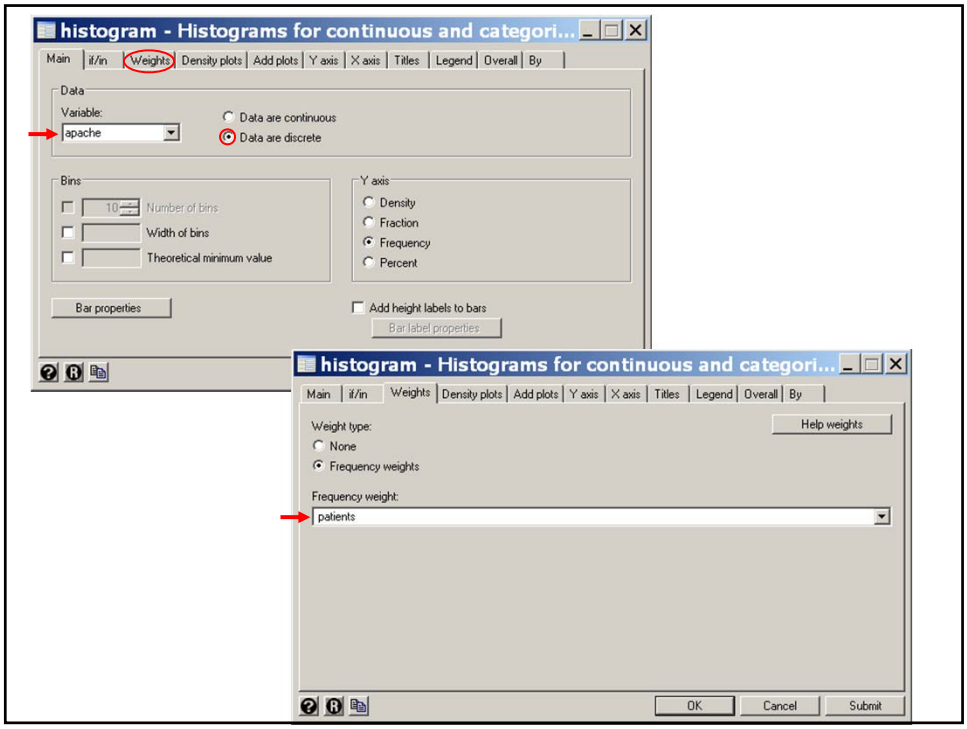
Overall, the fit appears quite good, although the regression curve comes close to the ends of the confidence intervals for some scores and is just outside when the APACHE score equals 18.

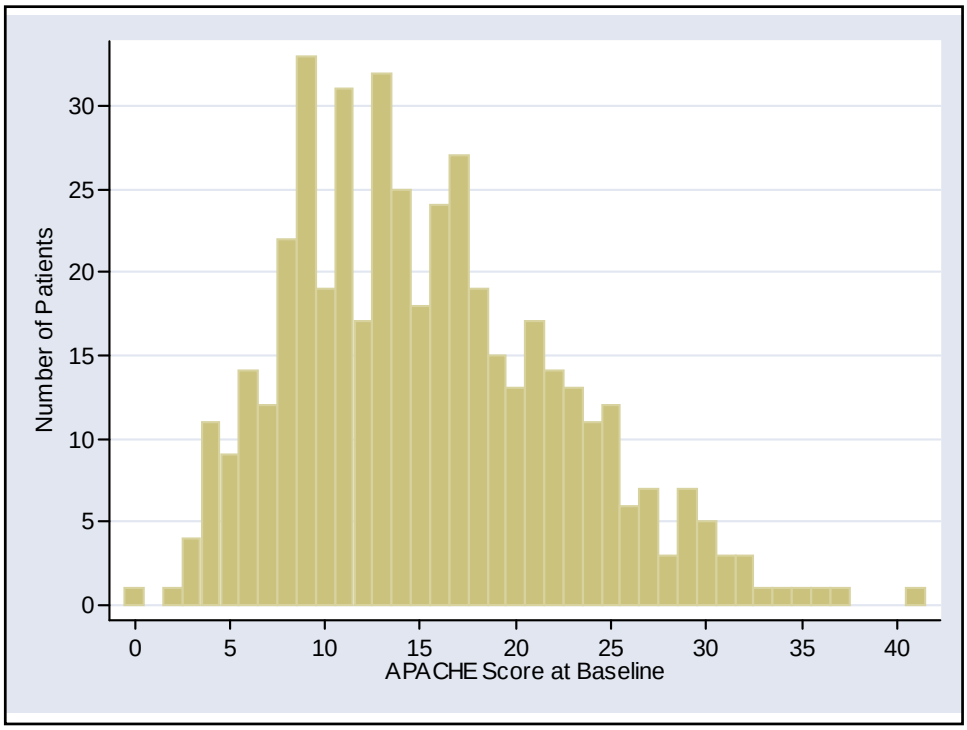
```
*
*  Graph number of patients with different APACHE II scores
*
. * Graphics > Histogram
. histogram apache [freq=patients], discrete frequency xlabel(0(5)40) /// {4}
>     ylabel(0(5)30, angle(0)) ytitle(Number of Patients)
(start=0, width=1)
```

{4} This command produces a histogram of the patient APACHE II scores.

discrete specifies that the data have a discrete number of values. It produces one bar per value unless **width** is also specified.

frequency specifies that the *y*-axis is to be **number of patients** rather than proportion of patients.


histogram - Histograms for continuous and categori...
Main
if/in
Weights
Density plots
Add plots
Y axis
X axis
Titles
Legend
Overall
By
Data
Variable:
apache
Data are continuous
Data are discrete
Bins
Number of bins
Width of bins
Theoretical minimum value
Y axis
Density
Fraction
Frequency
Percent
Bar properties
Add height labels to bars
Bar label properties
histogram - Histograms for continuous and categori...
Weight type:
None
Frequency weights
Help weights
Frequency weight:
patients
OK
Cancel
Submit


Number of Patients
0
5
10
15
20
25
30
APACHE Score at Baseline
0
5
10
15
20
25
30
35
40

12. Simple 2x2 Case-Control Studies

a) Example: Esophageal Cancer and Alcohol

Breslow & Day, Vol. I (1980) give the following results from the Ille-et-Vilaine case-control study of esophageal cancer and alcohol (Tuyns et al. 1977) .

Cases were 200 men diagnosed with esophageal cancer in regional hospitals between 1/1/1972 and 4/30/1974.

Controls were 775 men drawn from electoral lists in each commune.

Esophageal	Daily Alcohol Consumption		
Cancer	≥ 80g	< 80g	Total
Yes (Cases)	96	104	200
No (Controls)	109	666	775
Total	205	770	975

b) Review of Classical Case-Control Theory

Let m_i = number of cases ($i = 1$) or controls ($i = 0$)

d_i = number of cases ($i = 1$) or controls ($i = 0$) who are heavy drinkers.

Then the observed prevalence of heavy drinkers is

d_0/m_0 = 109/775 for controls and

d_1/m_1 = 96/200 for cases.

The observed prevalence of moderate or non-drinkers is

$(m_0 - d_0)/m_0$ = 666/775 for controls and

$(m_1 - d_1)/m_1$ = 104/200 for cases.

The observed odds that a case or control will be a heavy drinker is

$$(d_i / m_i) / [(m_i - d_i) / m_i] = d_i / (m_i - d_i)$$

= 109/666 and 96/104 for controls and cases, respectively.

The observed odds ratio for heavy drinking in cases relative to controls is

$$\hat{\psi} = \frac{d_1 / (m_1 - d_1)}{d_0 / (m_0 - d_0)} = \frac{96/104}{109/666} = 5.64$$

If the cases and controls are a representative sample from their respective underlying populations then

1. $\hat{\psi}$ is an unbiased estimate of the true odds ratio for heavy drinking in cases relative to controls in the underlying population.
2. This true odds ratio also equals the true odds ratio for esophageal cancer in heavy drinkers relative to moderate drinkers.

   Case-control studies would be pointless if this were not true.

Since esophageal cancer is rare $\hat{\psi}$ also estimates the **relative risk** of esophageal cancer in heavy drinkers relative to moderate drinkers.

Woolf's estimate of the standard error of the log odds ratio is

$$\text{se}_{\log(\hat{\psi})} = \sqrt{\frac{1}{d_0} + \frac{1}{m_0 - d_0} + \frac{1}{d_1} + \frac{1}{m_1 - d_1}}$$

and the distribution of $\log(\hat{\psi})$ is approximately normal.

Hence, if we let

$$\hat{\psi}_L = \hat{\psi} \exp\left[-1.96\,\text{se}_{\log(\hat{\psi})}\right]$$

and

$$\hat{\psi}_U = \hat{\psi} \exp\left[1.96\,\text{se}_{\log(\hat{\psi})}\right]$$

then $(\hat{\psi}_L, \hat{\psi}_U)$ is a 95% confidence interval for ψ .

13. Logistic Regression Models for 2x2 Contingency Tables

Consider the logistic regression model

$$\text{logit}(E(d_i / m_i)) = \alpha + \beta x_i \qquad \{3.9\}$$

where $E(d_i / m_i) = \pi_i =$ Probability of being a heavy drinker for cases ($i = 1$) and controls ($i = 0$).

and $x_i = \begin{cases} 1 = \text{ cases} \\ 0 = \text{ for controls} \end{cases}$

Then {3.9} can be rewritten

$$\text{logit}(\pi_i) = \log(\pi_i / (1 - \pi_i)) = \alpha + \beta x_i$$

Hence

$$\log(\pi_1 / (1 - \pi_1)) = \alpha + \beta x_1 = \alpha + \beta$$
$$\log(\pi_0 / (1 - \pi_0)) = \alpha + \beta x_0 = \alpha$$

since $x_1 = 1$ and $x_0 = 0$.

Subtracting these two equations gives

$$\log(\pi_1 / (1 - \pi_1)) - \log(\pi_0 / (1 - \pi_0)) = \beta$$

$$\log\left[\frac{\pi_1 / (1 - \pi_1)}{\pi_0 / (1 - \pi_0)}\right] = \log(\psi) = \beta$$ and hence the true odds ratio $\psi = e^{\beta}$

a) Estimating relative risks from the model coefficients

Our primary interest is in β. Given an estimate $\hat{\beta}$ of β then $\hat{\psi} = e^{\hat{\beta}}$

b) Nuisance parameters

α is called a **nuisance parameter**. This is one that is required by the model but is not used to calculate interesting statistics.

14. Analyzing Case-Control Data with Stata

The Ille-et-Vilaine data may be analyzed as follows:

```
. *  esoph_ca_cc1.log
. *
. *  Logistic regression analysis of Illes-et-Vilaine
. *  2x2 case-control data.
. *
. *  Enter 2x2 table by hand with editor
. *
```

```
. edit                                                          {1}

. list

     +-----------------------------+
     | cancer   alcohol  patients |
     |-----------------------------|
  1. |      0         0       666 |
  2. |      1         0       104 |
  3. |      0         1       109 |
  4. |      1         1        96 |
     +-----------------------------+

. label define yesno 0 "No" 1 "Yes"                             {2}

. label values cancer yesno                                     {3}

. label define dose 0 "< 80g" 1 ">= 80g"

. label values alcohol dose

. list                                                          {4}

     +-----------------------------+
     | cancer   alcohol  patients |
     |-----------------------------|
  1. |     No     < 80g       666 |
  2. |    Yes     < 80g       104 |
  3. |     No    >= 80g       109 |
  4. |    Yes    >= 80g        96 |
     +-----------------------------+
```

{1} Press the **Editor** button to access Stata's spreadsheet-like editor. **Enter** three variables named ***cancer, alcohol*** and ***patients*** as shown in the following *list* command.

{2} The *cancer* variable takes the value **0** for **controls** and **1** for **cases**. To define these values we first define a value **label** called *yesno* that links **0** with "**No**" and **1** with "**Yes**".

{3} We then use the ***label values*** command to link the variable ***cancer*** with the values label ***yesno***. Multiple variables can be assigned to the same values label.

{4} The ***list*** command now gives the value labels of the *cancer* and *alcohol* variables instead of their numeric values. The **numeric values** are still **available** for use in estimation commands.

```
. *
. *  Calculate the odds ratio for esophageal cancer
. *  associated with heavy drinking.
. *
. * Statistics > Epidemiology... > Tables... > Case-control odds ratio
. cc cancer alcohol [freq=patients], woolf
                                                                        {5}

                 | alcohol                 |             Proportion
                 |   Exposed   Unexposed   |     Total     Exposed
-----------------+-------------------------+----------------------
           Cases |        96         104   |       200      0.4800
        Controls |       109         666   |       775      0.1406
-----------------+-------------------------+----------------------
           Total |       205         770   |       975      0.2103
                 |                         |
                 |      Point estimate     |  [95% Conf. Interval]
                 |-------------------------+----------------------
      Odds ratio |          5.640085       |  4.000589    7.951467  (Woolf)    {6}
 Attr. frac. ex. |           .8226977      |  .7500368    .8742371  (Woolf)
 Attr. frac. pop |           .3948949      |
                 +-------------------------------------------------
                             chi2(1) =   110.26  Pr>chi2 = 0.0000
```

{5} Perform a **classical** case-control **analysis** of the data in the 2x2 table defined by *cancer* and *alcohol*. [***freq=patients***] gives the number of patients who have the specified values of *cancer* and *alcohol*. The **woolf** option specifies that the 95% confidence interval for the odds ratio is to be calculated using Woolf's method.

We could have entered one record per patient giving

666 records with cancer = 0 and alcohol = 0,
104 records with cancer = 1 and alcohol = 0,
109 records with cancer = 0 and alcohol = 1, and
96 records with cancer = 1 and alcohol = 1.

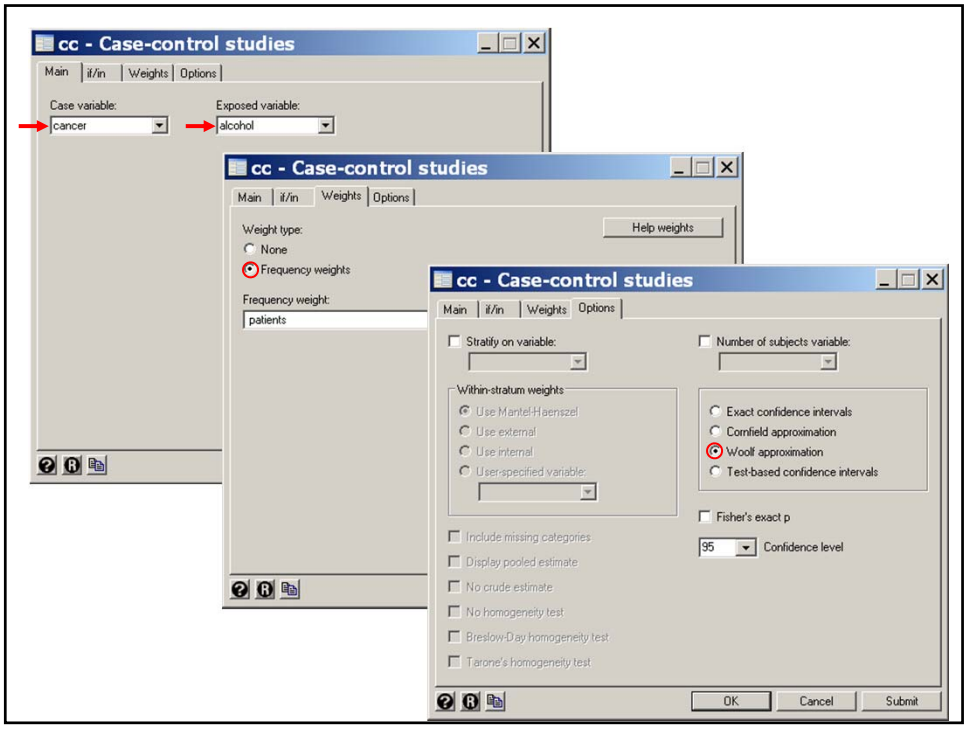
Then the command

```
cc cancer alcohol, woolf
```

would have given exactly the same results as those shown in this example.

N.B. We need to use the **[freq=patients]** command modifier whenever each record represents multiple patients. This will also be true in logistic regression and other commands.

{6} The estimated **odds ratio** is $\frac{96/104}{109/666} = 5.64$




```
. *
. * Now calculate the same odds ratio using logistic regression
. *
. * Statistics > Binary outcomes > Logistic regression
. logit alcohol cancer [freq=patients]                          {7}

Logistic regression                               Number of obs   =        975
                                                  LR chi2(1)      =      96.43
                                                  Prob > chi2     =     0.0000
Log likelihood =  -453.2224                       Pseudo R2       =     0.0962

------------------------------------------------------------------------------
     alcohol |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      cancer |   1.729899   .1752366     9.87   0.000     1.386442    2.073356
       _cons |  -1.809942   .1033238   -17.52   0.000    -2.012453   -1.607431
------------------------------------------------------------------------------
```

{7} This is the analogous ***logit*** command for simple logistic regression. If we had entered the data as

cancer	heavy	patients
0	109	775
1	96	200

Then we would have achieved the same analysis with the command

```
glm heavy cancer, family(binomial patients) link(logit)
```

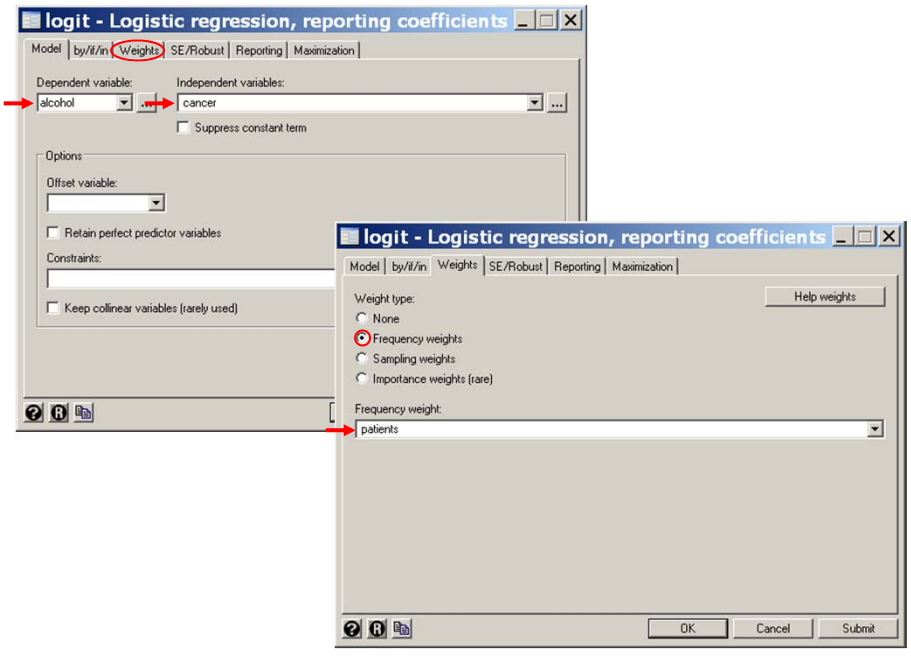
Both of these commands fit the model

$$\text{logit}(E(\text{alcohol})) = \alpha + \text{cancer}*\beta$$

giving **β = 1.73** = the **log odds ratio** of being a heavy drinker in cancer patients relative to controls. The **standard error** of **β** is **0.1752**

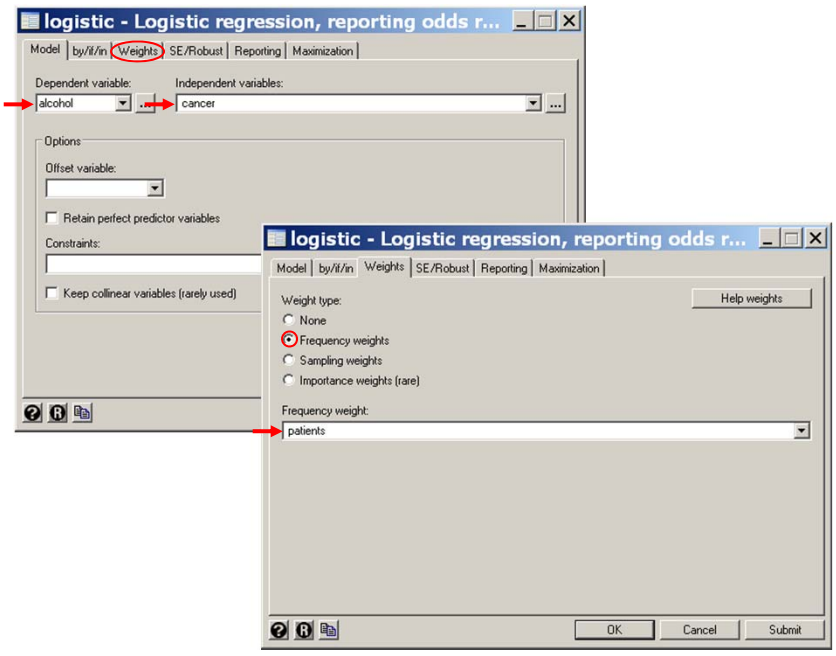
The **odds ratio** is exp(1.73) = 5.64.

The **95% confidence interval** for the odds ratio is exp(1.73 ±1.96*0.1752) = (4.00, 7.95)



```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
.  logistic alcohol cancer [freq=patients]                {8}

Logistic regression                                     No. of obs =
975
                  LR chi2(1) =     96.43
                  Prob > chi2=    0.0000
Log likelihood =   -453.2224      Pseudo R2  =    0.0962

------------------------------------------------------------------------------
 alcohol | Odds Ratio   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
  cancer |   5.640085   .9883491      9.87    0.000       4.000589    7.951467
------------------------------------------------------------------------------
```

{8} The ***logistic*** command calculates the odds ratio and its confidence interval directly.



a) Logistic and classical estimates of the 95% CI of the OR

The 95% confidence interval is

$$(5.64\exp(-1.96\times 0.1752),\ 5.64\exp(1.96\times 0.1752)) = (4.00,\ 7.95).$$

The classical limits using Woolf's method is

$$(5.64\exp(-1.96\times s),\ 5.64\exp(1.96\times s)) = (4.00,\ 7.95),$$

where $s^2 = 1/96 + 1/109 + 1/104 + 1/666 = 0.0307 = (0.1752)^2$.

Hence Logistic regression is in exact agreement with classical methods in this simple case.

In Stata the command

```
cc cancer alcohol [freq=patients], woolf
```

gives us Woolf's 95% confidence interval for the odds ratio. We will cover how to calculate confidence intervals using *glm* in the next chapter.

15. Regressing Disease Against Exposure

The simplest explanation of simple logistic regression is the one given above. Unfortunately, it does not generalize to multiple logistic regression where we are considering several risk factors at once. In order to make the next chapter easier to understand, let us return to simple logistic regression one more time.

Suppose we have a population who either are or are not exposed to some risk factor.

Let π'_j denote the true probability of disease in exposed (j = 1) and unexposed (j = 0) people.

We conduct a case-control study in which we select a representative sample of diseased (case) and healthy (control) subjects from the underlying population. That is, the selection is done in such a way that the probability that an individual is selected is unaffected by her exposure status.

Let m_j be the number of study subject who are (j = 1) or are not (j = 0) exposed,

d_j be the number of cases who are (j = 1) or are not (j = 0) exposed,

$x_j =$ j denote exposure status, and

π_j be the probability that a study subject is a case given that she is (j=1) or is not (j=0) exposed.

Consider the model

$$\text{logit}\,(E(d_j / m_j)) = \alpha + \beta x_j$$

This is a legitimate logistic regression model with $E(d_j / m_j) = \pi_j$. It can be shown, however, that this model can be rewritten as

$$\text{logit}(\pi'_j) = \alpha' + \beta x_j$$

where α' is a different constant. However, since α' cancels out in the odds ratio calculation, β estimates the log odds ratio for disease in exposed vs. unexposed members of the population as well as in our case-control sample.

Thus in building logistic regression models it makes sense to regress disease against exposure even though we have no estimate of the probability of disease in the underlying population.

16. What we have covered

- Simple logistic regression: Assessing the effect of a continuous variable on a dichotomous outcome
- How logistic regression parameters affect the probability of an event
  - $\pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$
  - exp(β) is the odds ratio for death associated with a unit increase in x.
- Probability, odds and odds ratios
- Generalized linear models: The relationship between linear and logistic regression $\text{logit}(E(d_i)) = \alpha + x_i \beta$
- Wald and Wilson confidence intervals for proportions
- Plotting probability of death with 95% confidence bands as a function of a continuous risk factor
- Review of classic 2x2 case-control studies
- Analyzing case-control studies with logistic regression

Cited References

Bernard, G. R., et al. (1997). The effects of ibuprofen on the physiology and survival of patients with sepsis. The Ibuprofen in Sepsis Study Group. *N Engl J Med* 336: 912-8.

Breslow, N. E. and N. E. Day (1980). Statistical Methods in Cancer Research: Vol. 1 - The Analysis of Case-Control Studies. Lyon, France, IARC Scientific Publications.

Tuyns, A. J., G. Pequignot, et al. (1977). Le cancer de L'oesophage en Ille-et-Vilaine en fonction des niveau de consommation d'alcool et de tabac. Des risques qui se multiplient. *Bull Cancer* 64: 45-60.

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.

IV. MULTIPLE LOGISTIC REGRESSION

- Extend simple logistic regression to models with multiple covariates
- Similarity between multiple linear and multiple logistic regression
- Multiple 2x2 tables and the Mantel-Haenszel test
  - Estimating an odds ratio that is adjusted for a confounding variable
- Using logistic regression as an alternative to the Mantel-Haenszel test
- Using indicator covariates to model categorical variables
- Making inferences about odds ratios derived from multiple parameters
- Analyzing complex data with logistic regression
  - Multiplicative models
  - Models with interaction
- Assessing model fit
  - Testing the change in model deviance in nested models
  - Evaluating residuals and influence
- Using restricted cubic splines in logistic regression models
  - Plotting the probability of an outcome with confidence bands
  - Plotting odds ratios and confidence bands

1. The Model

If the data is organized as one record per patient then the model is

$$\text{logit}\left(E(d_i)\right) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \qquad \{4.1\}$$

where

$x_{i1}, x_{I2}, \ldots, x_{ik}$ are covariates from the i^{th} patient

$\alpha, \beta_1, \ldots \beta_k$, are unknown parameters

$$d_i = \begin{cases} 1: & i^{\text{th}} \text{ patient suffers event of interest} \\ 0: & \text{otherwise} \end{cases}$$

If the data is organized as one record per unique combination of covariate values then the model is

$$\text{logit}\,(E(d_i / m_i)) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \qquad \{4.2\}$$

where m_i is the number of patients with covariate values x_{i1}, x_{i2}, …, x_{ik} and d_i is the number of events among these m_i subjects.

d_i is assumed to have a binomial distribution obtained from m_i dichotomous trials with probability of success $\pi(x_{i1}, x_{i2}, \ldots, x_{ik})$ on each trial.

Thus, the only difference between simple and multiple logistic regression is that the linear predictor is now $\alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$. As in simple logistic regression, the model has a logit link function; the random component, d_i / m_i has a binomial distribution.

2. Mantel-Haenszel Test of a Common Odds Ratio

The following data is from the Ille-et-Vilaine study of esophageal cancer and alcohol by Tuyns et al. (1977). This data is published in Appendix I of Breslow and Day Vol. I, who also provide an excellent and extensive discussion of this data set.

Age	Cancer	Daily Alcohol Consumption ≥ 80g	Daily Alcohol Consumption <80g		% ≥ 80g
25-34	Yes	1	0	1	100.00%
	No	9	106	115	7.83%
		10	106	116	8.62%
35-44	Yes	4	5	9	44.44%
	No	26	164	190	13.68%
		30	169	199	15.08%
45-54	Yes	25	21	46	54.35%
	No	29	138	167	17.37%
		54	159	213	25.35%
55-64	Yes	42	34	76	55.26%
	No	27	139	166	16.27%
		69	173	242	28.51%
65-74	Yes	19	36	55	34.55%
	No	18	88	106	16.98%
		37	124	161	22.98%

a) Confounding Variables

A **confounding variable** is one that is associated with both the disease and exposure of interest but which is not, in itself, a focus of our investigation.

Note mild evidence that age confounds the effect of alcohol on cancer risk.

b) Age-adjusted odds ratios

The following log file show how to calculate the common odds ratio for esophageal cancer associated with heavy alcohol use in five age strata. It thus calculates an **age-adjusted** odds ratio for esophageal cancer among heavy and light drinkers of similar age.

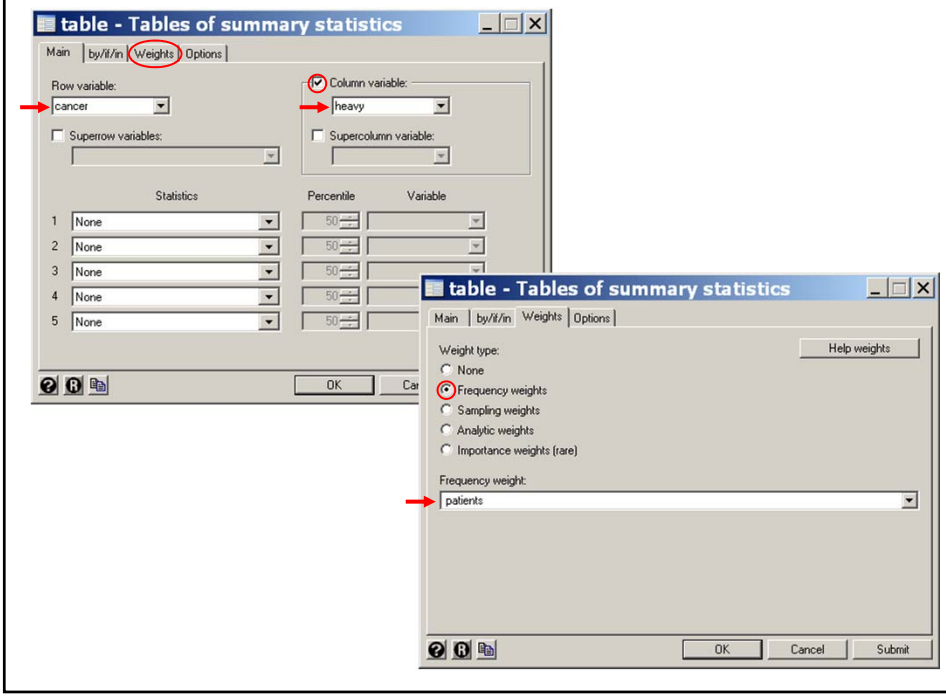
3. Deriving the Mantel-Haenszel test with Stata

```
  *  5.5.EsophagealCa.log
. *
. *  Calculate the Mantel-Haenszel age-adjusted odds ratio from
. *  the Ille-et-Vilaine study of esophageal cancer and alcohol
. *  (Breslow & Day 1980, Tuyns 1977).
. *
. use C:\WDDtext\5.5.EsophagealCa.dta, clear

. codebook age cancer heavy


age ---------------------------------------- Age (years)
                  type:  numeric (float)
                 label:  age

                 range:  [1,6]
units:  1
         unique values:  6        coded missing:  0 / 192

            tabulation:  Freq.   Numeric  Label
                            32         1  25-34
                            32         2  35-44
                            32         3  45-54
                            32         4  55-64
                            32         5  65-74
                            32         6  >= 75
```

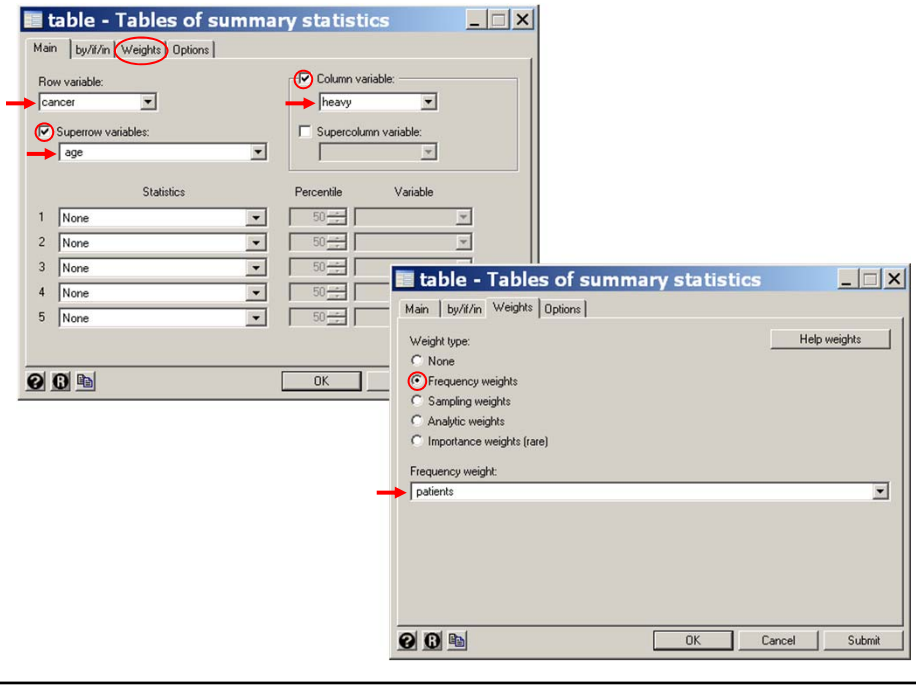
```
cancer ------------------------------- Esophageal Cancer
                  type:  numeric (float)
                 label:  yesno

                 range:  [0,1]
units:  1
         unique values:  2        coded missing:  0 / 192

            tabulation:  Freq.   Numeric  Label
                            96         0  No
                            96         1  Yes



heavy ------------------------- Heavy Alcohol Consumption
                  type:  numeric (float)
                 label:  heavy

                 range:  [0,1]
units:  1
         unique values:           coded missing:  0 / 192

            tabulation:  Freq.   Numeric  Label
                            96         0  < 80 gm
                            96         1  >= 80 gm
```

```
. * Statistics > Summaries... > Tables > Table of summary statistics (table).
 table heavy cancer [freq=patients]                                    {1}

----------+------------------
          |  Heavy Alcohol
Esophagea |   Consumption
l Cancer  | < 80 gm  >= 80 gm
----------+------------------
       No |     666       104
      Yes |     109        96
----------+------------------
```

{1} This **table** command gives 2×2 cross-tables of ***heavy*** by ***cancer***, and confirms *that EsophagealCancer.dta* is the correct data set.



```
. table  cancer heavy [freq=patients], by(age)

----------+-------------------
Age       |
(years)   |
and       |   Heavy Alcohol
Esophagea |    Consumption
l Cancer  |  < 80 gm  >= 80 gm
----------+-------------------
25-34     |
       No |      106         9
      Yes |                  1
----------+-------------------
35-44     |
       No |      164        26
      Yes |        5         4
----------+-------------------
45-54     |
       No |      138        29
      Yes |       21        25
----------+-------------------
55-64     |
       No |      139        27
      Yes |       34        42
----------+-------------------
65-74     |
       No |       88        18
      Yes |       36        19
----------+-------------------
>= 75     |
       No |       31
      Yes |        8         5
----------+-------------------
```



```
. * Statistics > Epidemiology... > Tables... > Case-control odds ratio
. cc heavy cancer [freq=patients], by(age)                                  {2}

 Age (years) |       OR      [95% Conf. Interval]    M-H Weight
-------------+-----------------------------------------------
       25-34 |          .            0          .             0 (exact)
       35-44 |   5.046154     .9268664   24.86538      .6532663 (exact)
       45-54 |   5.665025     2.632894   12.16536      2.859155 (exact)
       55-64 |   6.359477     3.299319   12.28473      3.793388 (exact)
       65-74 |   2.580247     1.131489   5.857261      4.024845 (exact)
       >= 75 |          .     4.388738          .             0 (exact)
-------------+-----------------------------------------------
       Crude |   5.640085     3.937435   8.061794                (exact)      {3}
M-H combined |   5.157623     3.562131   7.467743                             {4}
-------------+-----------------------------------------------
Test of homogeneity (Tarone)   chi2(5) =     9.30  Pr>chi2 = 0.0977          {5}

             Test that combined OR = 1:                                     {6}
                         Mantel-Haenszel chi2(1) =     85.01
                         Pr>chi2 =    0.0000
```

{2} The ***by(age)*** option causes **odds ratios** to be calculated for each **age strata**. No estimate is given for the youngest strata because there were no moderate drinking cases. No estimate is given for the oldest strata because there were no heavy drinking controls.

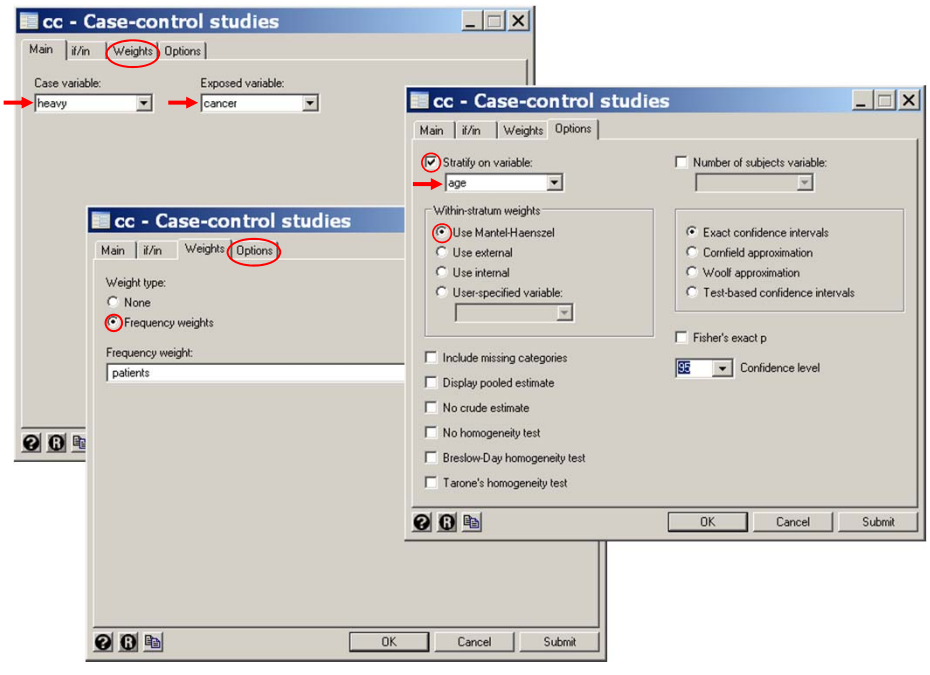
{3} The **crude** odds ratio is **5.64** which we derived in the last chapter. This odds ratio is obtained by ignoring the age strata.

The **exact 95% confidence interval** consists of all values of the odds ratio that **cannot be rejected** at the $P = 0.05$ level of statistical significance (see text, Section 1.4.7). The derivation of this interval uses a rather complex iterative formula (Dupont and Plummer 1999).

{4} The **Mantel-Haenszel** (M-H) estimate of the **common odds ratio** within all age strata is **5.16**. This is an **age-adjusted** estimate. It is slightly lower than the crude estimate, and is consistent with a mild confounding of age and drinking habits on the risk of esophageal cancer.

{5} The M-H estimate is only reasonable if the data is consistent with the hypothesis that the alcohol-cancer odds ratio does not vary with age. The **test for homogeneity** tests the null hypothesis that all age strata share a common odds ratio. This test is not significant, which suggests that the M-H estimate may be reasonable.

{6} The test of the **null hypotheses** that the odds ratio equals 1 is highly **significant**. Hence the association between heavy alcohol consumption and esophageal cancer can not be explained by chance. The argument for a **causal** relationship is strengthened by the **magnitude** of the **odds ratio**.



4. Effect Modifiers and Confounding Variables

a) Test of homogeneity of odds ratios

In the previous example the test for homogeneity of the odds ratio was not significant (see comment 5). Of course, lack of significance does not prove the null hypotheses, and it is prudent to look at the odds ratios from the individual age strata. In the preceding Stata output these values are fairly similar for all strata except ages 65-74, where the odds ratio drops to 2.6. This may be due to chance, or perhaps, to a hardy survivor effect. You must use your clinical judgment in deciding what to report.

Effect Modifier: A variable that influences the effect of a risk factor on the outcome variable.

The key differences between confounding variables and effect modifiers are:

i) Confounding variables are not of primary interest in our study while effect modifiers are.

ii) A variable is an important effect modifier if there is a meaningful interaction between it and the exposure of interest on the risk of the event under study.

5. Logistic Regression For Multiple 2×2 Contingency Tables

a) Estimating the common relative risk from the parameter estimates

Let

m_{jk} be the number of subjects in the j^{th} age strata who are (k = 1) or are not (k = 0) heavy drinkers.

d_{jk} be the number of cancers among these m_{jk} subjects.

x_k = k = 1 or 0 depending on their drinking status.

π_{jk} be the probability that someone in the j^{th} age strata who does (k = 1) or doesn't (k = 0) drink heavily develops cancer.

Consider the logistic regression model

$$\text{logit}\left(E\left(d_{jk}/m_{jk}\right)\right)=\alpha_j+x_k\beta \qquad \textbf{\{4.3\}}$$

where d_{jk} has a binomial distribution obtained from m_{jk} independent trials with probability of success with π_{jk} on each trial.

Then for any age strata j, $E\left(d_{jk}/m_{jk}\right)=\pi_{jk}$ and

$$\text{logit}\left(E\left(d_{j0}/m_{j0}\right)\right)=\text{logit}(\pi_{j0})=\log\left(\pi_{j0}/(1-\pi_{j0})\right)=\alpha_j \qquad \textbf{\{4.4\}}$$

Similarly

$$\text{logit}\left(E\left(d_{j1}/m_{j1}\right)\right)=\log\left(\pi_{j1}/(1-\pi_{j1})\right)=\alpha_j+\beta \qquad \textbf{\{4.5\}}$$

Subtracting equation {4.4} from equation {4.5} gives that

$$\log\left(\pi_{j1}/(1-\pi_{j1})\right)-\log\left(\pi_{j0}/(1-\pi_{j0})\right)=\beta \quad \text{or}$$

$$\log\left(\frac{\pi_{j1}/(1-\pi_{j1})}{\pi_{j0}/(1-\pi_{j0})}\right)=\log\psi=\beta$$

Hence, this model implies that the odds ratio for cancer is the same in all strata and equals exp(β).

This is an age-adjusted estimate of the cancer odds ratio

In practice we fit model {4.1} by defining indicator covariates

$$z_j=\begin{cases}1: & \text{if subjects are from the } j^{\text{th}} \text{ age strata}\\ 0: & \text{otherwise}\end{cases}$$

Then {4.3} becomes

$$\text{logit}\left(E\left(d_{jk}/m_{jk}\right)\right)=z_1\alpha_1+z_2\alpha_2+z_3\alpha_3+z_4\alpha_4+z_5\alpha_5+z_6\alpha_6+x_k\beta$$

Note that this model places no restraints of the effect of age on the odds of cancer and only requires that the within strata odds ratio be constant.

For example, a moderate drinker from the 3rd age stratum has log odds

$$\text{logit}\left(E\left(d_{3,0} / m_{3,0}\right)\right) = \alpha_3$$

While a moderate drinker from the first age stratum has

$$\text{logit}\left(E\left(d_{1,0} / m_{1,0}\right)\right) = \alpha_1$$

Hence the log odds ratio for stratum 3 versus stratum 1 is α_3 - α_1, which can be estimated independently of the cancer risk associated with age strata 2, 4, 5 and 6.

An equivalent model is

$$\text{logit}\left(E\left(d_{jk} / m_{jk}\right)\right) = \alpha + z_2\alpha_2 + z_3\alpha_3 + z_4\alpha_4 + z_5\alpha_5 + z_6\alpha_6 + x_k\beta \qquad \textbf{\{4.6\}}$$

For this model, a moderate drinker from the 3rd age stratum has log odds

$$\text{logit}\left(E\left(d_{3,0} / m_{3,0}\right)\right) = \alpha + \alpha_3$$

While a moderate drinker from the first age stratum has

$$\text{logit}\left(E\left(d_{1,0} / m_{1,0}\right)\right) = \alpha$$

Hence the log odds ratio for stratum 3 versus stratum 1 is

$$\left(\alpha + \alpha_3\right) - \alpha = \alpha_3$$

This is slightly preferable to our previous formulation in that it involves one parameter rather than 2.

An alternative model that we could have used is

$$\text{logit}\left(E\left(d_{jk} / m_{jk}\right)\right) = \text{age} \times \alpha + x_k \beta$$

However, this model imposes a linear relationship between age and the log odds for cancer. That is, the log odds ratio

for age stratum 2 vs stratum 1 is $2\alpha - \alpha = \alpha$

for age stratum 3 vs stratum 1 is $3\alpha - \alpha = 2\alpha$

⋮

for age stratum 6 vs stratum 1 is $6\alpha - \alpha = 5\alpha$

6. Analyzing Multiple 2×2 Contingency Tables

```
. *   5.9.EsophagealCa.ClassVersion.log
. *
. *  Calculate age-adjusted odds ratio from the Ille-et-Vilaine study
. *  of esophageal cancer and alcohol using logistic regression.
. *
. use C:\WDDtext\5.5.EsophagealCa.dta, clear
. *
. *  First, define indicator variables for the age strata 2 through 6
. *
```

```
. generate age2 = 0

. replace age2 = 1 if age == 2
(32 real changes made)
. generate age3 = 0

. replace age3 = 1 if age == 3
(32 real changes made)
. generate age4 = 0

. replace age4 = 1 if age == 4
(32 real changes made)
. generate age5 = 0

. replace age5 = 1 if age == 5
(32 real changes made)
. generate age6 = 0

. replace age6 = 1 if age == 6
(32 real changes made)
```

```
. * Statistics > Binary outcomes > Logistic regression
. logit cancer age2 age3 age4 age5 age6 heavy   [freq=patients]    {1}

Logistic regression                             No. of obs      =        975
                                                LR chi2(6)      =     200.57
                                                Prob > chi2     =     0.0000
Log likelihood   = -394.46094                   Pseudo R2       =     0.2027

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age2 |    1.542294   1.065895     1.45   0.148     -.546822     3.63141
        age3 |    3.198762    1.02314     3.13   0.002     1.193445     5.204079
        age4 |     3.71349   1.018531     3.65   0.000     1.717207     5.709774
        age5 |    3.966882   1.023072     3.88   0.000     1.961698     5.972066
        age6 |     3.96219   1.065024     3.72   0.000      1.87478     6.049599
       heavy |     1.66989   .1896018     8.81   0.000     1.298277     2.041503  {2}
       _cons |   -5.054348   1.009422    -5.01   0.000    -7.032778    -3.075917
------------------------------------------------------------------------------
```

The results of this logistic regression are similar to those obtained form the Mantel-Haenszel test. The age-adjusted odds ratio from this latter test was **5.16** as compared to **5.31** from logistic regression.

{1} By default, Stata adds a constant term to the model. Hence, this command uses model {4.6}.

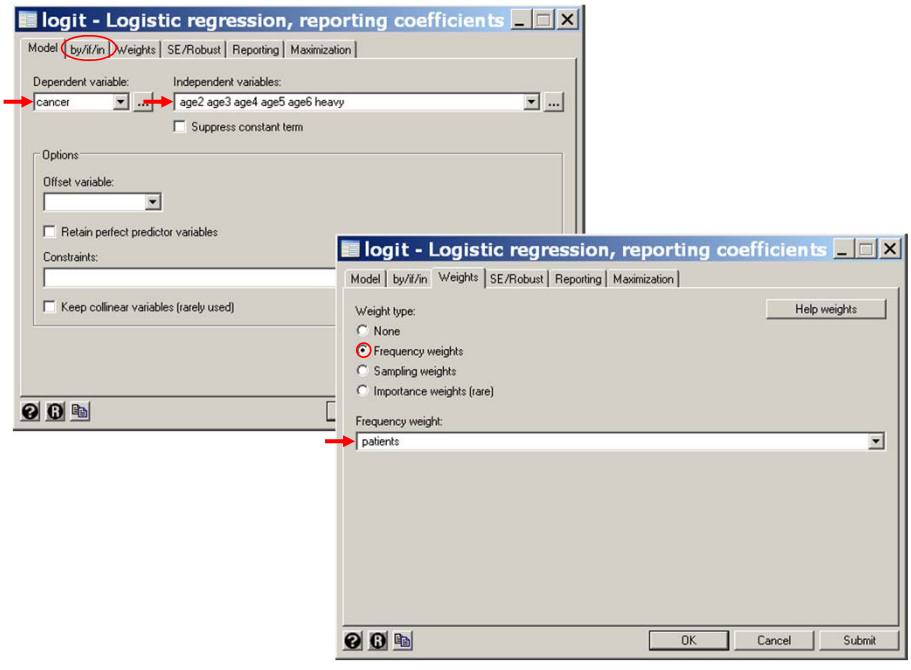
The *coef* option specifies that the model parameter estimates are to be listed as follows.

{2} The parameter estimate associated with *heavy* is 1.67 with a standard error of 0.1896. A 95% confidence interval for this interval is $1.67 \pm 1.96 \times 0.1896 = [1.30, 2.04]$.

The age-adjusted estimated odds ratio for cancer in heavy drinkers relative to moderate drinkers is

$\psi = \exp(1.67) = 5.31$ with a 95% confidence interval

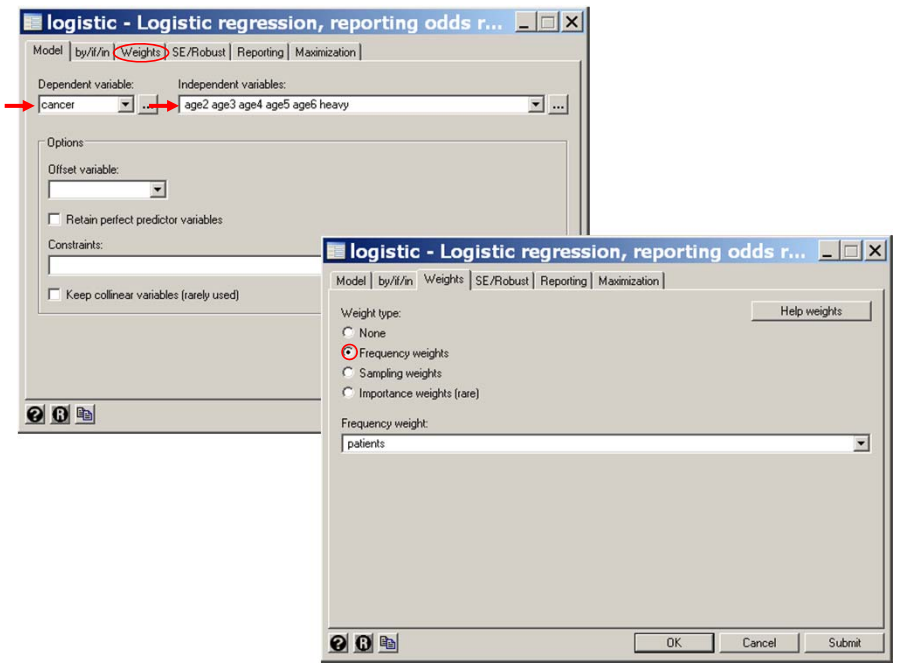
$[\exp(1.30), \exp(2.04)] = [3.66, 7.70]$.



```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic cancer age2 age3 age4 age5 age6 heavy
> [freq=patients]                                                    {3}

Logistic regression                               No. of obs      =        975
                                                  LR chi2(6)      =     200.57
                                                  Prob > chi2     =     0.0000
Log likelihood   = -394.46094                     Pseudo R2       =     0.2027
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        age2 |   4.675303   4.983382     1.45   0.148      .5787862    37.76602
        age3 |   24.50217   25.06914     3.13   0.002      3.298423    182.0131
        age4 |   40.99664   41.75634     3.65   0.000       5.56895    301.8028
        age5 |   52.81958   54.03823     3.88   0.000      7.777389    392.3155
        age6 |   52.57232   55.99081     3.72   0.000      6.519386    423.9432
       heavy |   5.311584   1.007086     8.81   0.000      3.662981    7.702174
------------------------------------------------------------------------------
```

{3} Without the *coef* option logistic does not output the constant parameter and exponentiates the other coefficients. This usually saves hand computation.

Note that the age adjusted odds ratio for heavy drinking is 5.31 with a 95% confidence interval of [3.7 – 7.7].



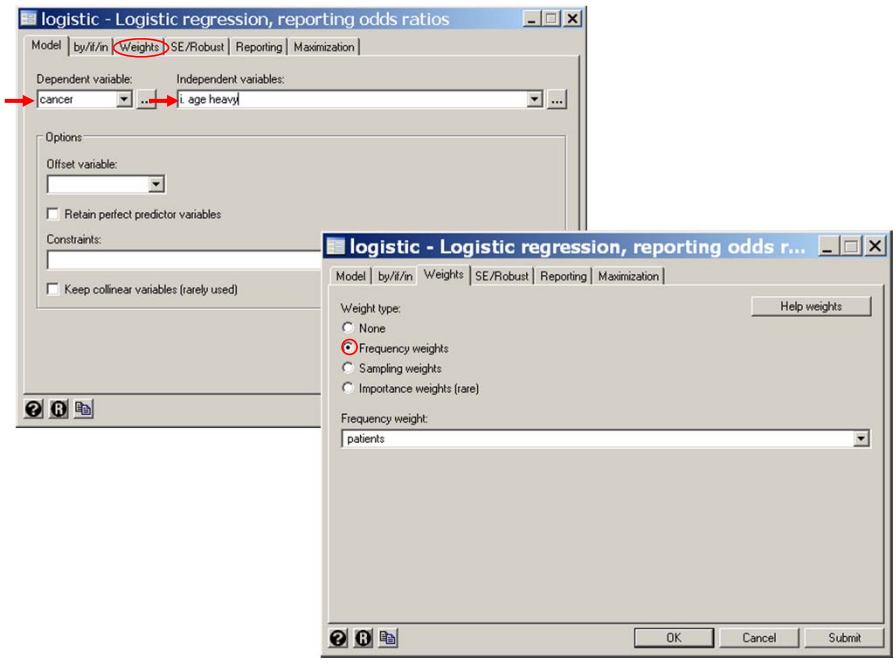
7. Handling Categorical Variables in Stata

In the preceding example, age is a categorical variable taking 6 values that is recorded as 5 separate indicator variables. It is very common to recode categorical variables in this way to avoid forcing a linear relationship on the effect of a variable on the response outcome. In the preceding example we did the recording by hand. It can also be done much faster using the *i.varname* syntax. We illustrate this by repeating the preceding analysis of model {4.3}.

```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic cancer i.age heavy [freq=patients]                              {1}

Logistic regression                               No. of obs      =        975
                                                  LR chi2(6)      =     200.57
                                                  Prob > chi2     =     0.0000
Log likelihood = -394.46094                       Pseudo R2       =     0.2027
```

{1} ***i.age*** indicates that *age* is to be recoded as five indicator variables (one for each value of *age*). These variables are named *2.age*, *3.age*, *4.age*, *5.age*, and *6.age*. By default the smallest value of *age* is not assigned a separate indicator variable and a constant term is included in the model giving

$$\text{logit}\left(E\left(d_{jk} / m_{jk}\right)\right) = \alpha + \alpha_j + x_k\beta : j = 2,\ldots,6;\ k = 0,1$$



cancer	Odds Ratio.	Std. Err.	z	P>|z|	[95% Conf.	Interval]	
age							
2	4.675303	4.983382	1.45	0.148	.5787862	37.76602	
3	24.50217	25.06914	3.13	0.002	3.298423	182.0131	
4	40.99664	41.75634	3.65	0.000	5.56895	301.8028	
5	52.81958	54.03823	3.88	0.000	7.111389	392.3155	
6	52.57232	55.99081	3.72	0.000	6.519386	423.9432	
heavy	5.311584	1.007086	8.81	0.000	3.662981	7.702174	{2}

{2} Note that the odds ratio estimate for *heavy* = **5.31** is the same as in the earlier analysis where the indicator variables were explicitly defined.

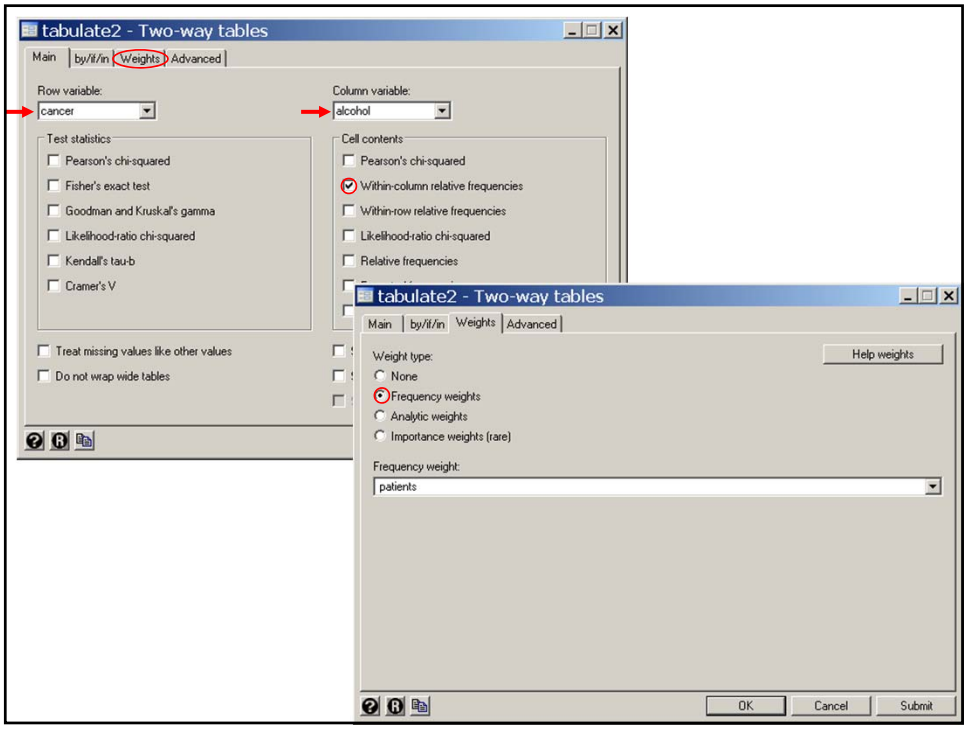
8. Example: Effect of Dose of Alcohol and Tobacco on Esophageal Cancer Risk

The Ille-et-Vilaine data set provides four different levels of consumption for both alcohol and tobacco. To investigate the joint effects of dose and alcohol on esophageal cancer risk we first tabulate the raw data.

```
. *  5.11.1.EsophagealCa.ClassVersion.log
. *
. *  Estimate age-adjusted risk of esophageal cancer due to dose of alcohol.
. *
. use C:\WDDtext\5.5.EsophagealCa.dta, clear
. *
. *  Show frequency tables of effect of dose of alcohol on esophageal cancer.
. *
```

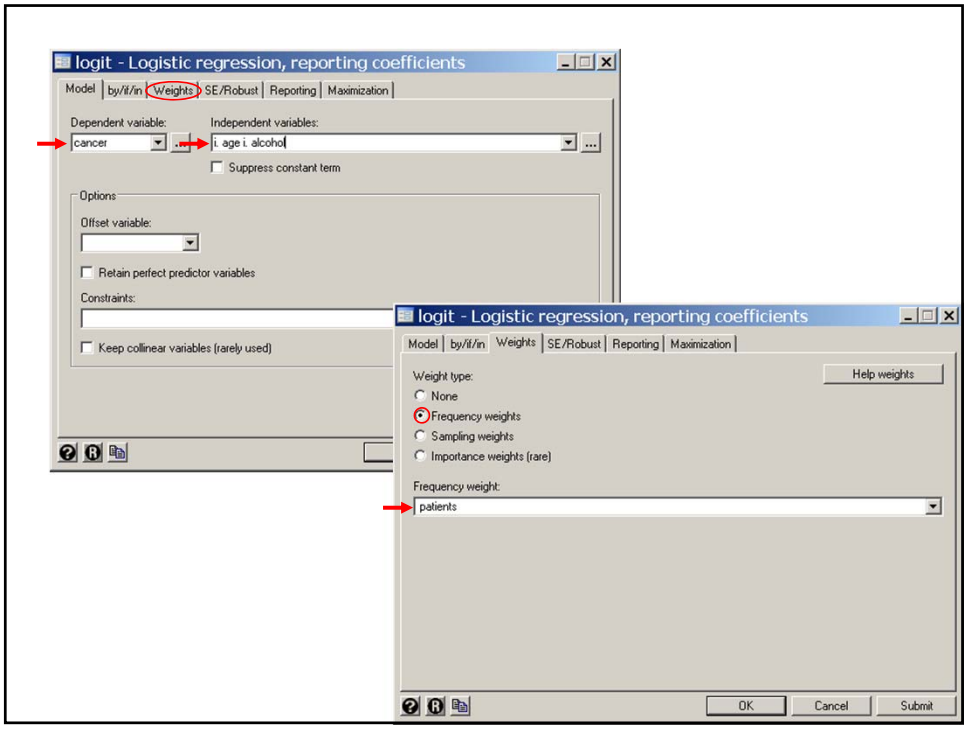
```
. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate cancer alcohol [freq=patients] , column                    {1}


+-------------------+
| Key               |
|-------------------|
|     frequency     |
| column percentage |
+-------------------+


Esophageal | Alcohol (gm/day)
Cancer     |    0-39   40-79   80-119   >= 120 |    Total
-----------+------------------------------------+--------
        No |     386      280        87        22 |      775
           |   93.01    78.87     63.04     32.84 |    79.49
-----------+------------------------------------+--------
       Yes |      29       75        51        45 |      200
           |    6.99    21.13     36.96     67.16 |    20.51
-----------+------------------------------------+--------
     Total |     415      355       138        67 |      975
           |  100.00   100.00    100.00    100.00 |   100.00
```

{1} The **tabulate** command produces one- and two-way frequency tables. The *column* option produces percentages of observations in each column.



```
. * Statistics > Binary outcomes > Logistic regression
. logit cancer i.age i.alcohol [freq=patients]

Logit estimates                                   No. of obs      =        975
                                                  LR chi2(8)      =     274.07
                                                  Prob > chi2     =     0.0000
Log likelihood   = -363.7080768                   Pseudo R2       =     0.2649

------------------------------------------------------------------------------
      cancer |      Coef.   Std. Err.       z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
          2  |   1.631112   1.080013     1.51   0.131    -.4856742    3.747899
          3  |   3.425834   1.038937     3.30   0.001     1.389555    5.462114
          4  |   3.943447   1.034622     3.81   0.000     1.915624    5.971269
          5  |   4.356767   1.041336     4.18   0.000     2.315786    6.397747
          6  |   4.424219     1.0914     4.05   0.000     2.285115    6.563324
             |
     alcohol |
          2  |    1.43431   .2447858     5.86   0.000     .9545384    1.914081 {2}
          3  |    2.00711   .2776153     7.23   0.000     1.462994    2.551226
          4  |   3.680012   .3763372     9.78   0.000     2.942405    4.417619
             |
       _cons |  -6.147181   1.041877    -5.90   0.000    -8.189223    -4.10514
------------------------------------------------------------------------------
```

{2} The parameter estimates of ***2.alcohol,*** *3.***alcohol*** and *4.***alcohol*** estimate the log-odds ratio for cancer associated with alcohol doses of **40-79** gm/day, **80-119** gm/day and **120+** gm/day, respectively. These log-odds ratios are derived with respect to people who drank **0-39** grams a day. They are all adjusted for age. All of these statistics are significantly different from zero ($P<0.0005$).




```
. * Statistics > Postestimation > Linear combinations of estimates
. lincom 3.alcohol - 2.alcohol, or                                        {3}

  ( 1) - [cancer] 2.alcohol + [cancer]3.alcohol = 0.0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   1.773226   .4159625     2.44   0.015     1.119669    2.808268
------------------------------------------------------------------------------
```

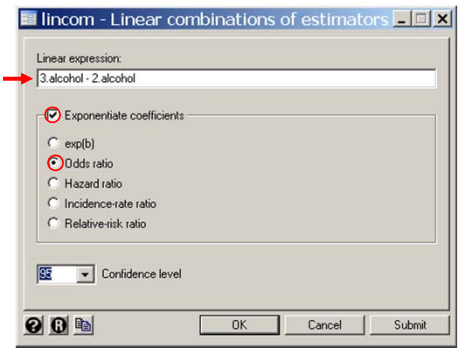
{3} In general, **lincom** calculates any **linear combination** of parameter estimates, tests the **null hypothesis** that the true value of this combination equals zero, and gives a **95% confidence interval** for this estimate.

The *or* option exponentiates the linear combination and calculates the corresponding confidence interval.

In this example ***3.alcohol*** – ***2.alcohol*** equals the log-odds ratio for cancer associated with drinking 8-119 gm/day compared to 40-79 gm/day. ***3.alcohol*** – ***2.alcoh*** = **2.001 – 1.434 = 0.573**, which is significantly different from zero with $P = 0.015$. The corresponding odds ratio is

exp[0.573] = 1.77. The 95% confidence interval for this difference is (1.1 – 2.8).

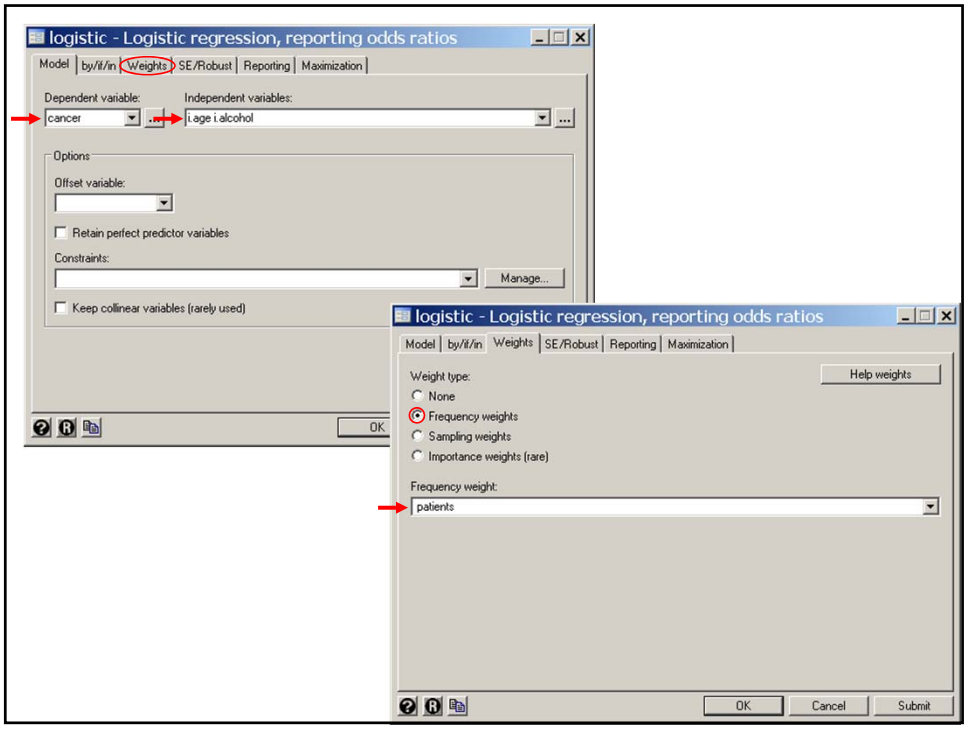
Note that the null hypothesis that a **log-odds ratio** equals **zero** is equivalent to the null hypothesis that the corresponding **odds ratio** equals **one.**




```
. lincom 4.alcohol - 3.alcohol, or

 ( 1) [cancer]3.alcohol + [cancer]4.alcohol = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   5.327606    1.95176     4.57   0.000     2.598339    10.92367
------------------------------------------------------------------------------
```

```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logisitc cancer i.age i.alcohol [freq=patients]                    {4}

Logit estimates                                   No. of obs      =       975
                                                  LR chi2(8)      =    274.07
                                                  Prob > chi2     =    0.0000
                                                  Pseudo R2       =    0.2649

Log likelihood   = -363.7080768
------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
           2 |   5.109555   5.518386     1.51   0.131     .6152822    42.43183
           3 |   30.74829   31.94554     3.30   0.001     4.013065    235.5949
           4 |   51.59613    53.3825     3.81   0.000     6.791178    392.0027
           5 |   78.00451   81.22889     4.18   0.000     10.13289    600.4908
           6 |   83.44761   91.07472     4.05   0.000     9.826812     708.623
             |
     alcohol |
           2 |   4.196747   1.027304     5.86   0.000     2.597471    6.780704
           3 |   7.441782   2.065953     7.23   0.000     4.318873    12.82282
           4 |   39.64687   14.92059     9.78   0.000     18.96139    82.8987
------------------------------------------------------------------------------
```

{4} *logistic* directly calculate the age adjusted odds ratio and 95% confidence interval for alcohol level 2 vs. level 1, level 3 vs. level 1 and level 4 vs. level 1.



By default, Stata includes a constant term in its regression models.

For this reason, when we convert a categorical variable into a number of indicator covariates we always have to leave one of the categories out to avoid multicolinearity.

For example, let

$$sex = \begin{cases} 1 \text{ for men} \\ 2 \text{ for women} \end{cases} \qquad 1.sex = \begin{cases} 1 \text{ for men} \\ 0 \text{ for women} \end{cases} \qquad 2.sex = \begin{cases} 0 \text{ for men} \\ 1 \text{ for women} \end{cases}$$

Then the linear predictor $\alpha + \beta_1 1.sex + \beta_2 2.sex$ takes the values

$\alpha + \beta_1$ for men and $\alpha + \beta_2$ for women.

This gives us three parameters to model the effects of two sexes. To obtain uniquely defined parameter estimates we must use one of the following models:

$$\beta_1 1.sex + \beta_2 2.sex$$

$$\alpha + \beta_2 2.sex$$

or

$$\alpha + \beta_1 1.sex$$

By default, the Stata syntax `i.varname` defines indicator covariates for all but the smallest value of `varname`.

If `varname` takes the values 1, 3, 5 and 10 and we want indicator covariates defined for each of these values except 5 we can use the syntax

`ib5.varname`

`5.11.EsophagealCa.ClassVersion.log` continues as follows.

```
. logistic cancer i.age ib2.alcohol [freq=patients]                          {5}

Logistic regression                               Number of obs   =        975
                                                  LR chi2(8)      =     262.07
                                                  Prob > chi2     =     0.0000
Log likelihood = -363.70808                       Pseudo R2       =     0.2649

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
          2  |   5.109555   5.518386     1.51   0.131     .6152822    42.43183
          3  |   30.74829   31.94554     3.30   0.001     4.013065    235.5949
          4  |   51.59613    53.3825     3.81   0.000     6.791178    392.0027
          5  |   78.00451   81.22889     4.18   0.000     10.13289    600.4908
          6  |   83.44761   91.07472     4.05   0.000     9.826812     708.623
             |
     alcohol |
          1  |   .2382798   .0583275    -5.86   0.000     .1474773    .3849898
          3  |   1.773226   .4159625     2.44   0.015     1.119669    2.808268 {6}
          4  |   9.447049   3.239241     6.55   0.000     4.824284    18.49948
------------------------------------------------------------------------------
```

{5} **ib2.alcohol** instructs Stata to include indicator covariates for each value of **alcohol** except **alcohol** = 2. This makes an alcohol value of 2 the baseline for odds ratios associated with this variable.

{6} The odds rato for level 3 drinkers compared to level 1 drinkers is 1.77, which is identical to the odds ratio obtained from the earlier **lincom** statement.

9. Making Inferences About Odds Ratio Derived from Multiple Parameters

In more complex multiple logistic regression models we need to make inferences about odds ratios that are estimated from multiple parameters.

A simple example was given in the preceding example where the log odds ratio for cancer associated with alcohol level 3 compared to alcohol level 2 was of the form

$$\beta_3 - \beta_2$$

To derive confidence intervals and perform hypothesis tests we need to be able to compute the standard errors of weighted sums of parameter estimates.

10. Estimating The Standard of Error of a Weighted Sum of Regression Coefficients

Suppose that we have a model with q parameters.

Let $b_1, b_2, \ldots, b_q$ be estimates of parameters $\beta_1, \beta_2, \ldots, \beta_q$

Let $c_1, c_2, \ldots, c_q$ be a set of known weights and let

$$f = \sum c_j b_j$$

For example, in the preceding logistic regression model there are 5 age parameters (*2.age*, *3.age*, …, *6.age*), three alcohol parameters (*2.alcohol*, *3.alcohol*, *4.alcohol*) and one constant parameter for a total of $q = 9$ parameters. Let us rename these parameters so that β_2 and β_3 represent *2.alcohol* and *3.alchol*, respectively.

Let

$$c_3 = 1, c_2 = -1, \text{ and } c_1 = c_4 = c_5 = \ldots = c_9 = 0$$

Then $f = b_3 - b_2 = 2.0071 - 1.4343 = 0.5728$

And exp (f) = exp(0.5728) = 1.773 is the odds ratio of level 3 drinkers relative to level 2 drinkers.

Let s_{jj} be the estimated variance of b_j: $j = 1, \ldots, q$ and let s_{ij} be the covariance of b_i and b_j for any $i \neq j$.

Then the variance of f equals:

$$s_f^2 = \sum_{i=1}^{q} \sum_{j=1}^{q} c_i c_j s_{ij} \qquad \{4.6\}$$

For large studies the 95% confidence interval for f is

$$f \pm 1.96 * \sqrt{s_f^2} = f \pm 1.96 s_f$$

When f estimates a log-odds ratio then the corresponding odds ratio is estimated by $\exp(f)$ with 95% confidence interval $\left[\exp(f - 1.96 s_f), \exp(f + 1.96 s_f)\right]$

11. The Estimated Variance-Covariance Matrix

The estimates of s_{ij} are written in a square array

$$\begin{bmatrix} s_{11}, & s_{12} & \ldots., & s_{1q} \\ s_{21}, & s_{22} & \ldots., & s_{2q} \\ . & & & . \\ . & & & . \\ . & & & . \\ s_{q1}, & s_{q2}, & \ldots, & s_{qq} \end{bmatrix}$$

which is called the estimated variance-covariance matrix.

In our example comparing level 3 drinkers to level 2 drinkers

$$s_f^2 = s_{33} + s_{22} - 2s_{23}$$

which gives $s_f = 0.2346$; this is the standard error of *3.alcohol –2.alcohol* given in the preceding example.

a) Estimating the Variance-Covariance Matrix with Stata

You can obtain the **variance-covariance matrix** in Stata using the ***estat vce*** post estimation command. However, the ***lincom*** command is so powerful and flexible that we will usually not need to do this explicitly. If you are working with other statistical packages you may need to calculate equation {4.6} explicitly

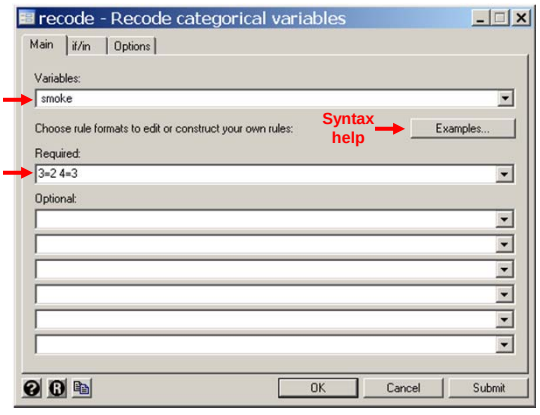
12. Example: Effect of Dose of Tobacco on Esophageal Cancer Risk

```
. *  5.12.EsophagealCa.ClassVersion.do
. *
. *  Estimate age-adjusted risk of esophageal cancer due to dose of tobacco.
. *
. use C:\WDDtext\5.5.EsophagealCa.dta, clear
. * Statistics > Summaries... > Tables > Two-way tables with measures...
. tabulate cancer tobacco [freq=patients] , column

+--------------------+
| Key                |
|--------------------|
|     frequency      |
| column percentage  |
+--------------------+
Esophageal |              Tobacco (gm/day)
    Cancer |       0-9      10-19      20-29      >= 30 |     Total
-----------+--------------------------------------------+----------
        No |       447        178         99         51 |       775
           |     85.14      75.42      75.00      62.20 |     79.49
-----------+--------------------------------------------+----------
       Yes |        78         58         33         31 |       200
           |     14.86      24.58      25.00      37.80 |     20.51
-----------+--------------------------------------------+----------
     Total |       525        236        132         82 |       975
           |    100.00     100.00     100.00     100.00 |    100.00
```

```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logisitic cancer i.age i.tobacco [freq=patients]

Logit regression                                  No. of obs      =        975
                                                  LR chi2(8)      =     157.68
                                                  Prob > chi2     =     0.0000
                                                  Pseudo R2       =     0.1594

Log likelihood   = -415.90235
------------------------------------------------------------------------------
     cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        age |
          2 |   6.035932   6.433686     1.69   0.092     .7472235    48.75713
          3 |   36.20831   37.10835     3.50   0.000     4.857896    269.8785
          4 |   61.79318   63.10432     4.04   0.000     8.349838    457.3019
          5 |   83.56952   85.86437     4.31   0.000     11.15506    626.0713
          6 |   60.45383   64.52449     3.84   0.000     7.462882    489.7124
            |
    tobacco |
          2 |   1.835482   .3781838     2.95   0.003     1.225655    2.748731 {1}
          3 |   1.945172    .487733     2.65   0.008     1.189947    3.179717
          4 |   5.706139   1.725688     5.76   0.000     3.154398     10.3221
------------------------------------------------------------------------------
```

{1} Note how **similar** the log-odds ratios for the **2nd** and **3rd** levels of tobacco exposure. If we had assigned a single parameter for ***tobacco*** we would have badly **overestimated** the odds ratio between levels **2** and **3**, and badly **underestimated** the odds ratio between levels **1** and **2** and between levels **3** and **4**.

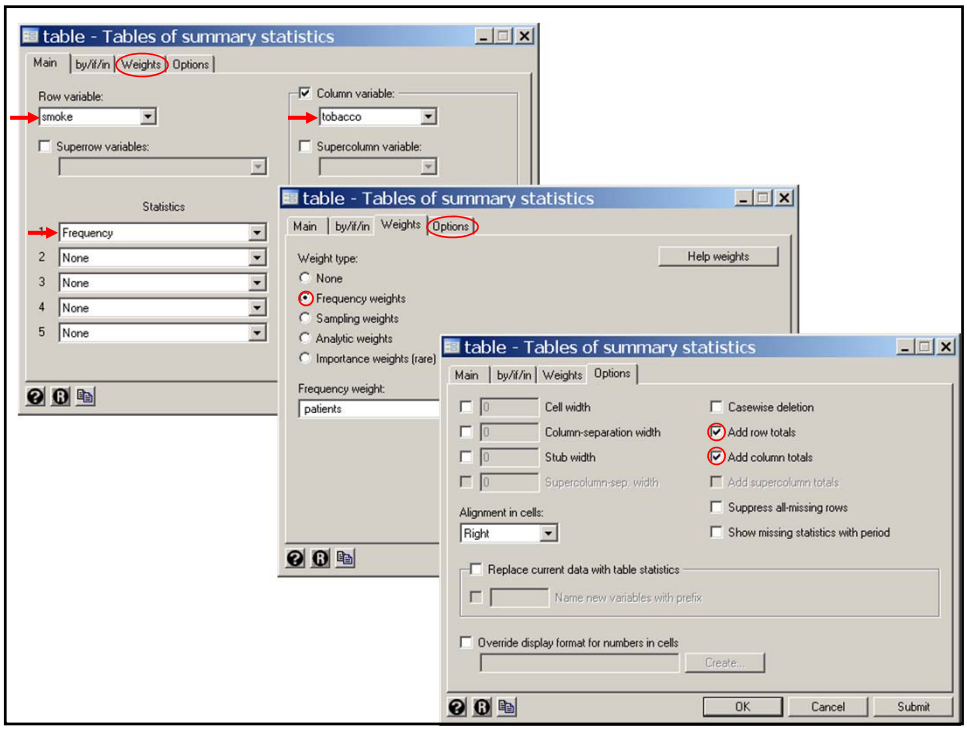
```
. generate smoke = tobacco

. * Data > Create... > Other variable-transformation... > Recode catigorical...
. recode smoke 3=2 4=3                                                    {2}
(96 changes made)
```



{2} We want to **combine** the 2nd and 3rd levels of tobacco exposure. We do this by defining a new variable called ***smoke*** that is identical to *tobacco* and then using the ***recode*** statement, which in this example changes values of *smoke* = 3 to *smoke* = 2, and values of *smoke* = 4 to *smoke* = 3.

```
. label variable smoke "Smoking (gm/day)"

. label define  smoke 1 "0-9" 2 "10-29" 3 ">= 30"

. label values smoke smoke

. * Statistics > Summaries... > Tables > Table of summary statistics (table).
. table smoke tobacco [freq=patients], row col                          {3}
-------------------------------------------
Smoking    |          Tobacco (gm/day)
(gm/day)   |   0-9  10-19  20-29  >= 30  Total
-----------+-------------------------------
      0-9  |   525                        525
    10-29  |          236    132         368
    >= 30  |                      82     82
           |
    Total  |   525    236    132     82   975
-------------------------------------------
```

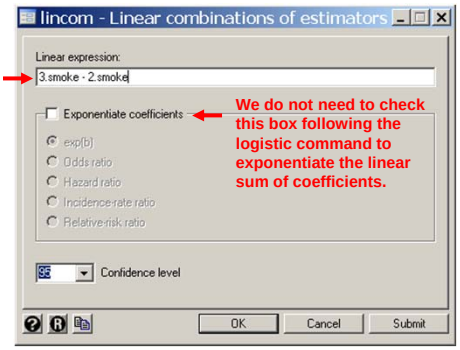
{3} This **table** statement shows that the previous *recode* statement worked.



```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic cancer i.age i.smoke [freq=patients]

Logistic regression                               Number of obs   =        975
                                                  LR chi2(7)      =     157.64
                                                  Prob > chi2     =     0.0000
Log likelihood = -415.92589                       Pseudo R2       =     0.1593

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |
          2  |   6.037092   6.434914     1.69   0.092      .7473691    48.76637
          3  |    36.2117   37.11182     3.50   0.000       4.85835    269.9038
          4  |   61.79965   63.11096     4.04   0.000      8.350705    457.3503
          5  |   83.52177   85.81492     4.31   0.000      11.14879    625.7078
          6  |   60.25337   64.30389     3.84   0.000      7.439742    487.9831
             |
       smoke |
          2  |   1.873669   .3421356     3.44   0.001      1.309972    2.679933  {4}
          3  |   5.704954   1.725242     5.76   0.000      3.153836    10.31965
------------------------------------------------------------------------------
. lincom 3.smoke - 2.smoke                                                  {5}

  ( 1)  - [cancer]2.smoke + [cancer]3.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   3.044803   .9116935     3.72   0.000      1.693118    5.475593
------------------------------------------------------------------------------
```

{4} There is a marked trend of **increasing** cancer **risk** with **increasing dose** of tobacco. Men who smoked 10-29 grams a day had **1.87** times the cancer risk of men who smoked less. Men who smoked more than 29 gm/day had **5.7** times the cancer risk of men who smoked less than 10 grams a day.

{5} The **odds ratio** for ≥ 30 gm/day of tobacco relative to 10-29 gm/day is 3.04 and is highly **significant**.



The next question is how do alcohol and tobacco interact on esophageal cancer risk?

```
. * 5.20.EsophagealCa.ClassVersionlog
. *
. * Regress esophageal cancers against age and dose of alcohol
. * and tobacco using a multiplicative model.
. *
. use 5.5.EsophagealCa.dta, clear
. sort tobacco

. * Statistics > Summaries... > Tables > Two-way tables with measures...
. by tobacco: tabulate cancer alcohol [freq=patients]
> , column                                                    {1}

-> tobacco=      0-9
Esophageal | Alcohol (gm/day)
Cancer     |    0-39    40-79   80-119   >= 120 |   Total
-----------+--------------------------------------+--------
        No |     252      145       42         8 |     447
           |   96.55    81.01    68.85     33.33 |   85.14
-----------+--------------------------------------+--------
       Yes |       9       34       19        16 |      78
           |    3.45    18.99    31.15     66.67 |   14.86
-----------+--------------------------------------+--------
     Total |     261      179       61        24 |     525
           |  100.00   100.00   100.00    100.00 |  100.00
```

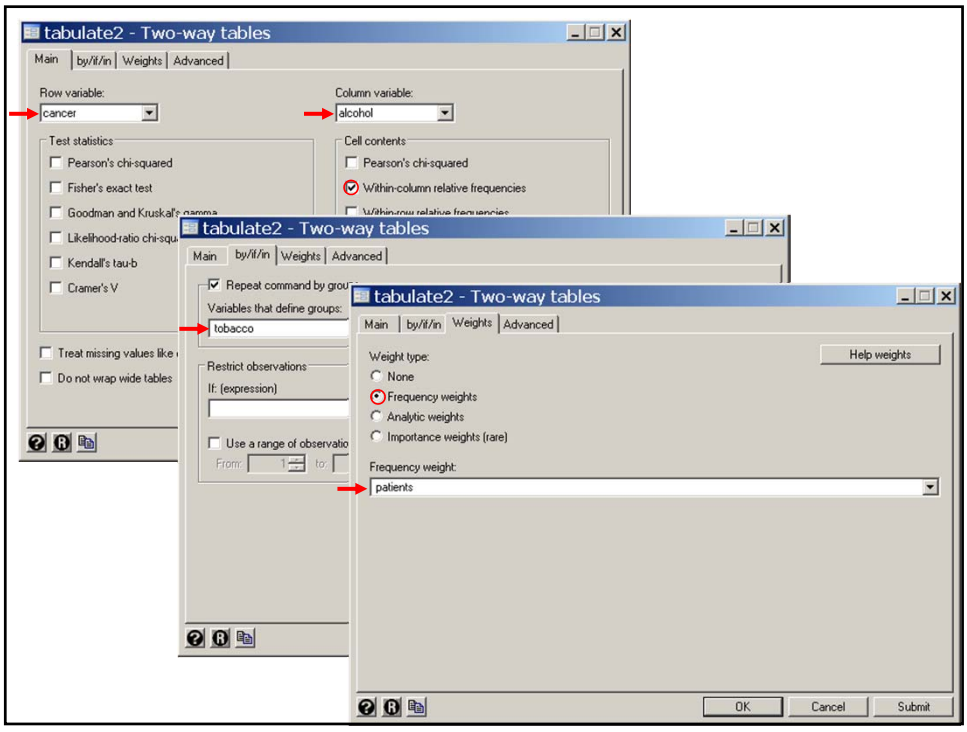
{1} ***by tobacco***: produces separate frequency tables for each value of *tobacco*. The data set must first be **sorted** by *tobacco*.

```
-> tobacco=    10-19
Esophageal | Alcohol (gm/day)
Cancer     |    0-39    40-79   80-119   >= 120 |   Total
-----------+--------------------------------------+--------
        No |      74       68       30         6 |     178
           |   88.10    80.00    61.22     33.33 |   75.42
-----------+--------------------------------------+--------
       Yes |      10       17       19        12 |      58
           |   11.90    20.00    38.78     66.67 |   24.58
-----------+--------------------------------------+--------
     Total |      84       85       49        18 |     236
           |  100.00   100.00   100.00    100.00 |  100.00

-> tobacco=   20-29
Esophageal | Alcohol (gm/day)
Cancer     |    0-39    40-79   80-119   >= 120 |   Total
-----------+--------------------------------------+--------
        No |      37       47       10         5 |      99
           |   88.10    75.81    62.50     41.67 |   75.00
-----------+--------------------------------------+--------
       Yes |       5       15        6         7 |      33
           |   11.90    24.19    37.50     58.33 |   25.00
-----------+--------------------------------------+--------
     Total |      42       62       16        12 |     132
           |  100.00   100.00   100.00    100.00 |  100.00
```

```
-> tobacco=    >= 30
Esophageal | Alcohol (gm/day)
Cancer     |    0-39    40-79   80-119   >= 120 |    Total
-----------+------------------------------------+--------
        No |      23       20        5        3 |       51
           |   82.14    68.97    41.67    23.08 |    62.20
-----------+------------------------------------+--------
       Yes |       5        9        7       10 |       31
           |   17.86    31.03    58.33    76.92 |    37.80
-----------+------------------------------------+--------
     Total |      28       29       12       13 |       82
           |  100.00   100.00   100.00   100.00 |   100.00
```

These tables show that the proportion of study subjects with cancer increases dramatically with increasing alcohol consumption for every level of tobacco consumption.

The proportion of cases also increases with increasing tobacco consumption for most levels of alcohol.



13. Multiplicative Model of Effect of Smoking and Alcohol on Esophageal Cancer Risk

Suppose that subjects either were or were not exposed to alcohol and tobacco and we did not include age in our model. Consider the model

$$\text{logit}\left(E\left(d_{ij} / m_{ij}\right)\right) = \alpha + x_i\beta_1 + y_j\beta_2$$

where $i = \begin{cases} 1: \text{ if patient drank} \\ 0: \text{ Otherwise} \end{cases}$

$j = \begin{cases} 1: \text{ if patient smoked} \\ 0: \text{ Otherwise} \end{cases}$

$x_i = i$
$y_j = j$

m_{ij} is the number of subjects with drinking status i and smoking status j.

d_{ij} is the number of cancers with drinking status i and smoking status j.

α, β_1 and β_2 are model parameters.

Thus the log-odds of a drinker with smoking status j is

$$\text{logit}\left(E\left(d_{1j} / m_{1j}\right)\right) = \alpha + \beta_1 + y_j\beta_2 \qquad \textbf{\{4.7\}}$$

The log-odds of a non-drinker with smoking status j is

$$\text{logit}\left(E\left(d_{0j} / m_{0j}\right)\right) = \alpha + y_j\beta_2$$

Subtracting equation {4.8} from {4.7} gives that **{4.8}**

$$\log\left(\frac{\pi_{1j} / (1-\pi_{1j})}{\pi_{0j} / (1-\pi_{0j})}\right) = \beta_1$$

where π_{ij} is the probability that someone with drinking status i and smoking status j develops cancer.

In other words, exp(β_1) is the odds ratio for cancer in drinkers compared to non-drinkers adjusted for smoking.

Note that this implies that the relative risk of drinking is the same in smokers and non-smokers.

By an identical argument, exp(β_2) is the odds ratio for cancer in smokers compared to non-smokers adjusted for drinking.

For people who both drink and smoke the model is

$$\text{logit}(E(d_{11}/m_{11})) = \alpha + \beta_1 + \beta_2 \qquad \textbf{\{4.9\}}$$

while for people who neither drink nor smoke the model is

$$\text{logit}(E(d_{00}/m_{00})) = \alpha \qquad \textbf{\{4.10\}}$$

Subtracting {4.9} from {4.10} give that the log-odds ratio for people who both smoke and drink relative to those who do neither is $\beta_1 + \beta_2$, and the corresponding odds ratio is exp(β_1) × exp(β_2).

Thus our model implies that the odds ratio of having both risk factors equals the product of the individual odds ratio for drinking and smoking.

It is for this reason that this is called a **multiplicative model**.

The multiplicative assumption is a very strong one that is often not justified. Let us see how it works with the Ille-et-Vilaine data set.

```
. *
. *  Regress cancer against age, alcohol and smoke.
. *  Use a multiplicative model
. *
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic cancer i.age i.alcohol i.smoke [freq=patients]              {1}

Logistic regression                               Number of obs   =        975
                                                  LR chi2(10)     =     285.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -351.96823                       Pseudo R2       =     0.2886
```

{1} This command fits a model with a constant parameter, **5 age** parameters **3 alcohol** parameters and **two tobacco** parameters. No parameter is given for the lowest strata associated with *age, alcohol* or *smoke*.

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]	
age							
2	7.262526	8.017757	1.80	0.073	.834391	63.21291	
3	43.65627	46.62635	3.54	0.000	5.381893	354.1263	
4	76.3655	81.33339	4.07	0.000	9.469377	615.8472	
5	133.7632	143.9793	4.55	0.000	16.22277	1102.93	
6	124.4262	139.5094	4.30	0.000	13.82058	1120.205	
alcohol							
2	4.213304	1.05191	5.76	0.000	2.582905	6.872854	{2}
3	7.222005	2.053957	6.95	0.000	4.135936	12.61077	
4	36.7912	14.17012	9.36	0.000	17.29434	78.26794	
smoke							
2	1.592701	.3200884	2.32	0.021	1.074154	2.361577	
3	5.159309	1.775207	4.77	0.000	2.628521	10.12679	

{2} The **odds ratio** for level **2** drinkers relative to level **1** drinkers **adjusted** for **age** and **smoking** is **4.21**.

```
. lincom 2.alcohol + 2.smoke

 ( 1)  [cancer]2.alcohol + [cancer]2.smoke = 0
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]	
(1)	6.710535	2.110331	6.05	0.000	3.623022	12.4292	{3}

```
. lincom 3.alcohol + 2.smoke

  ( 1)  [cancer]3.alcohol + [cancer]2.smoke = 0
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	11.5025	3.877641	7.25	0.000	5.940747	22.27118

```
. lincom 4.alcohol + 2.smoke

 ( 1)  [cancer]4.alcohol + [cancer]2.smoke = 0
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
(1)	58.59739	25.19568	9.47	0.000	25.22777	136.1061

{3} The cancer log-odds for a man in, say, the third age strata who is a level 2 drinker and level 2 smoker is

_cons + 3.age + 2.alcohol + 2.smoke

The cancer log-odds for a man in the same age strata who is a level 1 drinker and level 1 smoker is

_cons + 3.age

Subtracting these two log-odds and exponentiating gives that the odds ratio for men who are **both** level **2** drinkers and level **2** smokers relative to those who are level **1** drinkers and level **1** smokers is **6.71.**

```
. lincom 2.alcohol + 3.smoke

 ( 1)  [cancer]2.alcohol + [cancer]3.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   21.73774   9.508636     7.04   0.000     9.223106    51.23319
------------------------------------------------------------------------------

. lincom 3.alcohol + 3.smoke

 ( 1)  [cancer]3.alcohol + [cancer]3.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   37.26056   17.06685     7.90   0.000     15.18324    91.43957
------------------------------------------------------------------------------

. lincom 4.alcohol + 3.smoke

 ( 1)  [cancer]4.alcohol + [cancer]3.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   189.8171   100.9788     9.86   0.000     66.91353    538.4643
------------------------------------------------------------------------------
```

The preceding analyses are summarized in the following table.

Note that the multiplicative assumption holds.

E.g. $36.8 \times 5.16 = 190$

Table 4.1. Effect of Alcohol and Tobacco on Esophageal Cancer Risk

Multiplicative Model -- Adjusted to Age

Daily Alcohol Comsumption	Daily Tobacco Consumption					
	0-9 gm		10-29 gm		30gm	
	Odds Ratio	95% CI	Odds Ratio	95% CI	Odds Ratio	95% CI
0-39 gm	1.0*		1.59	(1.1 - 2.4)	5.16	(2.6 - 10)
40-79 gm	4.21	(2.6 - 6.9)	6.71	(3.6 - 12)	21.7	(9.2 - 51)
80-119 gm	7.22	(4.1 - 13)	11.5	(5.9 - 22)	37.3	(15 - 91)
120 gm.	36.8	(17 - 78)	58.6	(25 - 140)	190	(67 - 540)

* Denominator of odds ratios

This model suggests that combined heavy alcohol and tobacco consumption has an enormous effect on the risk of esophageal cancer.

To determine if this is real or a model artifact we need to look at a model that permits the cancer risk associated with combined risk factors to deviate from the multiplicative model.

14. Modeling the Effect of Alcohol and Tobacco on Cancer Risk with Interaction

Let us first return to the simple example where people either do or do not drink or smoke and where we do not adjust for age. Our multiplicative model was

$$\text{logit}\left(E\left(d_{ij} / m_{ij}\right)\right) = \alpha + x_i\beta_1 + y_j\beta_2 \qquad \textbf{\{4.11\}}$$

We allow alcohol and tobacco to have a synergistic effect on cancer odds by including a fourth parameter as follows

$$\text{logit}\left(E\left(d_{ij} / m_{ij}\right)\right) = \alpha + x_i\beta_1 + y_j\beta_2 + x_i y_j\beta_3 \qquad \textbf{\{4.12\}}$$

Then β_3 only enters the model for people who both smoke and drink. By the usual arguments…

β_1	is the log odds ratio for cancer associated with **alcohol among non-smokers**,
β_2	is the log odds ratio for cancer associated with **smoking among non-drinkers**,
$\beta_1 + \beta_3$	is the log odds ratio for cancer associated with **alcohol among smokers**,
$\beta_1 + \beta_2 + \beta_3$	is the log odds ratio for cancer associated with people who **smoke and drink** *compared* to those who are both **non-smokers and non-drinkers**.

We now apply this interpretation to the esophageal cancer data. *5.20.EsophagelaCa.ClassVersion.log* continues as follows:

```
. *
. *  Regress cancer against age, alcohol and smoke.
. *  Include alcohol-smoke interaction terms.
. *
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic cancer i.age alcohol##smoke [freq=patients],                    {1}

Logistic regression                               Number of obs   =        975
                                                  LR chi2(16)     =     290.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -349.29335                       Pseudo R2       =     0.2940
```

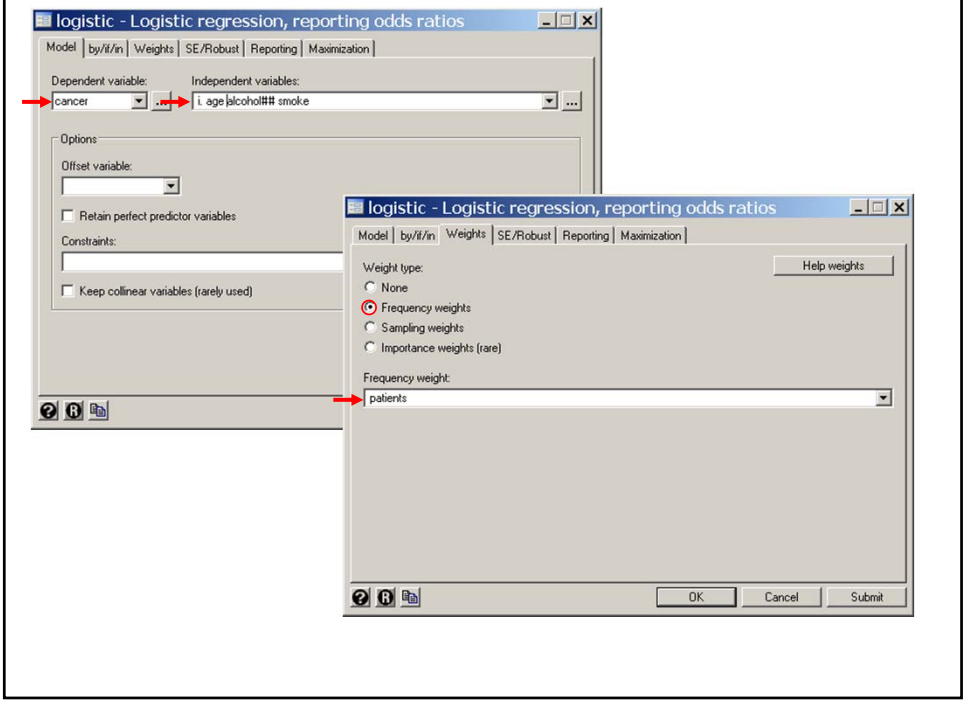
A separate parameter is fitted for each of these variables. In addition, the model specifies 5 parameters for the 5 age indicator variables and a constant parameter.

{1} The syntax ***alcohol##smoke*** defines the following categorical values:

```
      2.alcohol = 1 if alcohol = 2, and = 0 otherwise
      3.alcohol = 1 if alcohol = 3, and = 0 otherwise
      4.alcohol = 1 if alcohol = 4, and = 0 otherwise
      2.smoke   = 1 if smoke = 2,   and = 0 otherwise
      3.smoke   = 1 if smoke = 3,   and = 0 otherwise
   alcohol#smoke
      2 2 = 2.alcohol x 2.smoke
      2 3 = 2.alcohol x 3.smoke
      3 2 = 3.alcohol x 2.smoke
      3 3 = 3.alcohol x 3.smoke
      4 2 = 4.alcohol x 2.smoke
      4 3 = 4.alcohol x 3.smoke
```

cancer	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
age						
2	6.697614	7.41052	1.72	0.086	.7657997	58.57673
3	40.1626	42.67457	3.48	0.001	5.004744	322.3011
4	69.55115	73.73699	4.00	0.000	8.707117	555.5642
5	123.0645	131.6754	4.50	0.000	15.11374	1002.06
6	118.8368	133.2538	4.26	0.000	13.19724	1070.086
alcohol						
2	7.554406	3.043769	5.02	0.000	3.429574	16.64028
3	12.71358	5.825002	5.55	0.000	5.179306	31.20788
4	65.07188	39.54145	6.87	0.000	19.7767	214.108
smoke						
2	3.800862	1.703912	2.98	0.003	1.578671	9.151084
3	8.651205	5.569301	3.35	0.001	2.449667	30.55247
alcohol#smoke						
2 2	.3251915	.1746668	-2.09	0.036	.1134859	.9318294
2 3	.5033299	.4154539	-0.83	0.406	.0998302	2.53772
3 2	.3341452	.2008274	-1.82	0.068	.1028839	1.085233
3 3	.657279	.6598915	-0.42	0.676	.0918681	4.702563
4 2	.3731549	.301804	-1.22	0.223	.076462	1.821095
4 3	.3489097	.4210291	-0.87	0.383	.032777	3.714132

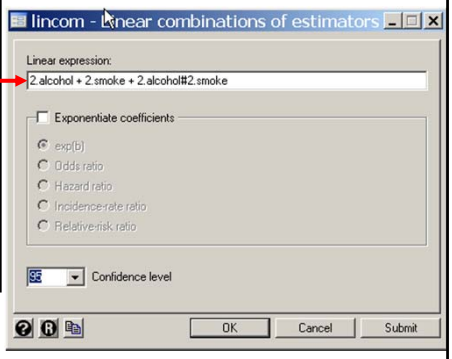
The highlighted odds ratios show age adjusted risks of drinking among level 1 smokers and smoking among level 1 drinkers



```
. lincom 2.alcohol + 2.smoke + 2.alcohol#2.smoke                    {2}

 ( 1)  [cancer]2.alcohol + [cancer]2.smoke + [cancer]2.alcohol#2.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   9.337306   3.826162     5.45   0.000     4.182379    20.84586
------------------------------------------------------------------------------
```

{2} This statement calculates the odds ratio for men in the second strata of *alcohol* and *smoke* relative to men in the first strata of both of these variables. This odds ratio of **9.33** is adjusted for age.

2.alcohol#2.smoke represents the parameter associated with the product of the covariates *2.alcohol* and *2.smoke*.




```
. lincom 2.alcohol + 3.smoke + 2.alcohol#3.smoke

 ( 1)  [cancer]2.alcohol + [cancer]3.smoke + [cancer]2.alcohol#3.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   32.89498   19.73769     5.82   0.000     10.14824    106.6274
------------------------------------------------------------------------------

. lincom 3.alcohol + 2.smoke + 3.alcohol#2.smoke

 ( 1)  [cancer]3.alcohol + [cancer]2.smoke + [cancer]3.alcohol#2.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   16.14675   7.152595     6.28   0.000     6.776802    38.47207
------------------------------------------------------------------------------
```

```
. lincom 3.alcohol + 3.smoke + 3.alcohol#3.smoke

 ( 1)  [cancer]3.alcohol + [cancer]3.smoke + [cancer]3.alcohol#3.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   72.29267   57.80896     5.35   0.000     15.08098    346.5446
------------------------------------------------------------------------------

. lincom 4.alcohol + 2.smoke + 4.alcohol#2.smoke

( 1)  [cancer]4.alcohol + [cancer]2.smoke + [cancer]4.alcohol#2.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   92.29212   53.97508     7.74   0.000     29.33307    290.3833
------------------------------------------------------------------------------

. lincom 4.alcohol + 3.smoke + 4.alcohol#3.smoke

 ( 1)  [cancer]4.alcohol + [cancer]3.smoke + [cancer]4.alcohol#3.smoke = 0

------------------------------------------------------------------------------
      cancer | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         (1) |   196.4188   189.1684     5.48   0.000     29.74417    1297.072
------------------------------------------------------------------------------
```

The following table summarizes the results of this analysis

Table 4.2. Effect of Alcohol and Tobacco on Esophageal Cancer Risk

Model with all 2-Way Interaction Terms -- Adjusted for Age

Daily Alcohol Comsumption	Daily Tobacco Consumption					
	0 – 9 gm		10 – 29 gm		≥ 30 gm	
	Odds Ratio	95% Confidence Interval	Odds Ratio	95% Confidence Interval	Odds Ratio	95% Confidence Interval
0 – 39 gm	1.0*		3.8	(1.6 – 9.2)	8.65	(2.4 – 31)
40 – 79 gm	7.55	(3.4 – 17)	9.34	(4.2 – 21)	32.9	(10 – 110)
80 – 119 gm	12.7	(5.2 – 31)	16.1	(6.8 – 38)	72.3	(15 – 350)
≥ 120 gm	65.1	(20 – 210)	92.3	(29 – 290)	196	(30 – 1300)

* Denominator of odds ratios

Tables 4.1 and 4.2 are quite **consistent**, and both indicate a dramatic increase in risk with increased drinking and smoking. Note that the **confidence intervals** are **wide**, particularly for the most heavily exposed subjects. The confidence intervals are **wider** in Table 4.2 because they are derived from a model with **more parameters.**

Which model is better?

Table 4.1. Effect of Alcohol and Tobacco on Esophageal Cancer Risk

Multiplicative Model -- Adjusted to Age

Daily Alcohol Comsumption	Daily Tobacco Consumption					
	0-9 gm		10-29 gm		30gm	
	Odds Ratio	95% CI	Odds Ratio	95% CI	Odds Ratio	95% CI
0-39 gm	1.0*		1.59	(1.1 - 2.4)	5.16	(2.6 - 10)
40-79 gm	4.21	(2.6 - 6.9)	6.71	(3.6 - 12)	21.7	(9.2 - 51)
80-119 gm	7.22	(4.1 - 13)	11.5	(5.9 - 22)	37.3	(15 - 91)
120 gm.	36.8	(17 - 78)	58.6	(25 - 140)	190	(67 - 540)

* Denominator of odds ratios

15. Model Fitting: Nested Models and Model Deviance

A model is said to be **nested** within a second model if the first model is a special case of the second.

For example, the multiplicative model {4.11} discussed before was

$$\text{logit}\left(E\left(d_{ij}/m_{ij}\right)\right) = \alpha + x_i\beta_1 + y_j\beta_2$$

while model {4.12} contained an interaction term and was

$$\text{logit}\left(E\left(d_{ij}/m_{ij}\right)\right) = \alpha + x_i\beta_1 + y_j\beta_2 + x_i y_j\beta_3$$

Model {4.11} is nested within model {4.12} since model {4.11} is a special case of model {4.12} with $\beta_3 = 0$.

The model **Deviance** D is a statistic derived from the likelihood function that measures goodness of fit of the data to a specific model. Let $\log(L)$ denote the maximum value of the log likelihood function. Then the deviance is given by

$$D = \text{K} - 2\log(L) \qquad \textbf{\{4.13\}}$$

for some constant K that is independent of the model parameters.

If the model is correct then for large sample sizes D has a χ^2 distribution with degrees of freedom equal to the number of observations minus the number of parameters. Regardless of the true model, D is a non-negative number. Large values of D indicate poor model fit; a perfect fit has $D = 0$.

Suppose that D_1 and D_2 are deviances from two models with model 1 nested in model 2. Then it can be shown that if model 1 is true then $\Delta D = D_1 - D_2$ has an approximately χ^2 distribution with the number of degrees of freedom equal to the number of parameters in model 2 minus the number of parameters in model 1.

Equivalently, $\Delta D = D_1 - D_2$

$$= K - 2\log(L_1) - (K - 2\log(L_2)) = 2(\log(L_2) - \log(L_1))$$

We use the reduction in deviance as a guide to building reasonable models for our data.

For example, in the multiplicative model of alcohol and tobacco levels analyzed above the log likelihood was

$\log(L) = -351.96823$

```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic cancer i.age i.alcohol i.smoke [freq=patients]

Logistic regression                               Number of obs   =        975
                                                  LR chi2(10)     =     285.55
                                                  Prob > chi2     =     0.0000
Log likelihood = -351.96823                       Pseudo R2       =     0.2886
```

The corresponding model with the 6 interaction terms has a log likelihood of

$\log(L) = -349.29335$

```
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic cancer i.age alcohol##smoke [freq=patients],

Logistic regression                               Number of obs   =        975
                                                  LR chi2(16)     =     290.90
                                                  Prob > chi2     =     0.0000
Log likelihood = -349.29335                       Pseudo R2       =     0.2940
```

For example, in the multiplicative model of alcohol and tobacco levels analyzed above the log likelihood was

$\log(L_1) = -351.96823$

The corresponding model with the 6 interaction terms has a log likelihood of

$\log(L_2) = -349.29335$

$\Delta D = 2\left(\log(L_2) - \log(L_1)\right)$

$= 2(-349.29335 + 351.96823)$

$= 5.35$

Since there are 6 more parameters in the interactive model than the multiplicative model, has a χ^2 distribution with 6 degrees of freedom under the independent model. We calculate the P value in Stata with the command

display chi2tail(6, 5.34976)

which gives P = .50.

Thus there is no statistical evidence to suggest that the multiplicative model is false, or that any meaningful improvement in the model fit can be obtained by adding interaction terms to the model.

So what results should we publish – **Table 4.1 or 4.2?**

In general, I am guided by deviance reduction statistics when deciding whether to include variables that may, or may not be true confounders, but that are not intrinsically of interest.

If I am interested in the joint effects of 2 or more variables, I usually include the interaction term unless the inclusion of the interaction parameter has almost no effect on the resulting relative risk estimates.

There are no hard and fast guidelines to model building other than that it is best not to include uninteresting variables in the model that have a trivial effect on the model deviance.

I think I personally would go with Table 4.2 over 4.1 in spite of the lack of evidence of interaction. The odds ratio for both ≥120 gm alcohol and ≥30 gm tobacco is so large that I would worry that we were being misled by not taking into account a small but real interaction term.

It would also be acceptable to say that we analyzed the data both ways, found no evidence of interaction, got comparable results and were presenting the multiplicative model results only.

Table 4.1. Effect of Alcohol and Tobacco on Esophageal Cancer Risk

Multiplicative Model -- Adjusted to Age

Daily Alcohol Comsumption	Daily Tobacco Consumption					
	0-9 gm		10-29 gm		30gm	
	Odds Ratio	95% CI	Odds Ratio	95% CI	Odds Ratio	95% CI
0-39 gm	1.0*		1.59	(1.1 - 2.4)	5.16	(2.6 - 10)
40-79 gm	4.21	(2.6 - 6.9)	6.71	(3.6 - 12)	21.7	(9.2 - 51)
80-119 gm	7.22	(4.1 - 13)	11.5	(5.9 - 22)	37.3	(15 - 91)
120 gm.	36.8	(17 - 78)	58.6	(25 - 140)	190	(67 - 540)

* Denominator of odds ratios

Table 4.2. Effect of Alcohol and Tobacco on Esophageal Cancer Risk

Model with all 2-Way Interaction Terms -- Adjusted for Age

Daily Alcohol Comsumption	Daily Tobacco Consumption					
	0 – 9 gm		10 – 29 gm		≥ 30 gm	
	Odds Ratio	95% Confidence Interval	Odds Ratio	95% Confidence Interval	Odds Ratio	95% Confidence Interval
0 – 39 gm	1.0*		3.8	(1.6 – 9.2)	8.65	(2.4 – 31)
40 – 79 gm	7.55	(3.4 – 17)	9.34	(4.2 – 21)	32.9	(10 – 110)
80 – 119 gm	12.7	(5.2 – 31)	16.1	(6.8 – 38)	72.3	(15 – 350)
≥ 120 gm	65.1	(20 – 210)	92.3	(29 – 290)	196	(30 – 1300)

* Denominator of odds ratios

16. Influence Analysis for Logistic Regression

Consider a logistic regression model with

J distinct covariate patterns

d_j events occur among n_j patients with the covariate pattern $x_{j1}, x_{j2}, \ldots x_{jq}$.

Let $\pi_j = \pi\left[x_{j1}, x_{j2}, \ldots, x_{jq}\right]$ denote the probability that a patient with the j^{th} pattern of covariate values suffers an event.

Then d_j has a binomial distribution with

expected value $n_j\pi_j$

standard error $\sqrt{n_j\pi_j\left(1-\pi_j\right)}$

Hence

$\left(d_j - n_j\pi_j\right) / \sqrt{n_j\pi_j\left(1-\pi_j\right)}$

will have a mean of 0 and a standard error of 1.

Let $\hat{\pi}_j = \dfrac{\exp\left[\hat{\alpha} + \hat{\beta}_1 x_{j_1} + \hat{\beta}_2 x_{j2} + \ldots + \hat{\beta}_q x_{jq}\right]}{1 + \exp\left[\hat{\alpha} + \hat{\beta}_1 x_{j_1} + \hat{\beta}_2 x_{j2} + \ldots + \hat{\beta}_q x_{jq}\right]}$

be the estimate of π_j obtained by substituting the maximum likelihood parameter estimates into the logistic probability function.

Then the **residual** for the j^{th} covariate pattern is $d_j - n_j\hat{\pi}_j$

The **Pearson residual** is $r_{j(Pearson)} = \left(d_j - n_j\hat{\pi}_j\right) / \sqrt{n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}$

which should have a mean of 0 and a standard deviation of 1 if the model is correct and if $\sqrt{n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)}$ is a good estimate of the standard error of $d_j - n_j\hat{\pi}_j$.

The **leverage** h_j is analogous to leverage in linear regression.

It measures to potential of a covariate pattern to influence our parameter estimates if the associated residual is large.

For our purposes we can define h_j by the formula

$$\begin{aligned}\text{var}\left[d_j - n_j\hat{\pi}_j\right] &= n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)\left(1-h_j\right) \\ &\cong \text{var}\left[d_j - n_j\pi_j\right]\left(1-h_j\right)\end{aligned}$$

In other words, 100(1-h_j) is the percent reduction in the variance of the j^{th} residual due to the fact that the estimate of $n_j\hat{\pi}_j$ is pulled towards d_j.

The value of h_j lies between 0 and 1.

When h_j is very small d_j has almost no effect on its estimated expected value $n_j\hat{\pi}_j$.

When h_j is close to 1, then $d_j \cong n_j\hat{\pi}_j$. This implies that both the residual $d_j - n_j\hat{\pi}_j$ and its variance will be close to zero.

The **standardized Pearson residual** for the j^{th} covariate pattern is the residual divided by its standard error. That is,

$$r_{sj} = \frac{d_j - n_j\hat{\pi}_j}{\sqrt{n_j\hat{\pi}_j\left(1-\hat{\pi}_j\right)\left(1-h_j\right)}} = \frac{r_{j(Pearson)}}{\sqrt{1-h_j}}$$

This residual is analogous to the studentized residual for linear regression.

r_{sj} has mean 0 and standard error 1

is not necessarily normally distributed when n_j is small.

The square of the standardized Pearson residual is denoted

$$\Delta X_j^2 = r_{sj}^2 = r_{j(Pearson)}^2 / \left(1-h_j\right)$$

We will use the critical value $\left(z_{0.025}\right)^2 = 1.96^2 = 3.84$ as a very rough guide to identifying large values of ΔX_j^2.

Approximately 95% of these squared residuals should be <u>less than</u> 3.84 if the logistic regression model is correct.

The $\Delta\hat{\beta}_j$ **influence statistic** is a measure of the influence of the jth covariate pattern on all of the parameter estimates taken together. It equals $\Delta\hat{\beta}_j = r_{sj}^2 h_j / \left(1-h_j\right)$

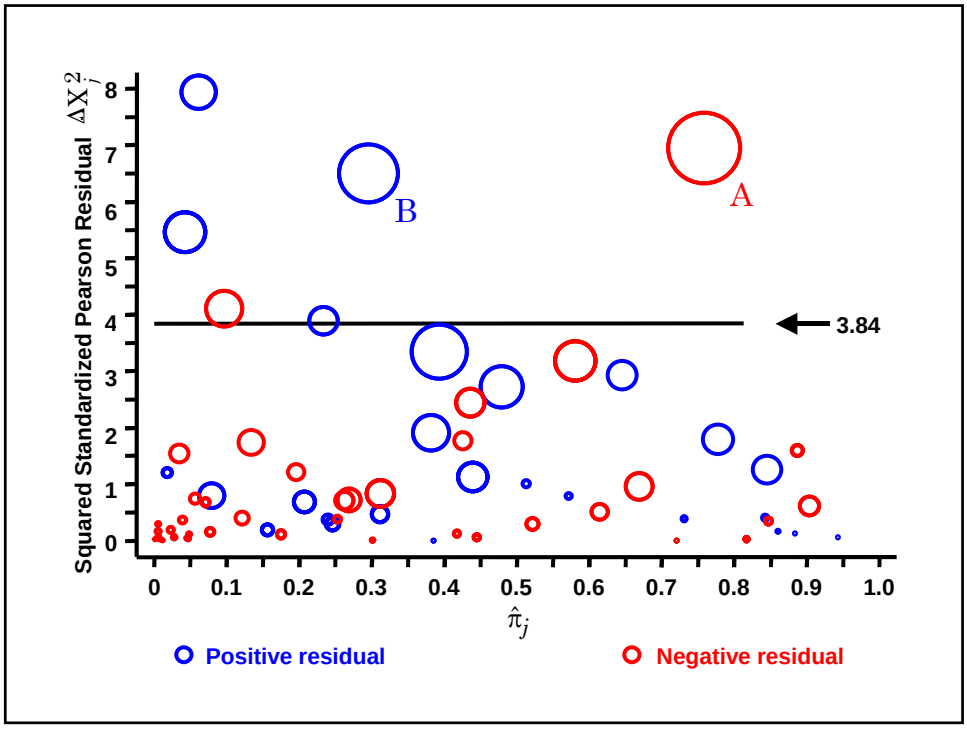
Note that $\Delta\hat{\beta}_j$ increases with both the magnitude of the standardized residual and the size of the leverage.

It is analogous to Cook's distance for linear regression.

Covariate patterns associated with large values of ΔX_j^2 and $\Delta\hat{\beta}_j$ merit special attention.

The following plot is for our model of alcohol and tobacco dose with interaction terms and plots ΔX_j^2 against $\hat{\pi}_j$

The area of the circles is proportional to $\Delta\hat{\beta}_j$



There are 68 unique covariate patterns in this data set.

5% of 68 equals 3.4

There are 6 residuals greater than 3.84.

There are 2 large squared residuals with high influence.

Residual A is associated with patients who are age 55 – 64 and consume, on a daily basis, at least 120 gm of alcohol 0 – 9 gm of tobacco.

Residual B is associated with patients who are age 55 – 64 and consume, on a daily basis, 0 – 39 gm of alcohol and at least 30 gm of tobacco.

The $\Delta\beta_j$ influence statistics associated with residuals A and B are 6.16 and 4.15, respectively.

NOTE:

In linear regression observations with high influence are due to a single patient and we have the option of deleting the patient

In logistic regression covariate patters with high influence indicate poor model fit. However, we usually do not have the option of deleting the pattern if it represents a sizable number of patients.

Daily Drug Consumption		Complete Data		Deleted Covariate Pattern			
				A†		B‡	
Tobacco	Alcohol	Odds Ratio	95% Confidence Interval	Odds Ratio	Percent Change from Complete Data	Odds Ratio	Percent Change from Complete Data
0 – 9 gm	0 – 39 gm	1.0*		1.0*		1.0*	
0 – 9 gm	40 – 79 gm	7.55	(3.4 – 17)	7.53	-0.26%	7.70	2.0%
0 – 9 gm	80 – 119 gm	12.7	(5.2 – 31)	12.6	-0.79%	13.0	2.4%
0 – 9 gm	≥ 120 gm.	65.1	(20 – 210)	274	321%	66.8	2.6%
10 – 29 gm	0 – 39 gm	3.80	(1.6 – 9.2)	3.77	-0.79%	3.86	1.6%
10 – 29 gm	40 – 79 gm	9.34	(4.2 – 21)	9.30	-0.43%	9.53	2.0%
10 – 29 gm	80 – 119 gm	16.1	(6.8 – 38)	16.0	-0.62%	16.6	3.1%
10 – 29 gm	≥ 120 gm.	92.3	(29 – 290)	95.4	3.4%	94.0	1.8%
≥ 30gm	0 – 39 gm	8.65	(2.4 – 31)	8.66	0.12%	1.88	-78%
≥ 30gm	40 – 79 gm	32.9	(10 – 110)	33.7	2.4%	33.5	1.8%
≥ 30gm	80 – 119 gm	72.3	(15 – 350)	73.0	0.97%	74.2	2.6%
≥ 30gm	≥ 120 gm.	196	(30 – 1300)	198	1.02%	203	3.6%

* Denominator of odds ratios

† Patients age 55 – 64 who drink at least 120 gm a day and smoke 0 – 9 gm a day deleted

‡ Patients age 55 – 64 who drink 0 – 39 gm a day and smoke at least 30 gm a day deleted

Table 4.1. Effect of Alcohol and Tobacco on Esophageal Cancer Risk

Daily Alcohol Comsumption	Daily Tobacco Consumption					
	0-9 gm		10-29 gm		30gm	
	Odds Ratio	95% CI	Odds Ratio	95% CI	Odds Ratio	95% CI
Multiplicative Model -- Adjusted to Age						
0-39 gm	1.0*		1.59	(1.1 - 2.4)	5.16	(2.6 - 10)
40-79 gm	4.21	(2.6 - 6.9)	6.71	(3.6 - 12)	21.7	(9.2 - 51)
80-119 gm	7.22	(4.1 - 13)	11.5	(5.9 - 22)	37.3	(15 - 91)
120 gm.	36.8	(17 - 78)	58.6	(25 - 140)	190	(67 - 540)
Model with all 2-Way Interaction Terms -- Adjusted for Age						
0 – 39 gm	1.0*		3.8	(1.6 – 9.2)	8.65	(2.4 – 31)
40 – 79 gm	7.55	(3.4 – 17)	9.34	(4.2 – 21)	32.9	(10 – 110)
80 – 119 gm	12.7	(5.2 – 31)	16.1	(6.8 – 38)	72.3	(15 – 350)
≥120 gm	65.1	(20 – 210)	92.3	(29 – 290)	196	(30 – 1300)

* Denominator of odds ratios

17. What is the best model?

We have 975 patients,
200 cases,
68 unique covariate patterns
17 parameters in the interactive model.

Over-fitting is certainly a concern

Still the effect of dose of tobacco and alcohol on risk is very marked, which makes the interactive model tempting to use.

It is a pity that age, alcohol and tobacco were categorized before we received this data. It is always a mistake to throw such data away.

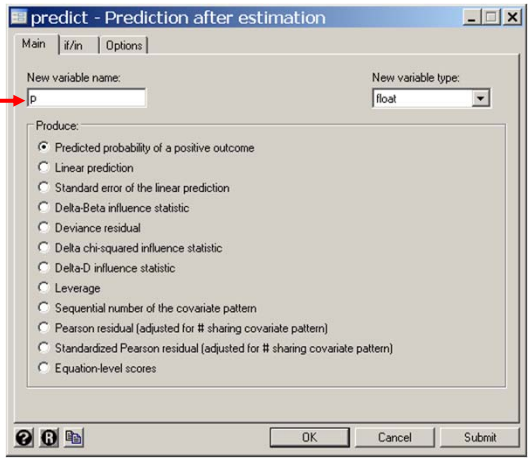
If we had the continuous data we could fit a cubic spline model with 1 constant parameter
6 spline parameters: 2 each for age alcohol and tobacco
4 interaction parameters for a total of
11 parameters, which would be more reasonable.

18. Residual analysis with Stata

5.20.EsophagelaCa.ClassVersion.log continues as follows

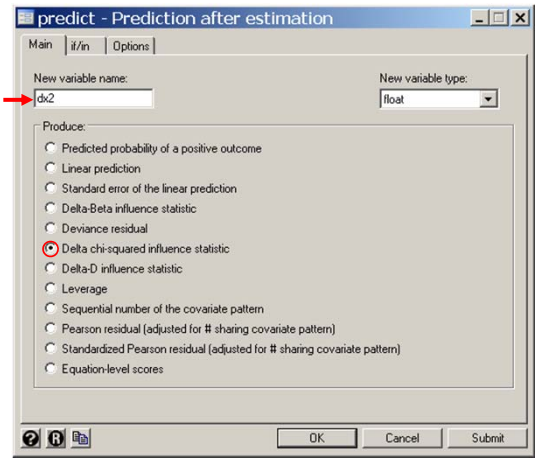
```
. *
. *  Perform residual analysis
. *
. * Statistics > Postestimation > Predictions, residuals, etc.
. predict p, p                                                    {1}
```

{1} The ***p*** option in this ***predict*** command defines the variable ***p*** to equal $\hat{\pi}_j$. In this and the next two ***predict*** commands the name of the newly defined variable is the same as the command option.



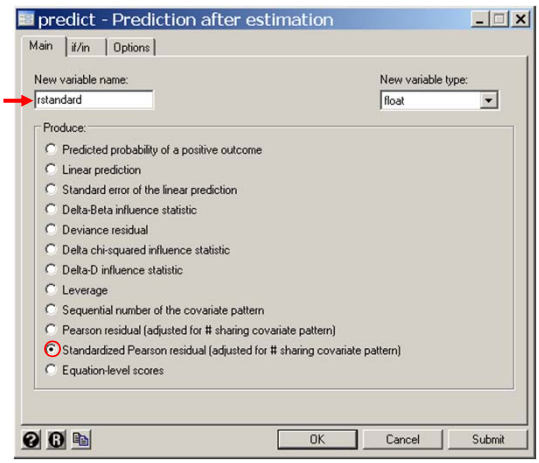

```
. predict dx2, dx2                                                {2}
(57 missing values generated)
```

{2} Define the variable ***dx2*** to equal ΔX_j^2. All records with the **same** covariate **pattern** are given the **same value** of *dx2*.



```
. predict rstandard, rstandard                                          {3}
(57 missing values generated)
```

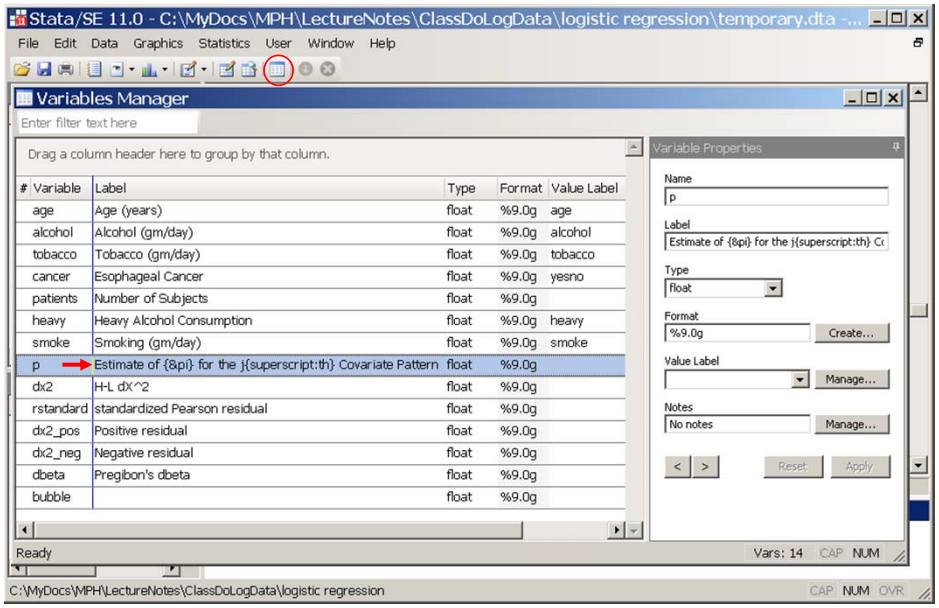
{3} Define ***rstandard*** to equal the **standardized Pearson residual** r_{sj}.

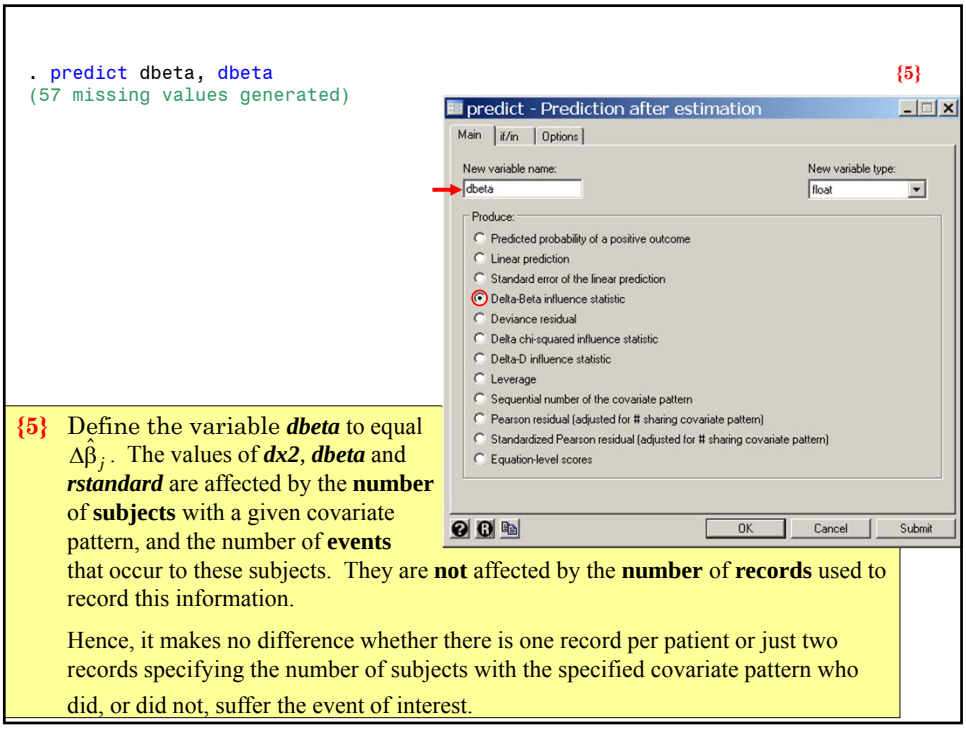



```
. generate dx2_pos = dx2 if rstandard >= 0                              {4}
(137 missing values generated)

. generate dx2_neg = dx2 if rstandard <  0
(112 missing values generated)

. label variable dx2_pos "Positive residual"

. label variable dx2_neg "Negative residual"

. label variable p                          ///
      "Estimate of {&pi} for the j{superscript:th} Covariate Pattern"   {5}
```

{4} We are going to draw a scatterplot of ΔX_j^2 against $\hat{\pi}_j$. We would like to **color code** the plotting symbols to indicate if the residual is **positive** or **negative**. This command defines ***dx2_pos*** to equal ΔX_j^2 if and only if r_{sj} is non-negative. The next command defines ***dx2_neg*** to equal ΔX_j^2 if r_{sj} is negative.

{5} Greek lettters, superscripts, italics, etc can be entered in variable labels. {&pi} enters the letter π into the label. {superscript:th} writes the letters "th" as a superscript.





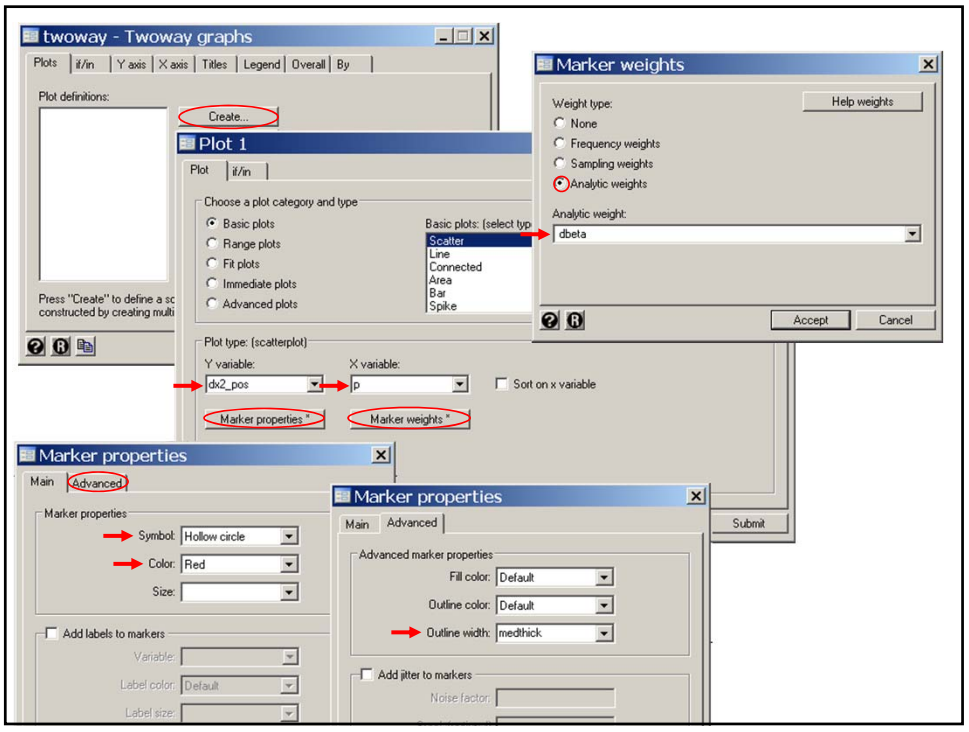
{5} Define the variable ***dbeta*** to equal $\Delta\hat{\beta}_j$. The values of ***dx2***, ***dbeta*** and ***rstandard*** are affected by the **number** of **subjects** with a given covariate pattern, and the number of **events** that occur to these subjects. They are **not** affected by the **number** of **records** used to record this information.

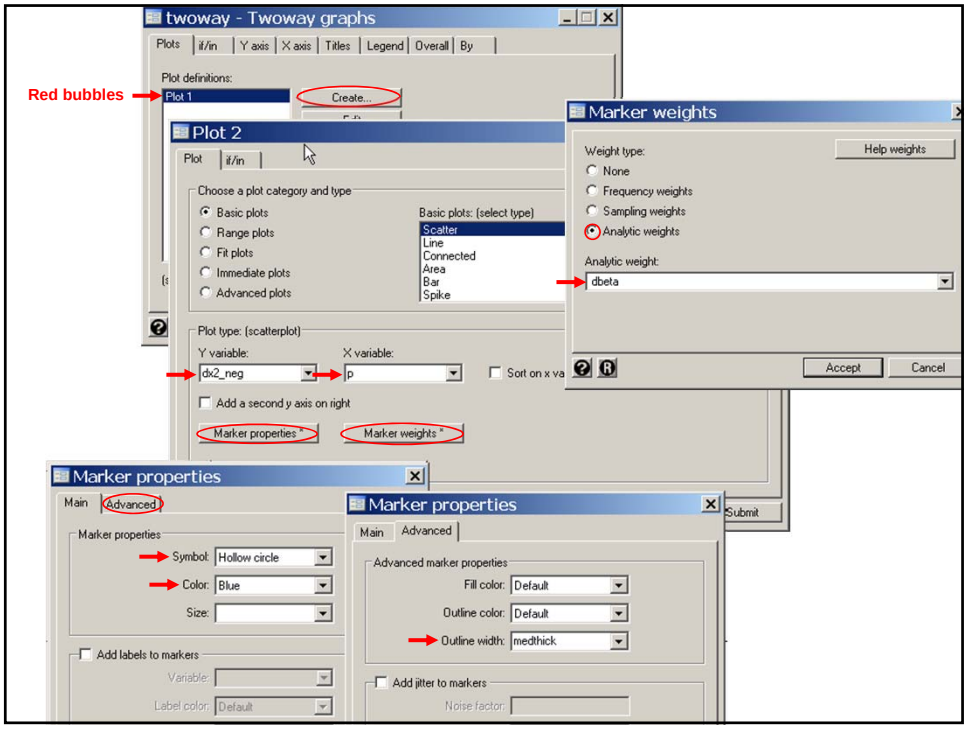
Hence, it makes no difference whether there is one record per patient or just two records specifying the number of subjects with the specified covariate pattern who did, or did not, suffer the event of interest.

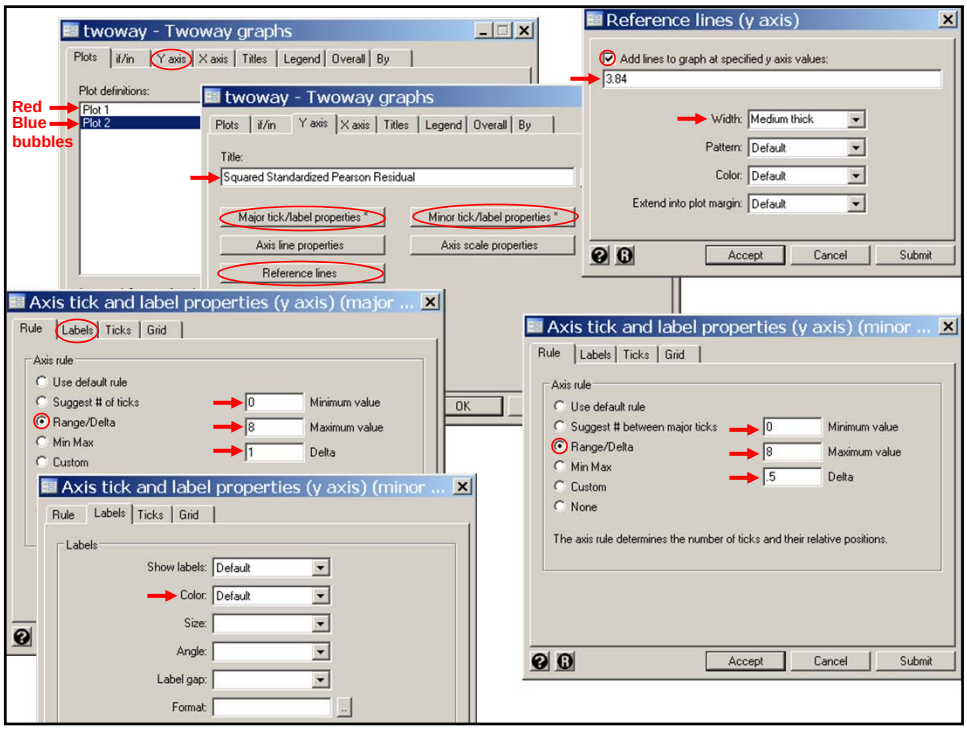
```
. scatter        dx2_pos p [weight=dbeta]                                ///    {6}
>       , msymbol(Oh) mlwidth(medthick) mcolor(red)                       ///    {7}
>     || scatter dx2_neg p [weight=dbeta]                                 ///
>       , msymbol(Oh) mlwidth(medthick) mcolor(blue)                      ///
>     ||, ylabel(0(1)8, angle(0))                                         ///
>         ymtick(0(.5)8) yline(3.84, lwidth(medthick))                    ///
>         xlabel(0(.1)1) xmtick(0(.05)1)                                   ///
>         ytitle("Squared Standardized Pearson Residual") xscale(titlegap(2))
(analytic weights assumed)
(analytic weights assumed)
```

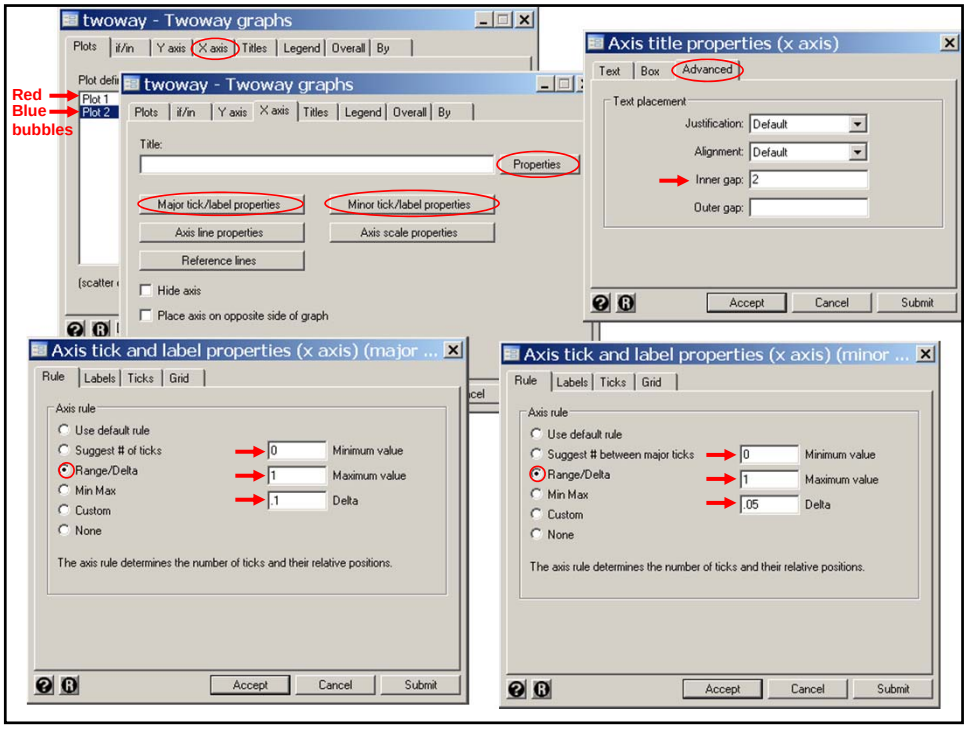
{6} This graph produces a scatterplot of ΔX_j^2 against $\hat{\pi}_j$ that is shown in the next slide. The ***[weight=dbeta]*** command modifier causes the plotting symbols to be circles whose **area** is **proportional** to the variable ***dbeta***. We plot both **dx2_pos** and **dx2_neg** against **p** in order to be able to assign **different colors** to values of ΔX_j^2 that are associated with positive or negative residuals.

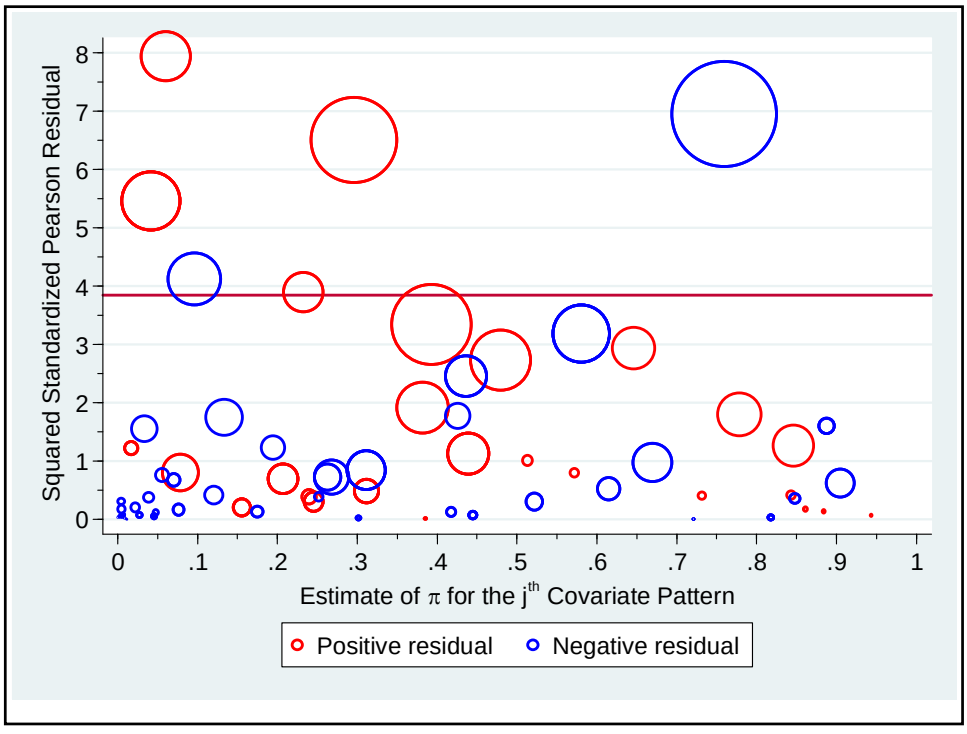
{7} **mlwidth** defines the width of the marker lines. This is, the width of the circles. **mcolor** defines the marker color.




twoway - Twoway graphs
Red bubbles
Plot 1
Create...
Plot 2
Marker weights
Weight type:
None
Frequency weights
Sampling weights
Analytic weights
Analytic weight:
dbeta
Help weights
Accept
Cancel
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Area
Bar
Spike
Plot type: (scatterplot)
Y variable:
dx2_neg
X variable:
p
Add a second y axis on right
Marker properties *
Marker weights *
Marker properties
Main
Advanced
Marker properties
Symbol:
Hollow circle
Color:
Blue
Size:
Add labels to markers
Variable:
Label color:
Default
Marker properties
Main
Advanced
Advanced marker properties
Fill color:
Default
Outline color:
Default
Outline width:
medthick
Add jitter to markers
Noise factor:
Submit


twoway - Twoway graphs
Red
Blue
bubbles
Plot 1
Plot 2
Reference lines (y axis)
Add lines to graph at specified y axis values:
3.84
Width:
Medium thick
Pattern:
Default
Color:
Default
Extend into plot margin:
Default
Accept
Cancel
Submit
Title:
Squared Standardized Pearson Residual
Major tick/label properties *
Minor tick/label properties *
Axis line properties
Axis scale properties
Reference lines
Axis tick and label properties (y axis) (major ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
Minimum value
Maximum value
Delta
OK
Axis tick and label properties (y axis) (minor ...
Suggest # between major ticks
None
The axis rule determines the number of ticks and their relative positions.
Show labels:
Default
Color:
Size:
Angle:
Label gap:
Format:


twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Red
Blue
bubbles
Plot 1
Plot 2
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines
Hide axis
Place axis on opposite side of graph
Axis title properties (x axis)
Text
Box
Advanced
Text placement
Justification:
Alignment:
Inner gap:
Outer gap:
Accept
Cancel
Submit
Axis tick and label properties (x axis) (major ...
Rule
Labels
Ticks
Grid
Axis rule
Use default rule
Suggest # of ticks
Range/Delta
Min Max
Custom
None
Minimum value
Maximum value
Delta
The axis rule determines the number of ticks and their relative positions.
Axis tick and label properties (x axis) (minor ...
Suggest # between major ticks


Squared Standardized Pearson Residual
0
1
2
3
4
5
6
7
8
0
.1
.2
.3
.4
.5
.6
.7
.8
.9
1
Estimate of π for the j^th Covariate Pattern
Positive residual
Negative residual

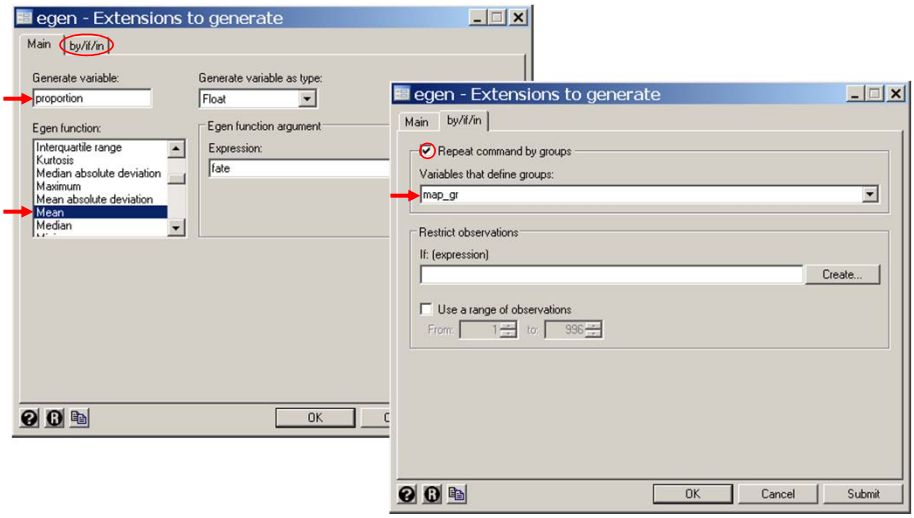
19. Restricted Cubic Splines and Logistic Regression

In the following example we use restricted cubic splines to model the effect of baseline MAP on hospital mortality in the SUPPORT data set.

```
. * SUPPORTlogisticRCS.log
. *
. *  Regress mortal status at discharge against MAP
. *  in the SUPPORT data set (Knaus et al. 1995).
. *
. use "C:\WDDtext\3.25.2.SUPPORT.dta" , replace

. *
. *  Calculate the proportion of patients who die in hospital
. *   stratified by MAP.
. *
. generate map_gr = round(map,5)                                    {1}
. sort map_gr
. label variable map_gr "Mean Arterial Pressure (mm Hg)"
. * Data > Create or change data > Create new variable (extended)
. by map_gr: egen proportion = mean(fate)                           {2}
```

{1} *round(map, 5)* rounds *map* to the nearest integer divisible by 5.

{2} This command defines *proportion* to equal the average value of *fate* over all records with the same value of *map_gr*. Since *fate* is a zero-one indicator variable, *proportion* will be equal to the proportion of patients with the same value of *map_gr* who die (have *fate* = 1). This command requires that the data set be sorted by the *by* variable (*map_gr*).

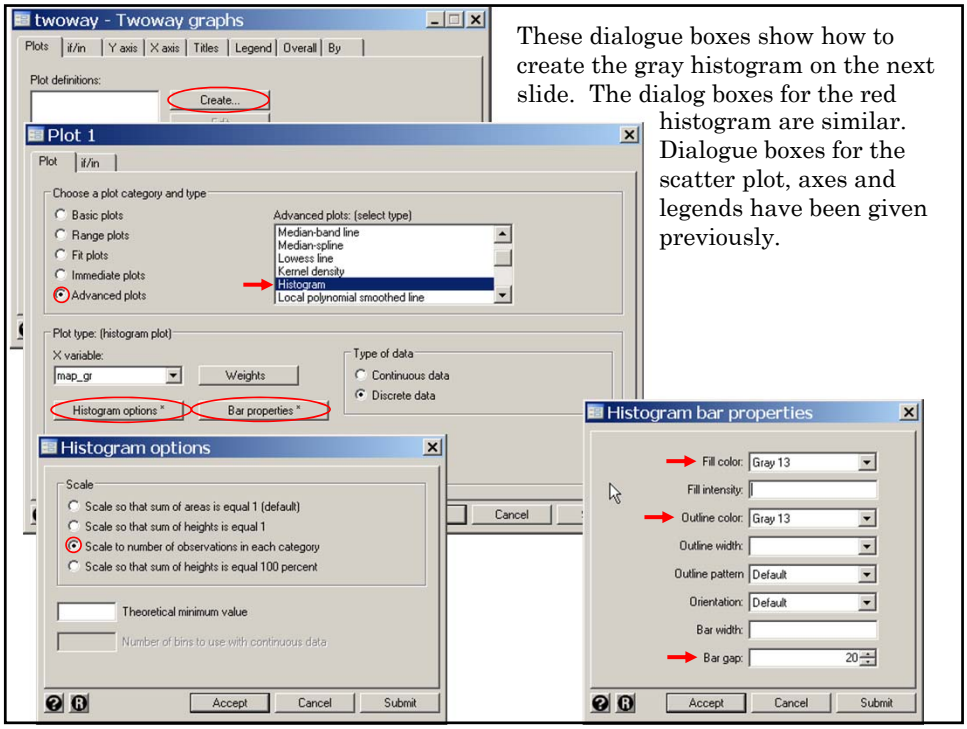



```
. generate rate = 100*proportion
. label variable rate "Observed In-Hospital Mortality Rate (%)"
. generate deaths = map_gr if fate
(747 missing values generated)
```

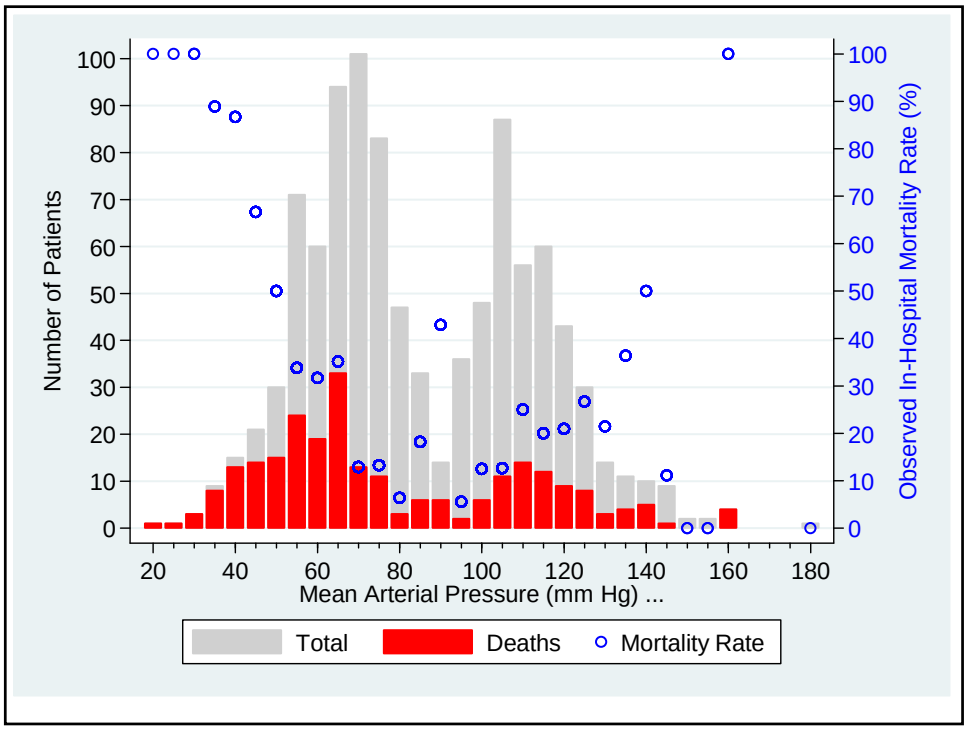
```
. *
. *  Draw an exploratory graph showing the number of patients,
. *  the number of deaths and the mortality rate for each MAP.
. *
. twoway histogram map_gr, discrete frequency color(gs13) gap(20) /// {3}
>     || histogram deaths, discrete frequency color(red) gap(20)  /// {4}
>     || scatter rate map_gr, yaxis(2) symbol(Oh) color(blue)     ///
>     ,  xlabel(20 (20) 180) ylabel(0(10)100, angle(0))           ///
>        xmtick(25 (5) 175) ytitle(Number of Patients)            ///
>        ylabel(0 (10) 100, angle(0) labcolor(blue) axis(2))      ///
>        ytitle(,color(blue) axis(2))                             ///
>        legend(order(1 "Total" 2 "Deaths" 3 "Mortality Rate" )   ///
>        rows(1))
```

{3} The command *twoway histogram map_gr* produces a histogram of the variable *map_gr*. The *discrete* option specifies that a bar is to be drawn for each distinct value of *map_gr*; *frequency* specifies that the *y*-axis will be the number of patients at each value of *map_gr*; *color(gs13)* specifies that the bars are to be light gray and *gap(20)* reduces the bar width by 20\% to provide separation between adjacent bars.

{4} This line of this command overlays a histogram of the number of in-hospital deaths on the histogram of the total number of patients.



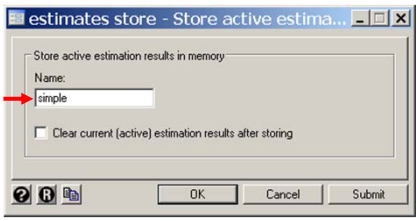
These dialogue boxes show how to create the gray histogram on the next slide. The dialog boxes for the red histogram are similar. Dialogue boxes for the scatter plot, axes and legends have been given previously.



```
. *
. *  Regress in-hospital mortality against MAP using simple
. *  logistic regression.
. *
. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic fate map                                                    {5}

Logistic regression                               Number of obs   =        996
                                                  LR chi2(1)      =      29.66
                                                  Prob > chi2     =     0.0000
Log likelihood = -545.25721                       Pseudo R2       =     0.0265

------------------------------------------------------------------------------
        fate | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         map |   .9845924   .0028997    -5.27   0.000     .9789254    .9902922
------------------------------------------------------------------------------

. *  Statistics > Postestimation > Manage estimation results > Store in memory
. estimates store simple                                               {6}
```

{5} This command regresses *fate* against *map* using simple logistic regression.

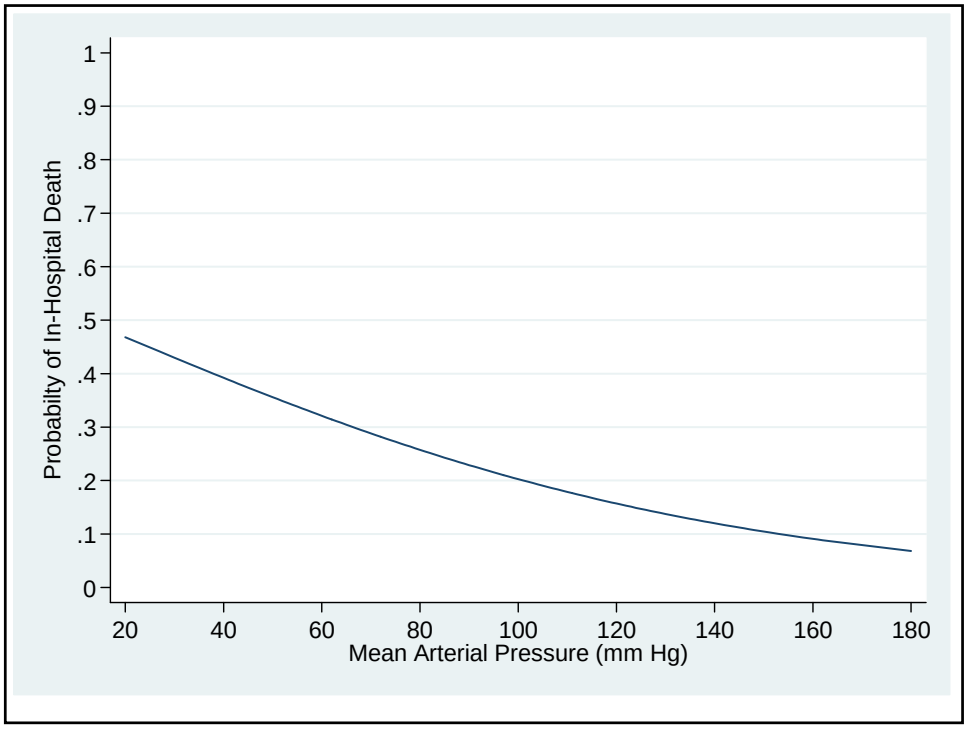
{6} This command stores parameter estimates and other statistics from the most recent regression command. These statistics are stored under the name *simple*. We will use this information later to calculate the change in model deviance.




```
. predict p,p                                                          {7}

. label variable p "Probabilty of In-Hospital Death"

. line p map, ylabel(0(.1)1, angle(0)) xlabel(20(20)180)
```

{7} The p option of this predict command defines p equal to the predicted probability of in-hospital death under the model. That is

$$p = \exp[\alpha + \beta \times map_i] / (1 + \exp[\alpha + \beta \times map_i]) = \text{logit}^{-1}[\alpha + \beta \times map_i]$$



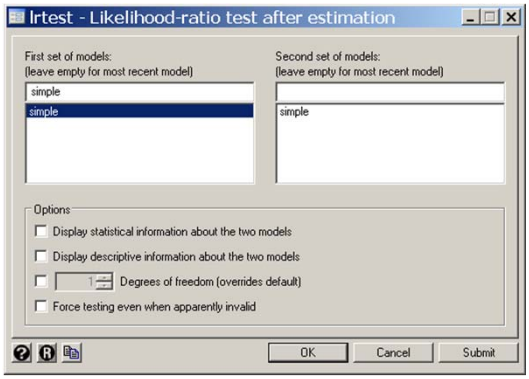

```
. *  Variables Manager
. drop p

. *
. *  Repeat the preceding model using restricted cubic splines
. *  with 5 knots at their default locations.
. *
. * Data > Create... > Other variable-creation... > linear and cubic...
. mkspline _Smap = map, cubic displayknots

             |     knot1      knot2      knot3      knot4      knot5
-------------+------------------------------------------------------
         map |        47         66         78        106        129

. * Statistics > Binary outcomes > Logistic regression (reporting odds ratios)
. logistic fate _S*                                                    {8}

Logistic regression                               Number of obs   =        996
                                                  LR chi2(4)      =     122.86 {9}
                                                  Prob > chi2     =     0.0000
Log likelihood = -498.65571                       Pseudo R2       =     0.1097

------------------------------------------------------------------------------
        fate | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      _Smap1 |   .8998261   .0182859    -5.19   0.000     .8646907    .9363892
      _Smap2 |    1.17328   .2013998     0.93   0.352      .838086    1.642537
      _Smap3 |     1.0781   .7263371     0.11   0.911     .2878645    4.037664
      _Smap4 |   .6236851   .4083056    -0.72   0.471     .1728672    2.250185
------------------------------------------------------------------------------
```

{8} Regress *fate* against MAP using a 5-knot RCS logistic regression model.

{9} Testing the null hypothesis that mortality is unrelated to MAP under this model is equivalent to testing the null hypothesis that all of the parameters associated with the spline covariates are zero. The likelihood ratio χ^2 statistic to test this hypothesis equals 122.86. It has four degrees of freedom and is highly significant $P < 0.00005$.

```
. *
. *  Test null hypotheses that the logit of the probability of
. *  in-hospital death is a linear function of MAP.
. *
. * Statistics > Postestimation > Tests > Likelihood-ratio test
. lrtest simple .                                                    {10}

Likelihood-ratio test                             LR chi2(3)  =     93.20
(Assumption: simple nested in .)                  Prob > chi2 =    0.0000
```

{10} This *lrtest* command calculates the likelihood ratio test of the null hypothesis that there is a linear relationship between the log odds of in-hospital death and baseline MAP. This is equivalent to testing the null hypothesis that *_Smap2* = *_Smap3* = *_Smap4* = 0. The *lrtest* command calculates the change in model deviance between two nested models. In this command, *simple* is the name of the model output saved by the previous *estimates store* command (see Comment 6). The period (.) refers to the estimates from the most recently executed regression command. The user must insure that the two models specified by this command are nested. The change in model deviance equals 93.2. Under the null hypothesis that the simple logistic regression model is correct this statistic will have an approximately chi-squared distribution with three degrees of freedom. The *P* value associated with this statistic is (much) less than 0.00005.




```
. display 2*(545.25721 -498.65571)                                    {11}
93.203
```

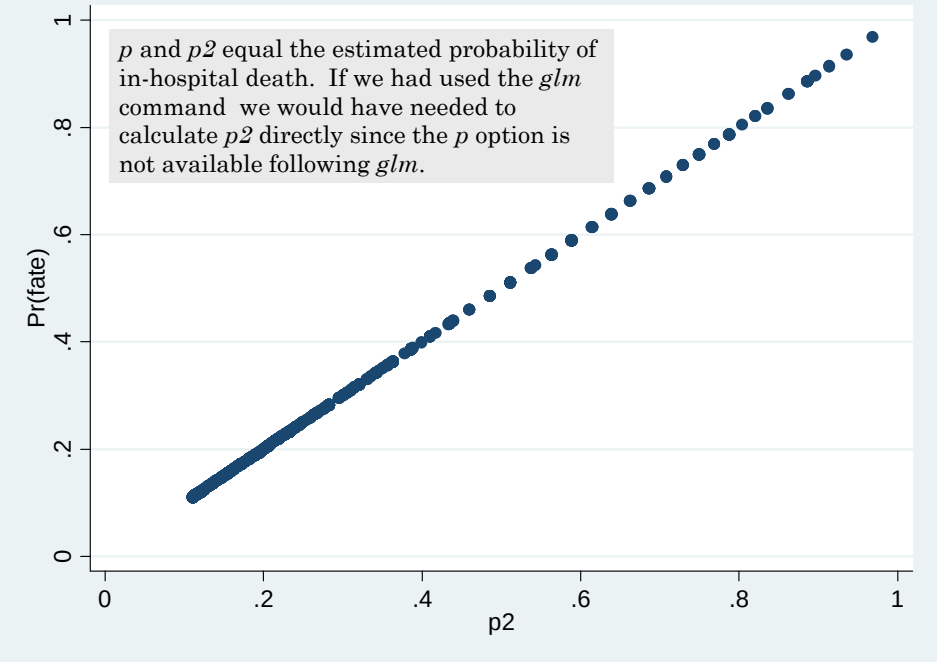
{11} Here we calculate the change in model deviance by hand from the maximum values of the log likelihood functions of the two models under consideration. Note that this gives the same answer as the preceding *lrtest* command.

N.B. We can always test the validity of a simple logistic regression model by running a RCS model with k knots and then testing the null hypothesis of whether the second through $k\text{-}1^{th}$ spline covariate parameters are simultaneously zero. In other words, we test the null hypothesis that the simple logisitic regression model is valid by testing the null hypothesis that the second through $k\text{-}1^{th}$ spline covariate parameters are simultaneously zero.

If we run a three-knot model then testing whether the second spline covariate parameter is zero is equivalent to testing the validity of the simple logistic regression model.

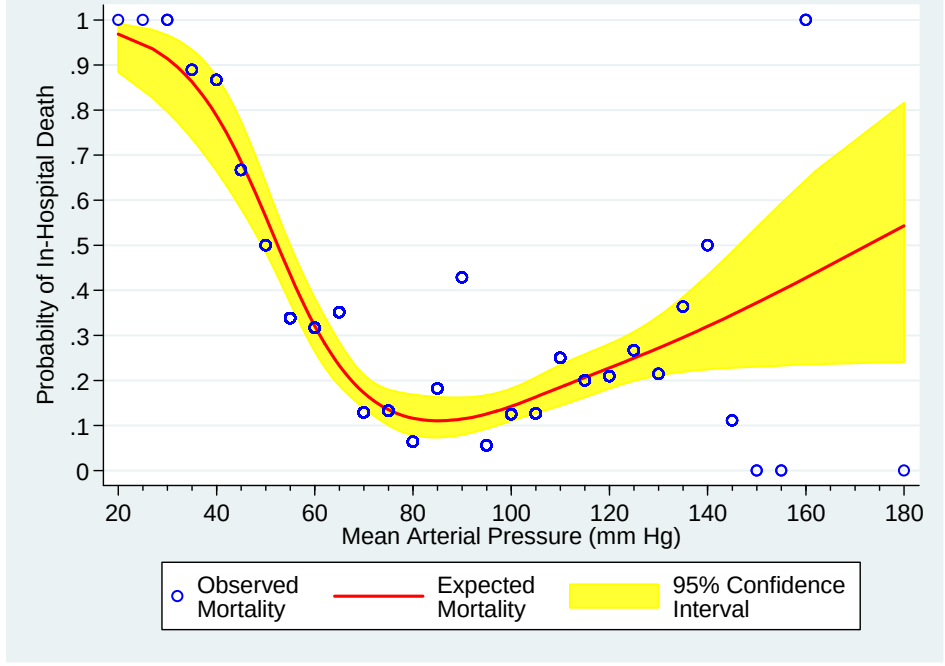
```
. *
. *  Plot the estimated probability of death against MAP together
. *  with the 95% confidence interval for this curve.  Overlay
. *  the MAP-specific observed mortality rates.
. *
. predict p,p                                                          {12}
. predict logodds, xb
. predict stderr, stdp
. generate p2 = exp(logodds)/(1+exp(logodds))
. *
. *  The values of p and p2 are identical.
. *
. scatter p p2
```

{12} The variable *p* is the estimated probability of in-hospital death from model our 5-knot RCS model.



```
. generate lodds_lb = logodds - 1.96*stderr
. generate lodds_ub = logodds + 1.96*stderr
. generate ub_p = exp(lodds_ub)/(1+exp(lodds_ub))                              {13}
. generate lb_p = exp(lodds_lb)/(1+exp(lodds_lb))
. twoway rarea lb_p ub_p map, color(yellow)                                    ///
>     || line p map, lwidth(medthick) color(red)                               ///
>     || scatter proportion map_gr,  symbol(Oh) color(blue)                    ///
>     ,  ylabel(0(.1)1, angle(0)) xlabel(20 (20) 180)                          ///
>        xmtick(25(5)175) ytitle(Probabilty of In-Hospital Death)              ///
>        legend(order(3 "Observed" "Mortality" 2 "Expected" "Mortality" ///
>            1 "95% Confidence" "Interval") rows(1))
```

{13} The variables *lb_p* and *ub_p* are the lower and upper 95% confidence bounds for *p*, respectively.



```
. *
. *  Determine the spline covariates at MAP = 90
. *
. list _S* if map == 90                                                {14}

     +---------------------------------------+
     | _Smap1     _Smap2     _Smap3     _Smap4 |
     |---------------------------------------|
575. |     90   11.82436   2.055919   .2569899 |
                                                          {output omitted}
581. |     90   11.82436   2.055919   .2569899 |
     +---------------------------------------+

. *
. *  Let or1 = _Smap1 minus the value of _Smap1 at 90.
. *  Define or2, or3 and or3 in a similar fashion.
. *
. generate or1 = _Smap1 - 90

. generate or2 = _Smap2 - 11.82436

. generate or3 = _Smap3 -  2.055919

. generate or4 = _Smap4 -   .2569899
```

{14} List the values of the spline covariates for the seven patients in the data set with a baseline MAP of 90. Only one or these identical lines of output are shown here.

N.B. $\text{logodds}[map] = \alpha + \beta_1 map + \beta_2 \_ Smap2(map) + \cdots + \beta_4 \_ Smap4(map)$

$\text{logodds}[90] = \alpha + \beta_1 \times 90 + \beta_2 \_ Smap2(90) + \cdots + \beta_4 \_ Smap4(90)$

$\text{logodds}[map] - \text{logodds}[90] = \beta_1 or1 + \beta_2 or2 + \beta_3 or3 + \beta_4 or4$

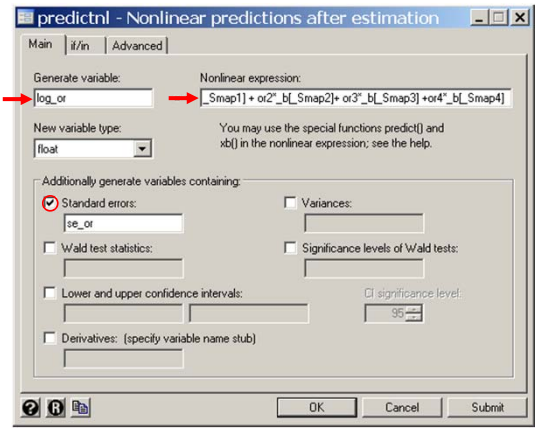
exp[logodds[*map*] - logodds[90]] = odds ratio of a patient with MAP = *map* compared to a patient with a MAP = 90 by the usual argument.

```
. *
. *  Calculate the log odds ratio for in-hospital death
. *  relative to patients with MAP = 90.
. *
. * Statistics > Postestimation > Nonlinear predictions
. predictnl log_or = or1*_b[_Smap1] + or2*_b[_Smap2]       ///        {15}
>     + or3*_b[_Smap3] +or4*_b[_Smap4], se(se_or)                     {16}
```

{15} Define *log_or* to be the mortal log odds ratio for the i^{th} patient in comparison to patients with a MAP of 90. The parameter estimates from the most recent regression command may be used in *generate* commands and are named *_b*[*varname*]. For example, in this RCS model *_b*[*_Smap2*] = $\hat{\beta}_2$ = 1.17328; *or2* = *_Smap2* - 11.82436.

The command *predictnl* may be used to estimate non-linear functions of the parameter estimates. It is also very useful for calculating linear combinations of these estimates as is illustrated here.

{16} The option *se(se_or)* calculates a new variable called *se_or* which equals the standard error of the log odds ratio.



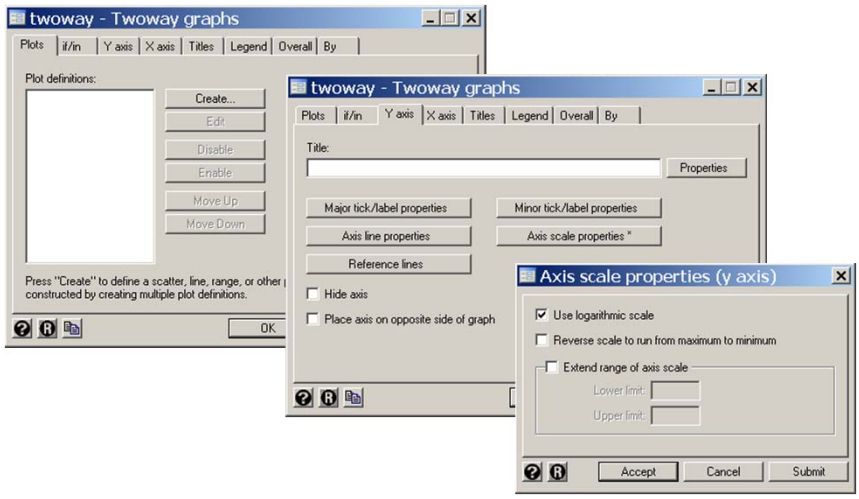
```
. generate lb_log_or = log_or - 1.96*se_or
. generate ub_log_or = log_or + 1.96*se_or
. generate or = exp(log_or)                                          {17}
. generate lb_or = exp(lb_log_or)                                    {18}
. generate ub_or = exp(ub_log_or)
```

{17} The variable *or* equals the odds ratio for in-hospital death for each patient relative to that for a patient with MAP = 90.

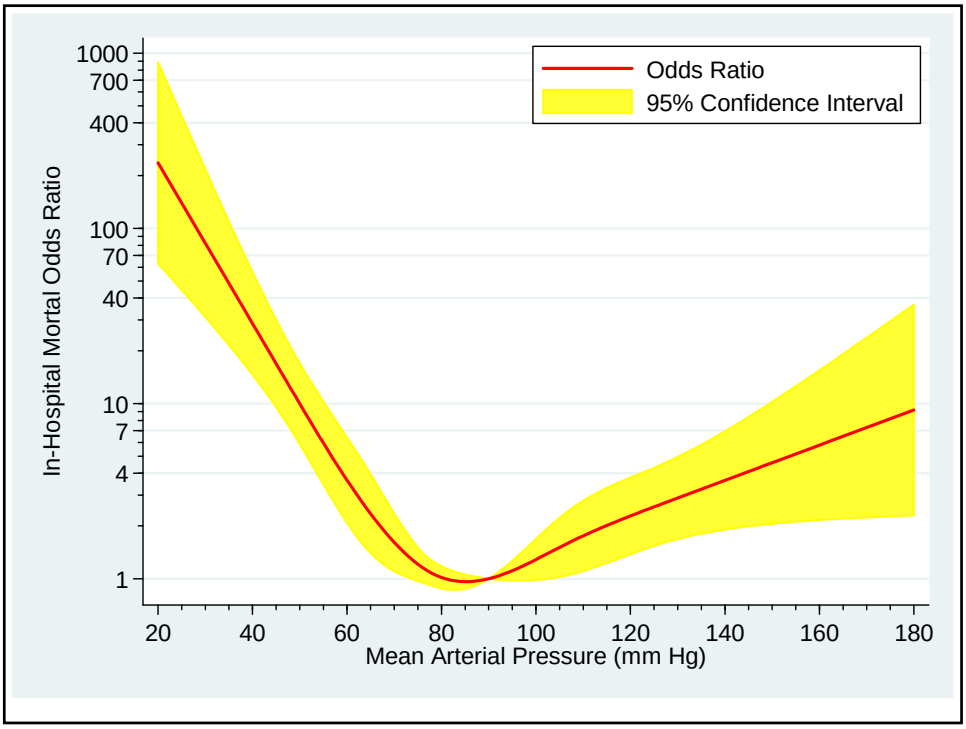
{18} The variables *lb_or* and *ub_or* equal the lower and upper bounds of the 95% confidence interval for this odds ratio

```
.
. twoway rarea lb_or ub_or map, color(yellow)              ///
>     || line or map, lwidth(medthick) color(red)          ///
>     ,  ylabel(1 (3) 10 40(30)100 400(300)1000, angle(0)) ///
>        ymtick(2(1)10 20(10)100 200(100)900) yscale(log)  ///        {19}
>        xlabel(20 (20) 180) xmtick(25 (5) 175)            ///
>        ytitle(In-Hospital Mortal Odds Ratio)             ///
>        legend(ring(0) position(2) order(2 "Odds Ratio"   ///
>            1 "95% Confidence Interval") cols(1))
```

{19} *yscale(log)* plots the *y*-axis on a logarithmic scale.



These dialogue boxes illustrate how to select a lograthmic scale for the *y*-axis



20. Frequency Matched Case-Control Studies

We often have access to many more potential control patients than case patients for case-control studies. If the distribution of some important confounding variable, such as age, differs markedly between cases and control, we may wish to adjust for this variable when designing the study. One way to do this is through **frequency matching**. The cases and potential controls are stratified into a number of groups based on, say, age. We then randomly select from each stratum the same number of controls as there are cases in the stratum. The data can then be analyzed by logistic regression with a classification variable to indicate the strata (see the analysis of the esophageal cancer and alcohol data in this chapter, Section 5 and 6).

It is important, however, to keep the strata fairly large if logistic regression is to be used for the analysis. Otherwise the estimates of the parameters of real interest may be seriously biased. Breslow and Day (Vol. I, p. 251-253) recommend that the strata be large enough so that each stratum contains at least 10 cases and 10 control. Even strata this large can lead to appreciable bias if the odds ratio being estimated is greater then 2.

a) Conditional logistic regression analysis

Sometimes there are more than one important confounders that we would like to adjust for in the design of our study.

In this case, we typically match each case patient to one or more controls with the same values of the confounding variables. This approach is often quite reasonable. However, it usually leads to strata (matched pairs or sets of patients) that are too small to be analyzed accurately with logistic regression. In this case, an alternate technique called conditional logistic regression should be used. This technique is discussed in Breslow and Day, Vol. I. In Stata, the *clogit* command may be used to implement these analyses.

21. What we have covered

- Extend simple logistic regression to models with multiple covariates
- Similarity between multiple linear and multiple logistic regression

  $\text{logit}\,(E(d_i)) = \alpha + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$
- Multiple 2x2 tables and the Mantel-Haenszel test
  - Estimating an odds ratio that is adjusted for a confounding variable
- Using logistic regression as an alternative to the Mantel-Haenszel test
- Using indicator covariates to model categorical variables

  i.varname notation in Stata

  ib#.varname notation in Stata
- Making inferences about odds ratios derived from multiple parameters

  The Stata *lincom* command
- Analyzing complex data with logistic regression
  - Multiplicative models
  - Models with interaction
- Assessing model fit
  - Testing the change in model deviance in nested models
  - Evaluating residuals and influence
- Using restricted cubic splines in logistic regression models
  - Plotting the probability of an outcome with confidence bands
  - Plotting odds ratios and confidence bands

    The Stata *predictnl* command

Cited References

Breslow, N. E. and N. E. Day (1980). Statistical Methods in Cancer Research: Vol. 1 - The Analysis of Case-Control Studies. Lyon, France, IARC Scientific Publications.

Knaus,W.A., Harrell, F.E., Jr., Lynn, J., Goldman, L., Phillips, R.S., Connors, A.F., Jr. et al. The SUPPORT prognostic model. Objective estimates of survival for seriously ill hospitalized adults. Study to understand prognoses and preferences for outcomes and risks of treatments. *Ann Intern Med.* 1995; 122:191-203.

Tuyns, A. J., G. Pequignot, et al. (1977). Le cancer de L'oesophage en Ille-et-Vilaine en fonction des niveau de consommation d'alcool et de tabac. Des risques qui se multiplient. *Bull Cancer* 64: 45-60.

For additional references on these notes see.

Dupont WD. *Statistical Modeling for Biomedical Researchers: A Simple Introduction to the Analysis of Complex Data. 2nd ed.* Cambridge, U.K.: Cambridge University Press; 2009.