I. REVIEW OF BIOSTATISTICS 1 AND SIMPLE LINEAR REGRESSION

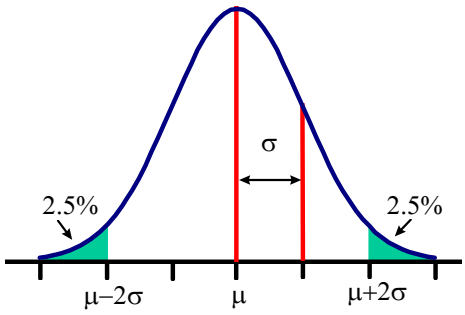
- Distinction between a parameter and a statistic
- The normal distribution
- Inference from a known sample about an unknown target population
- Simple linear regression: Assessing simple relationships between two continuous variables
- Interpreting the output from a linear regression program. Analyzing 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

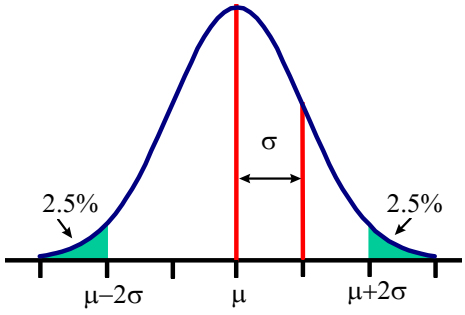
I. REVIEW OF BIOSTATISTICS 1 AND SIMPLE LINEAR REGRESSION

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.

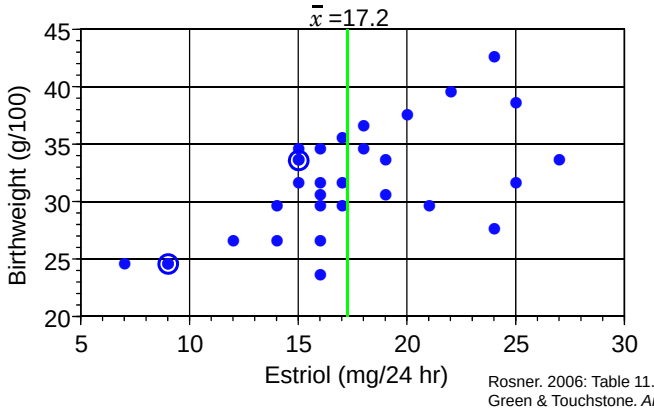
d) **Standard deviation** $\sigma = \sqrt{\sigma^2}$

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.



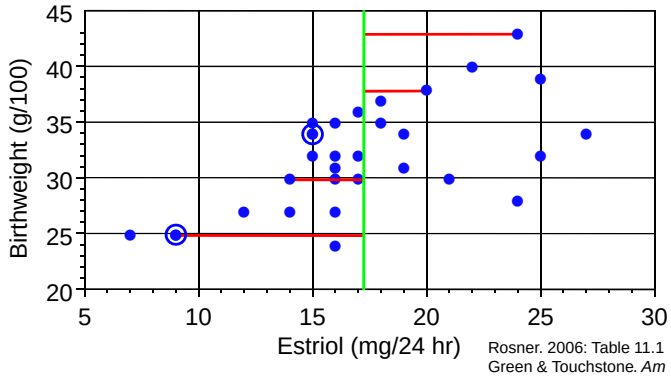
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

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.



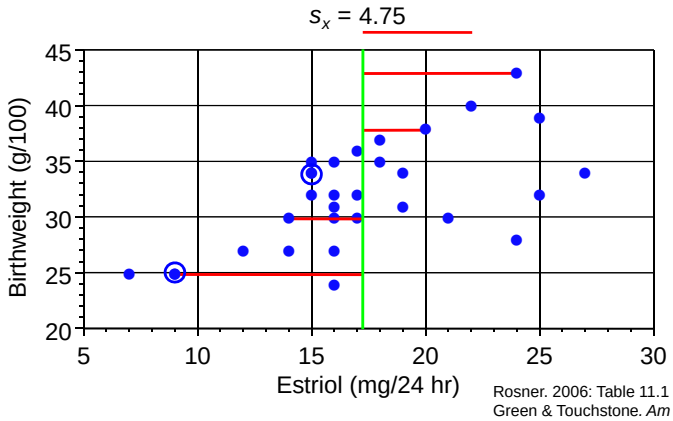
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

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

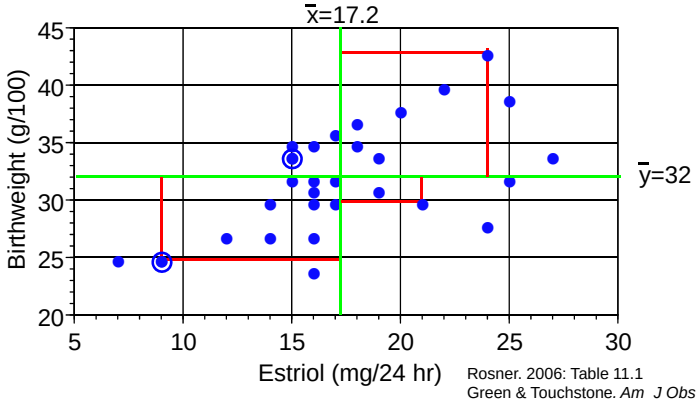
A statistic is **unbiased estimate** of a parameter if its expected value equals the parameter.

$\bar{x}$ 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**

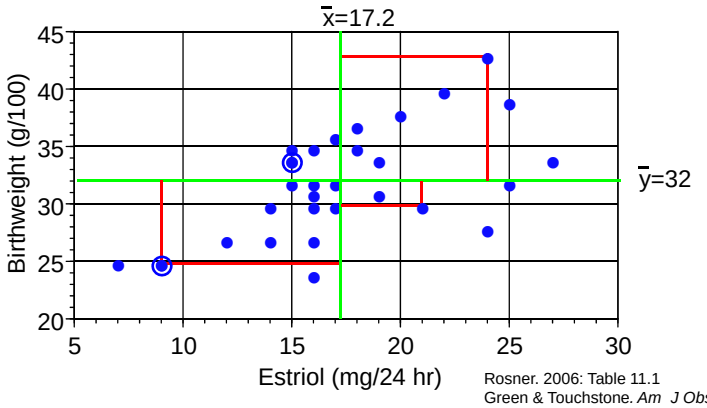
σ_{xy} = expected mean product of paired residuals

b) Sample covariance

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

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

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

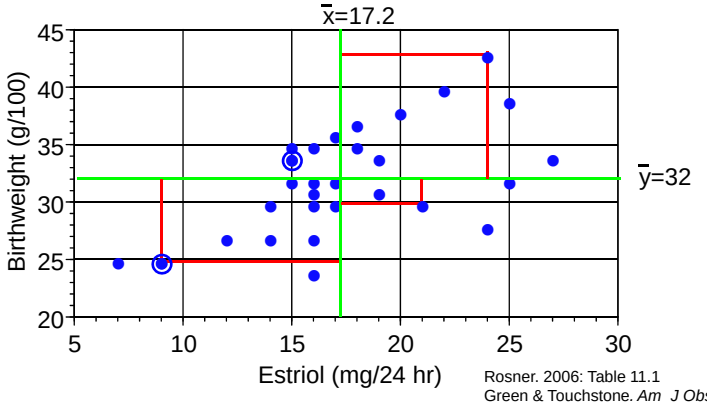


Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

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

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

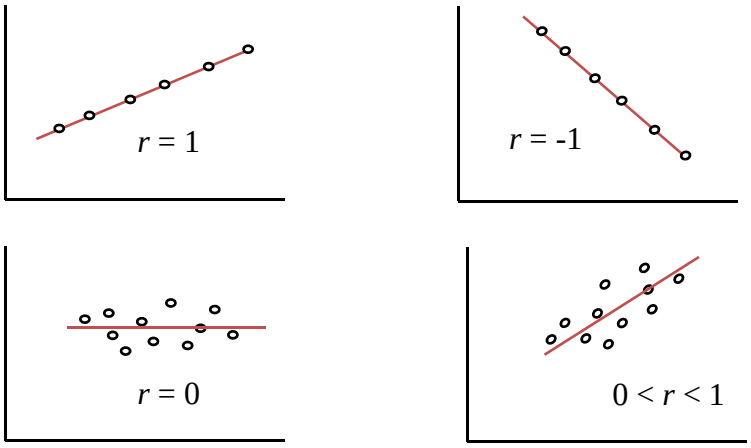
iii) $\sigma_{xy} > 0$ if most pairs fall in the upper right or lower left quadrant of the figure.

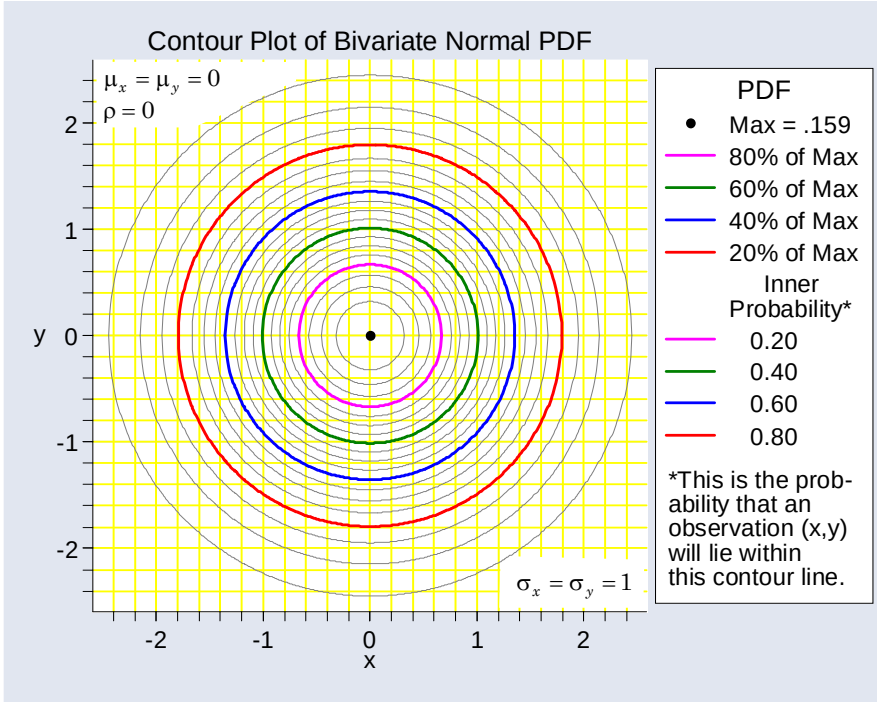


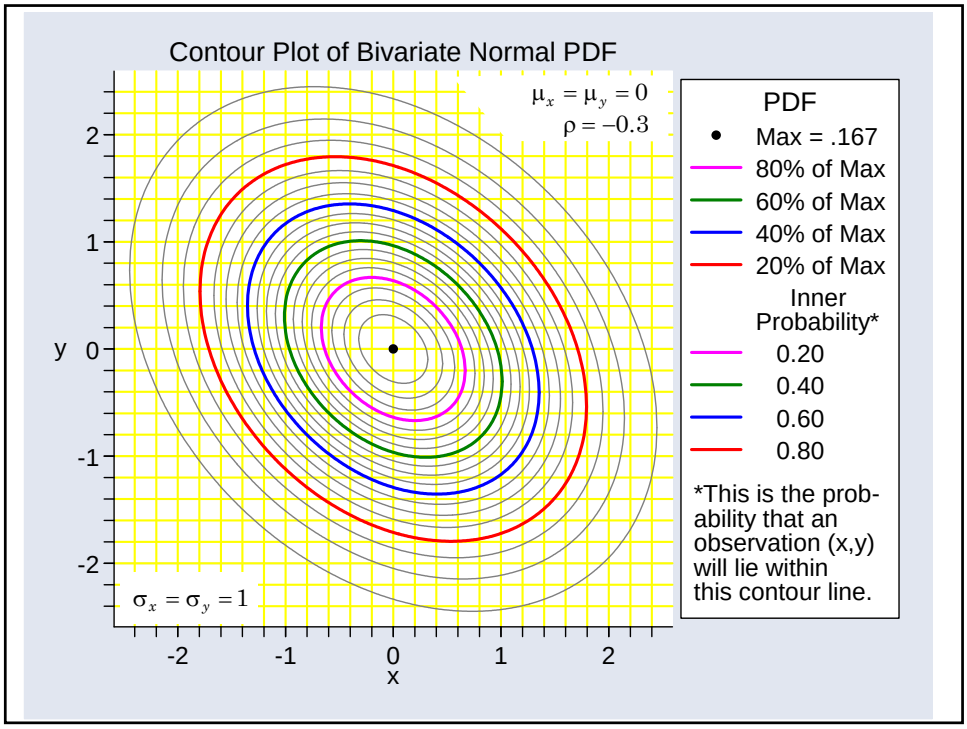
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

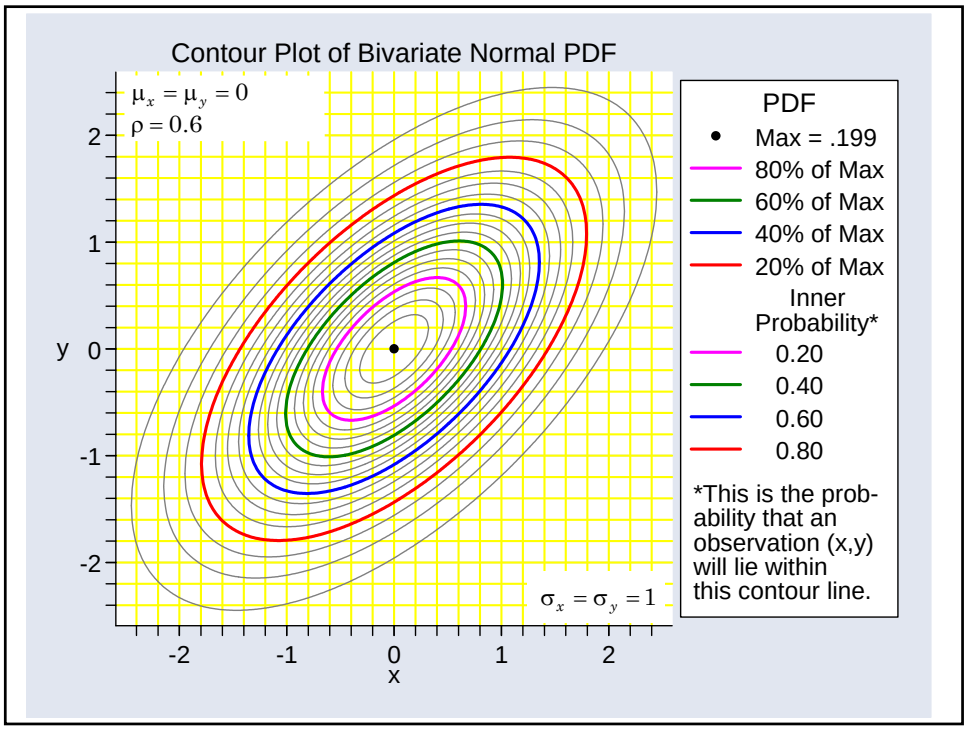
c) Correlation coefficient ρ

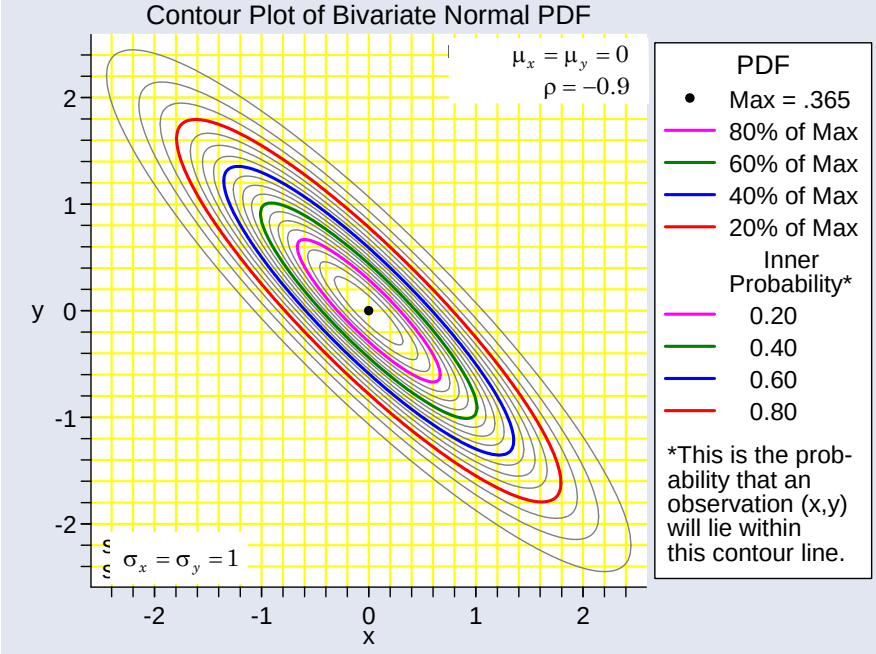
$\rho = \sigma_{xy} / (\sigma_x \sigma_y)$ is estimated by $r = s_{xy} / (s_x s_y)$ **{1.2}**






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

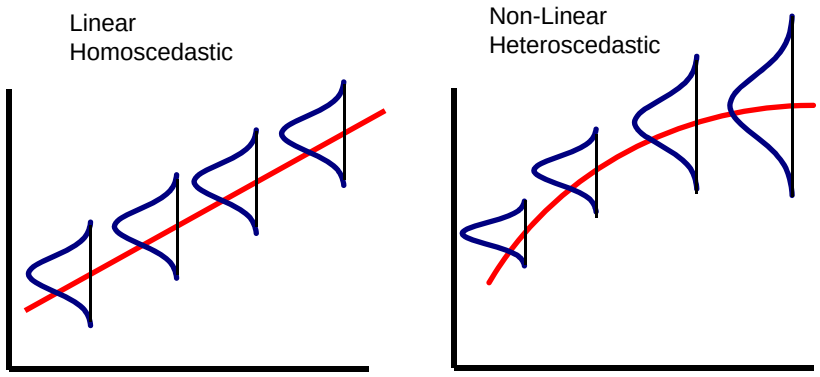
ε_i is assumed to be normally and independently distributed with mean 0 and standard deviation σ

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

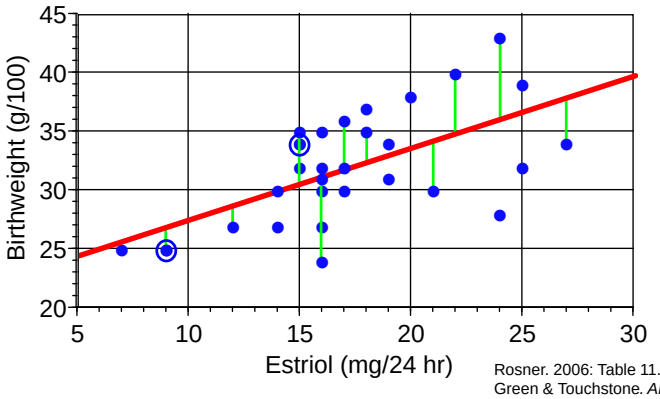
α and β are model parameters.

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

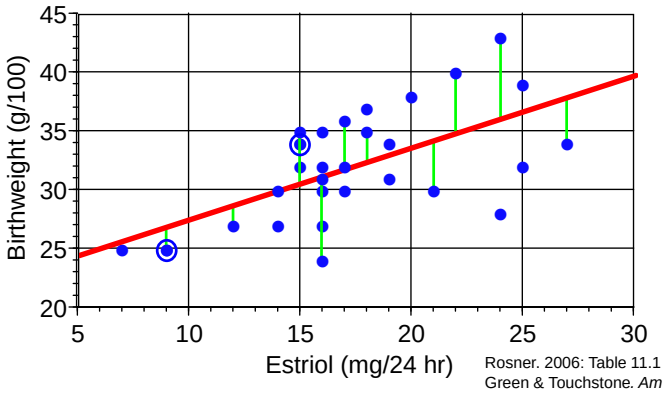
b) Implications of linearity and homoscedasticity.



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*



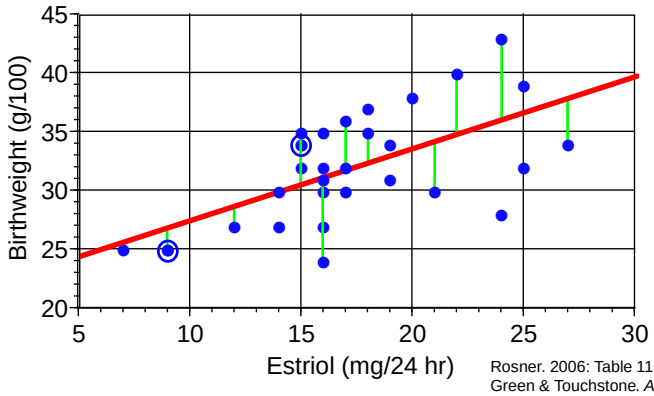
Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

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 $\alpha + \beta x$.

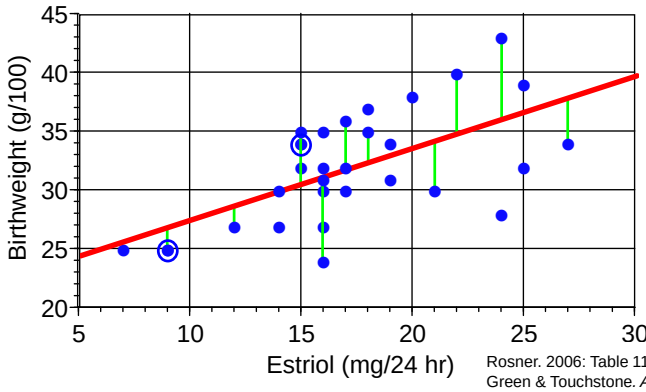
ii) Since $a = \bar{y} - b\bar{x}$

$\hat{y} = a + bx$ 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 $r \rightarrow 1$

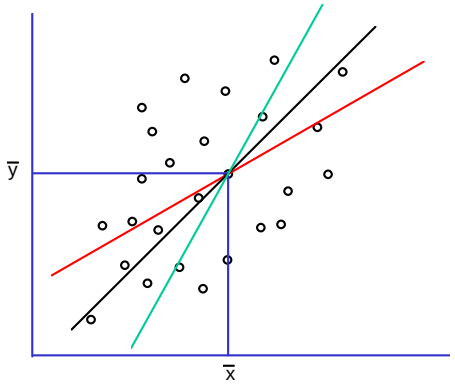


Rosner. 2006: Table 11.1
Green & Touchstone. *Am J Obs Gyn 1963;85:1-9*

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

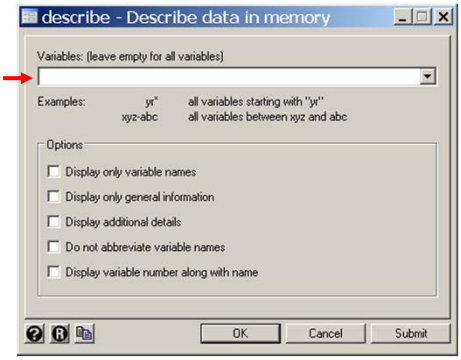
and clicking on Example Logs and Data from Lecture Notes.

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**

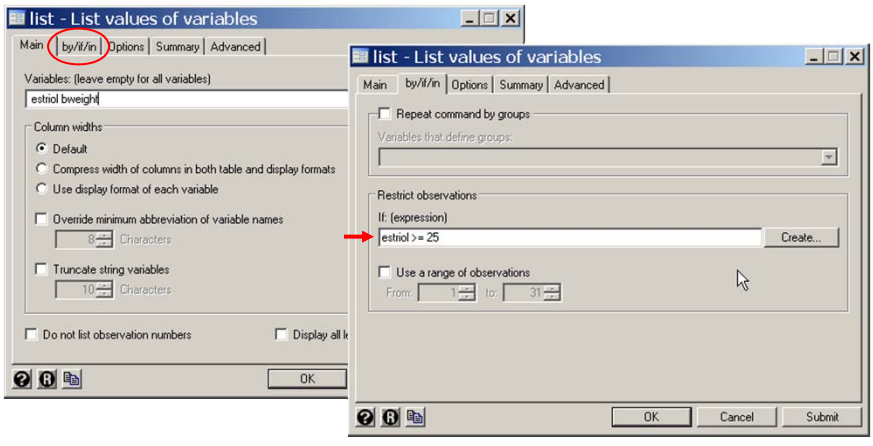
{3} The use command reads the Rosner data into memory. The previous contents of memory are purged.



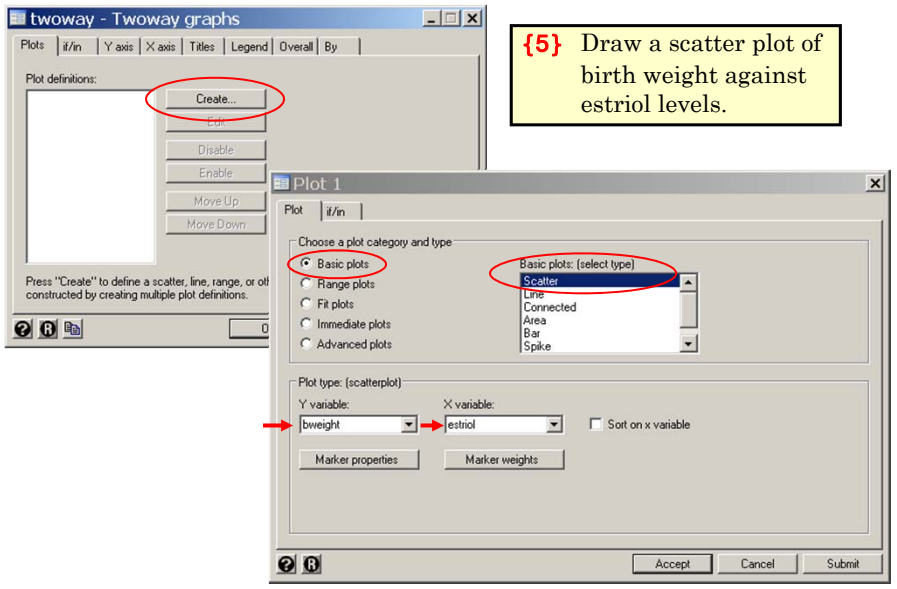
```
. * Data > Describe data > List data
. list estriol bweight if estriol >= 25                              {4}

      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.

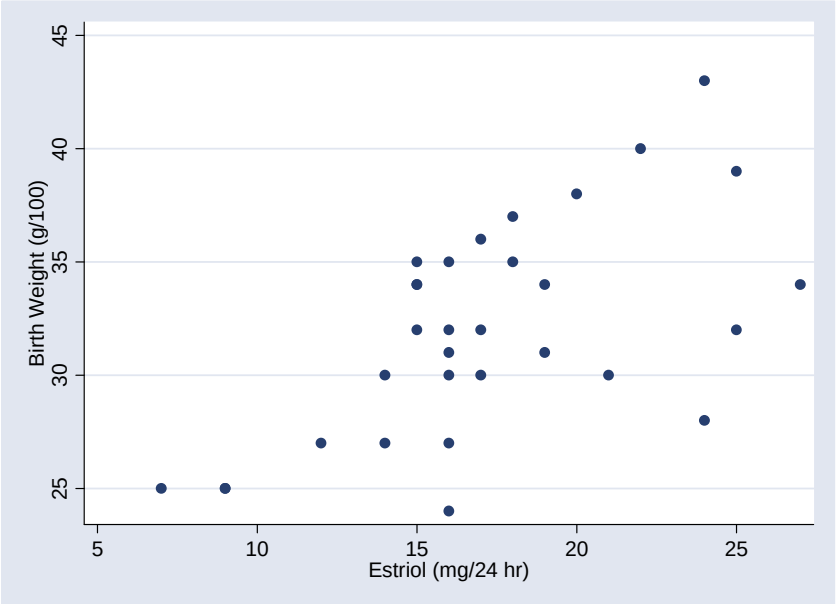
There is no obvious distinction between command modifiers like *if* and command options. In general, modifiers apply to most Stata commands while options are specific to a given class of commands. See you manuals and command summary in the text.



```
. * 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

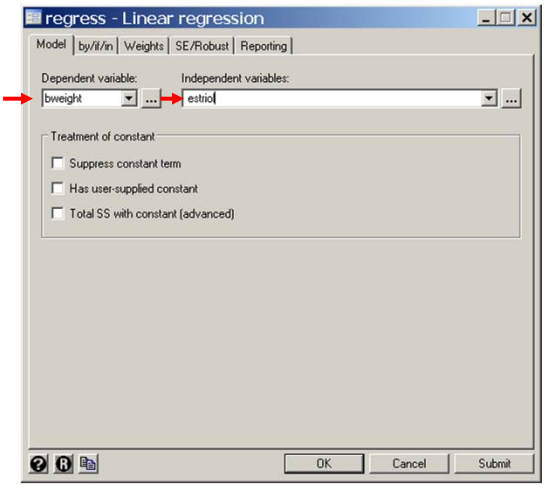
a) Running a simple linear regression program

```
. *  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.

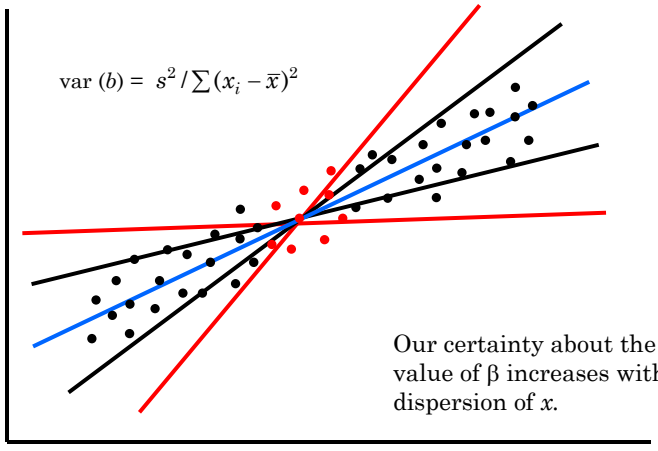
ii) The null hypothesis that β = 0 can be rejected with P< 0.0005

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.

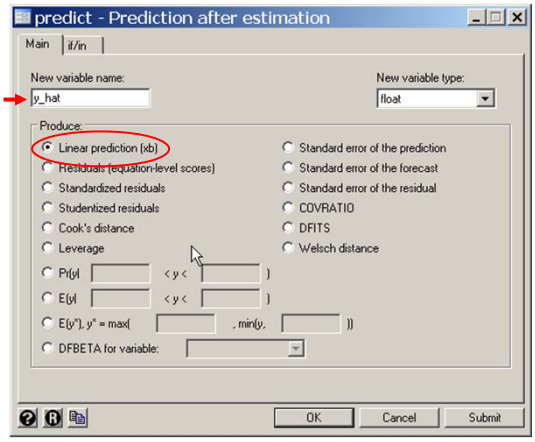
```
  Source |       SS           df          MS              Number of obs =      31
---------+------------------------------------           F(  1,    29) =   17.16
   Model |  250.574476         1     250.574476          Prob > F      =  0.0003
Residual |  423.425524 n-2 = 29 s²  = 14.6008801         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.

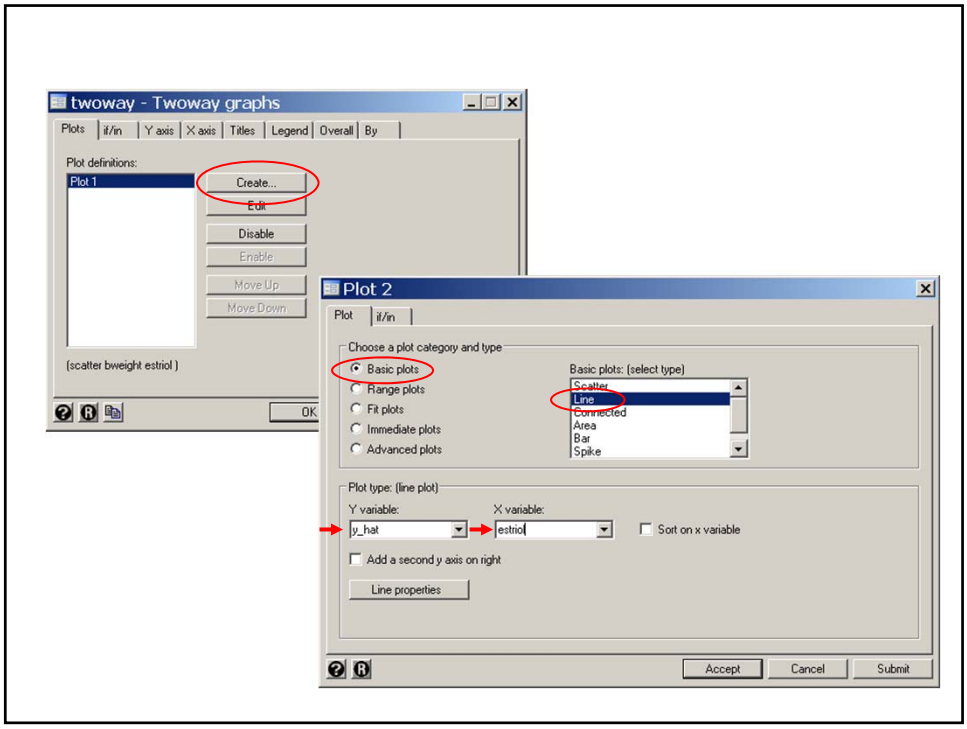


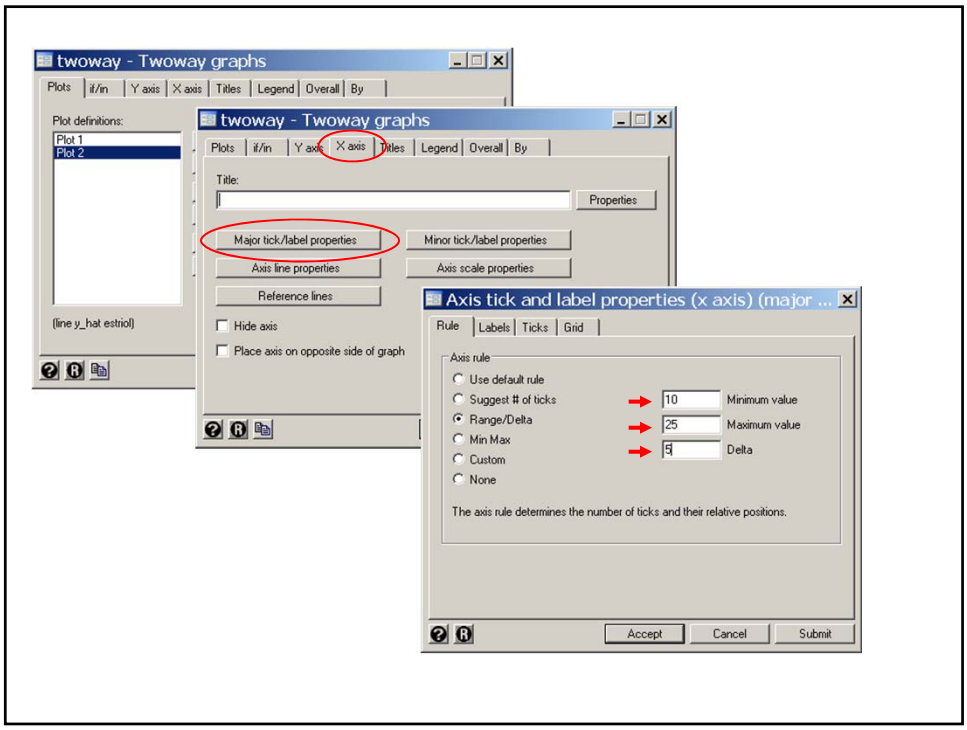

```
. * Graphics > Twoway graph (scatter, line, etc.)
. twoway scatter bweight estriol                                    ///    {2}
>     || line y_hat estriol                                         ///    {3}
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")   {4}
```

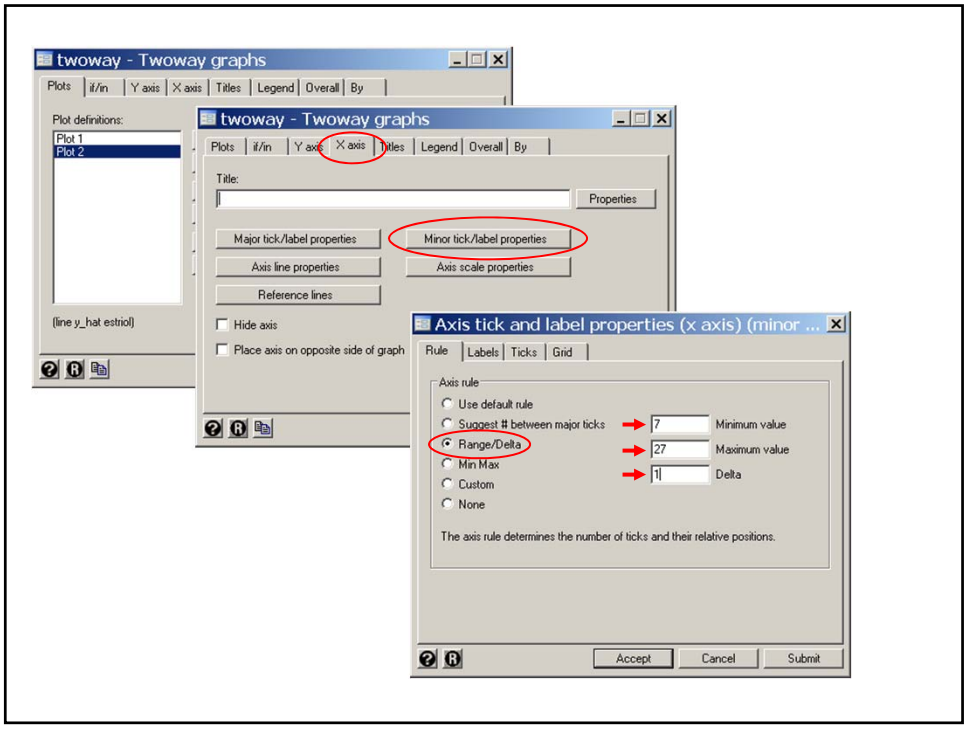
{2} This command is written over several lines to improve legibility. The three slashes (///) indicate that this command continues on the next line. These slashes are permitted in do files but not in the Command window.

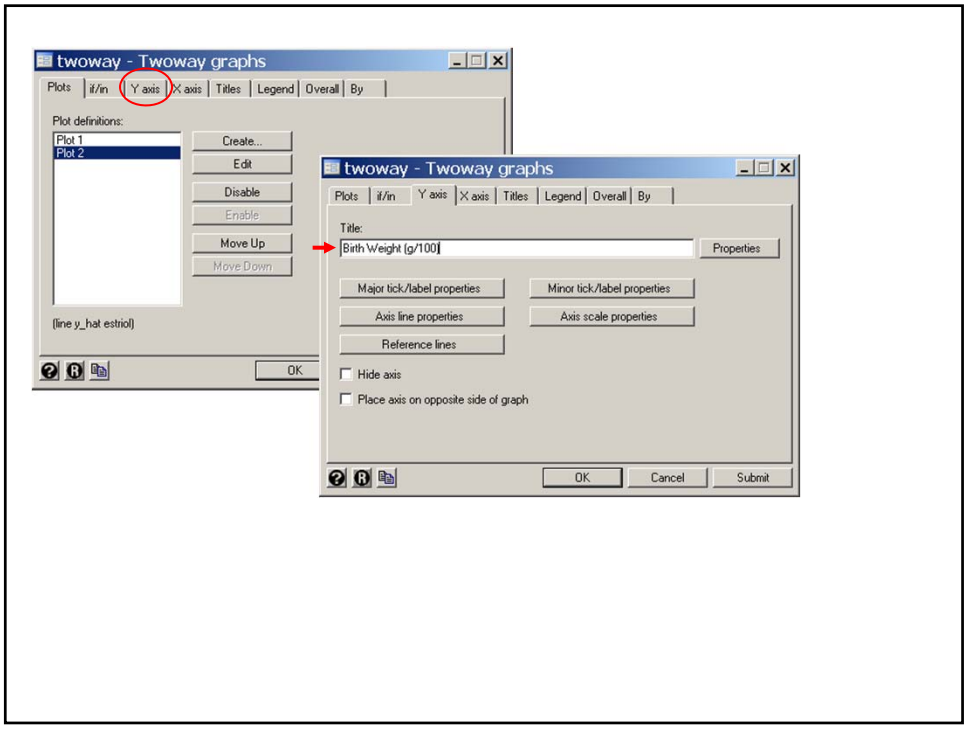
{3} A double bar (||) indicates that another graph is to be overlaid on top of the preceding one. *line y_hat estriol* indicates that *y_hat* is to be plotted against *estriol* with the points connected by a straight line.

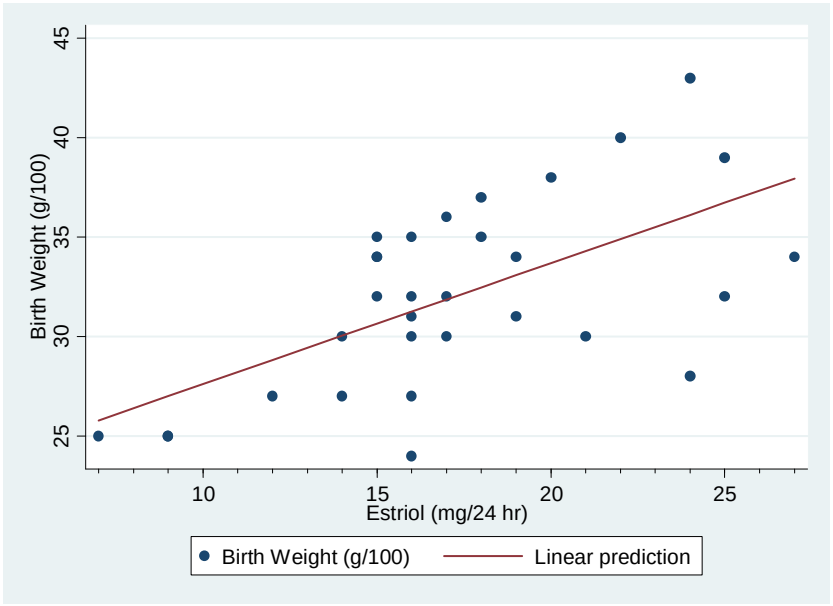
{4} This *xlabel* option labels the x-axis from 10 to 25 in steps of 5. *xmtick* adds tick marks from 7 to 27 in unit steps. *ytitle* gives a title to the y-axis


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

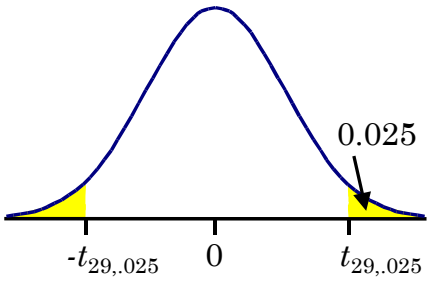

twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
Edit
Disable
Enable
Move Up
Move Down
(line y_hat estriol)
OK
Title:
Birth Weight (g/100)
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale 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*.

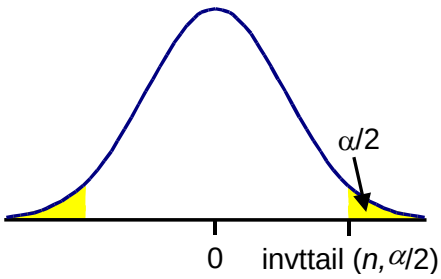
That is $std\_p = \sqrt{\text{var}(\hat{y}(x))}$

{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) =

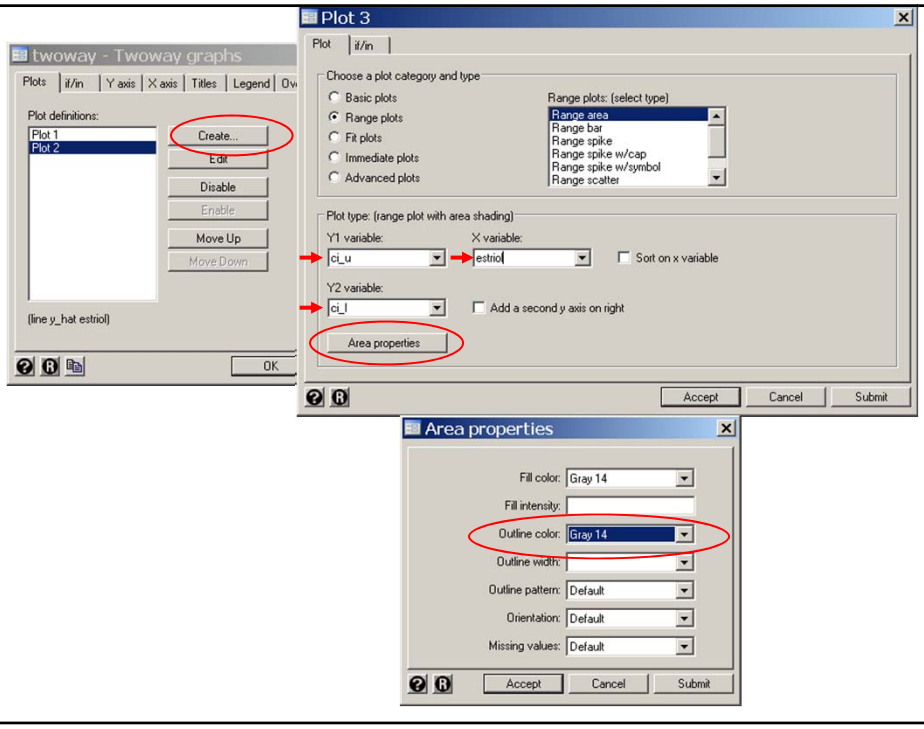
invttail(29, 0.025) = $t_{29,0.025} = 2.045$

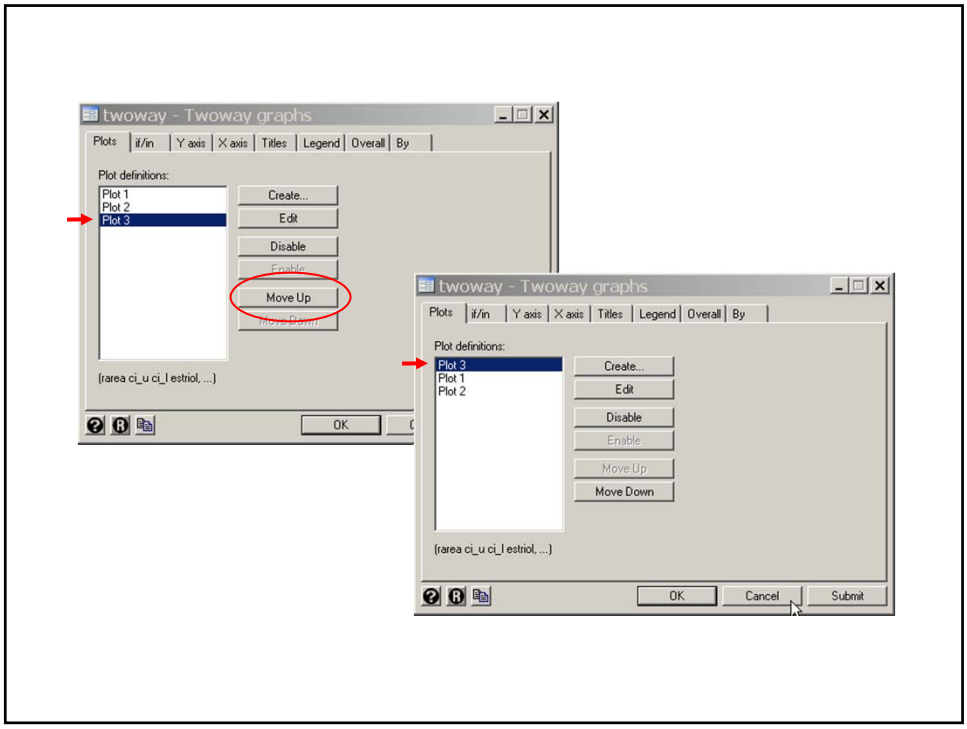


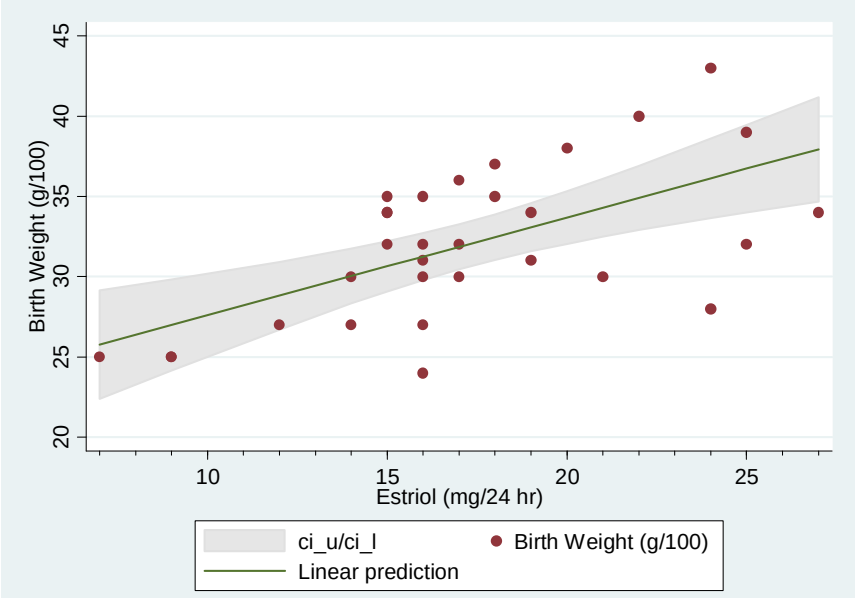
```
. twoway rarea   ci_u ci_l estriol, color(gs14)                    /// {3,4}
>     || scatter bweight estriol                                     ///
>     || line y_hat  estriol                                         ///
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")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

{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.



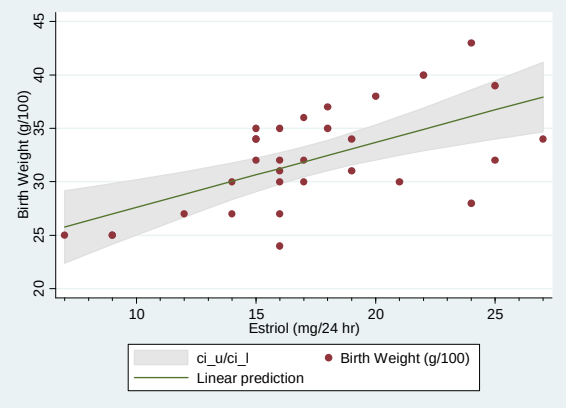




N.B. More than 5% of the observations lie outside of the 95% CI band. Band is for the regression line not the observations.

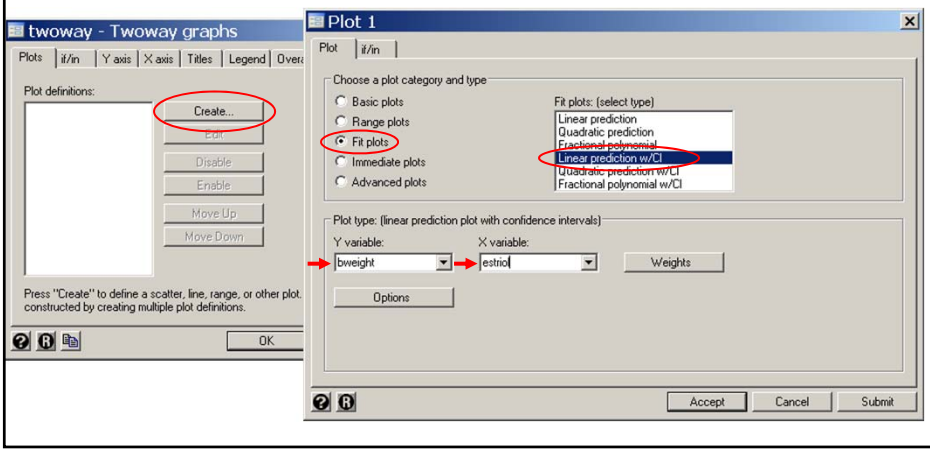
$$\hat{y} \pm t_{n-2,0.025}\sqrt{\text{var}(\hat{y}(x))}$$ Is the 95% confidence band for $\hat{y}(x)$

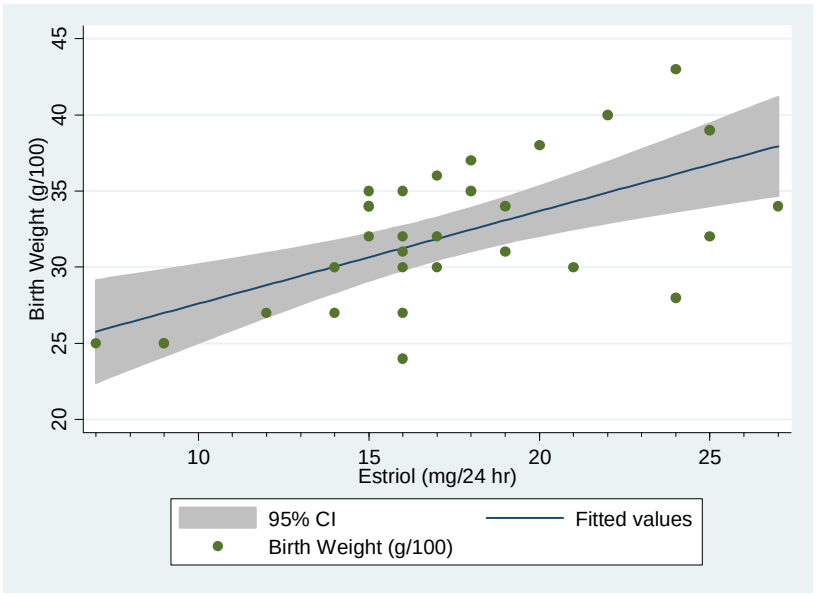
```
. predict std_p , stdp
. generate ci_u = y_hat + invttail(_N-2,0.25)*std_p
. generate ci_l = y_hat - invttail(_N-2,0.25)*std_p
. twoway rarea ci_u ci_l estriol, bcolor(gs14)
>     || scatter bweight estriol                                          ///
>     || line y_hat  estriol                                              ///
>     ,  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*.



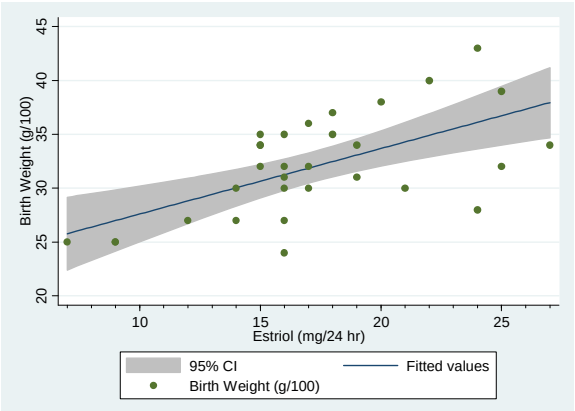


The variance of $\hat{y}(x)$ is

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

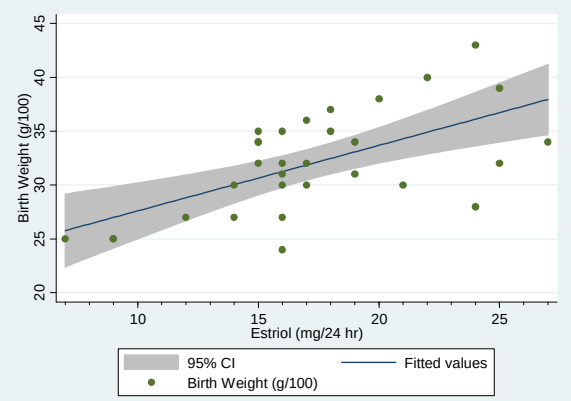
The 95% confidence interval for $\hat{y}(x)$ is

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



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

- 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.

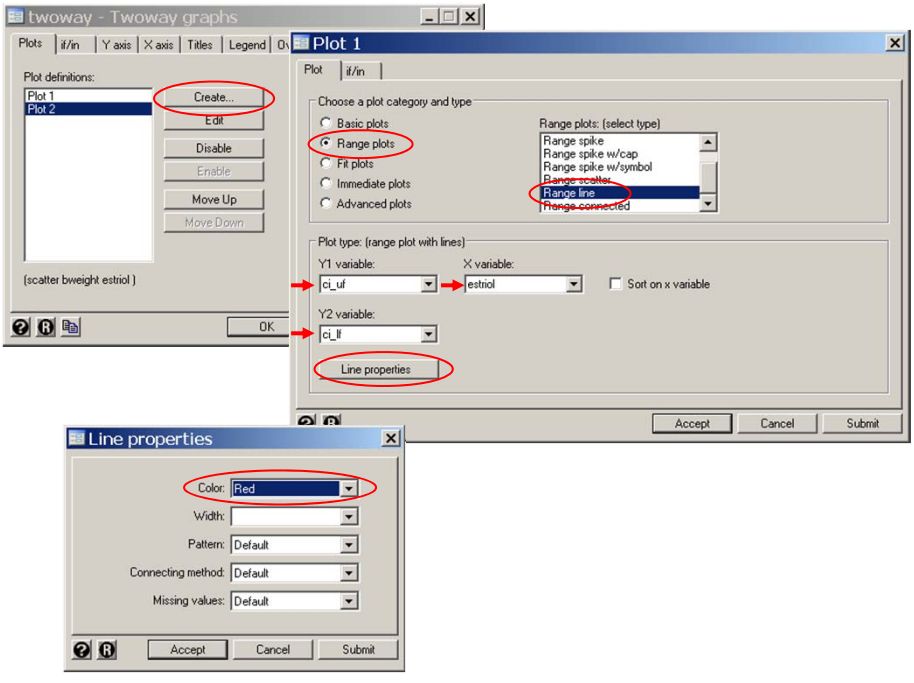
That is *std_f* = $\sqrt{\text{var}(\hat{y}) + s^2}$

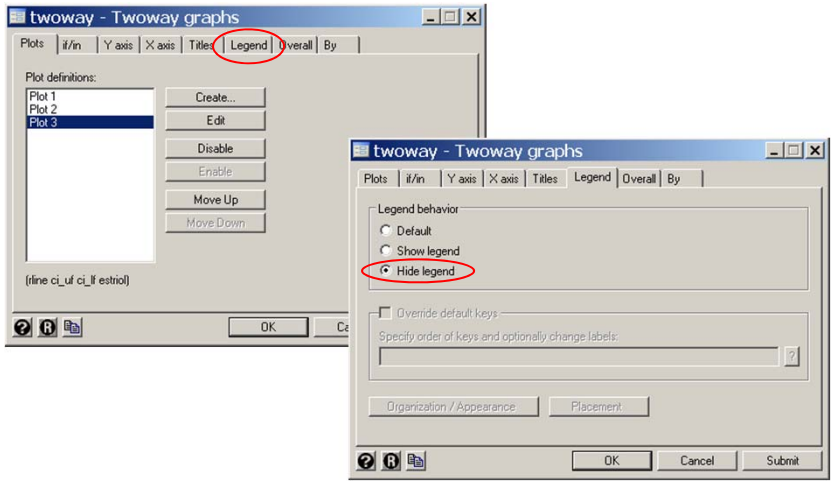
{2} **rline** plots both *ci_uf* and *ci_lf* against *estriol*.

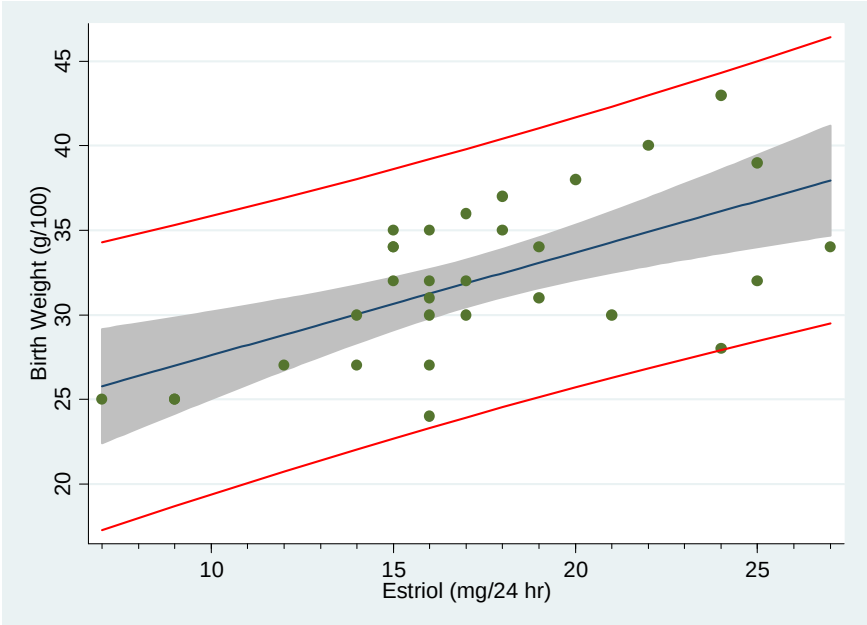
color specifies the color of the plotted lines.

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.

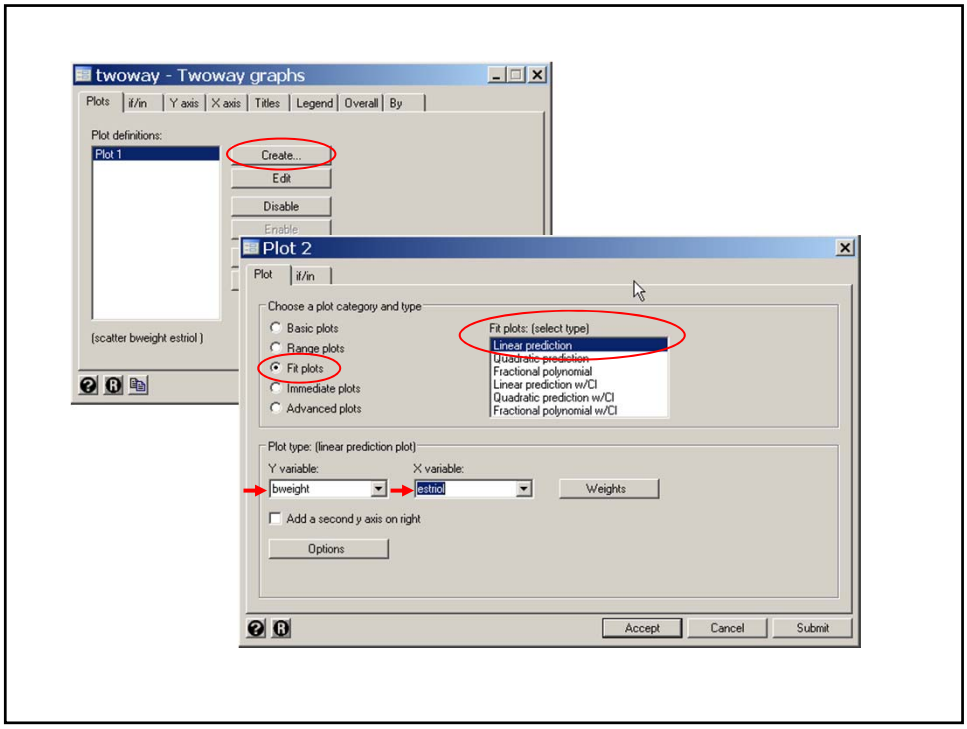
The listing of *Birth_Weight*.LR.*log* continues as follows

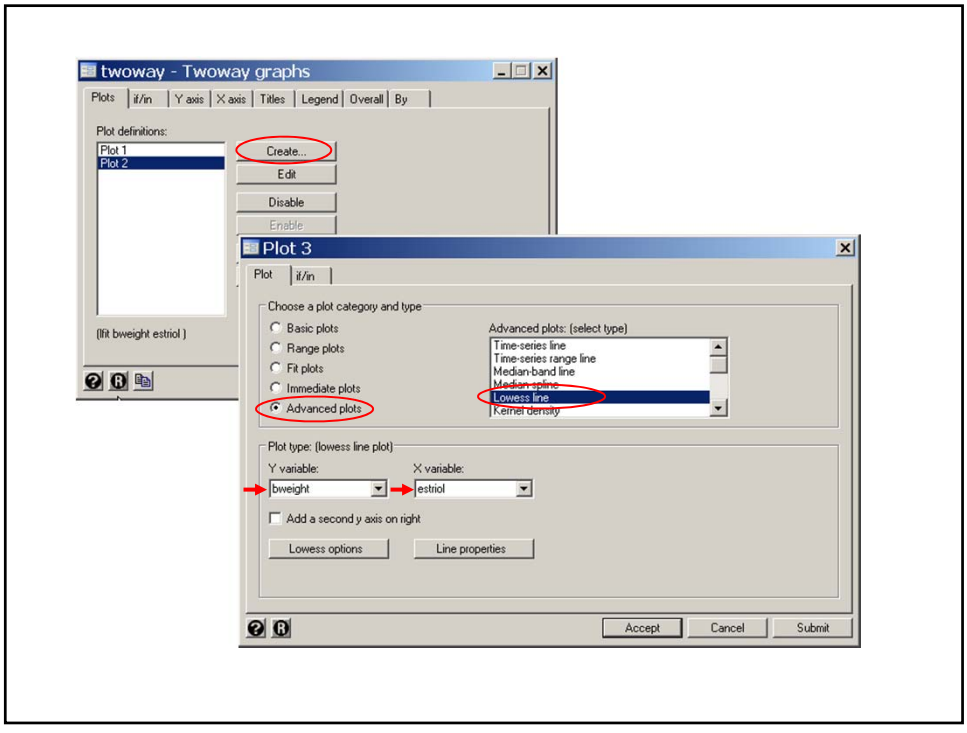
```
.twoway scatter bweight estriol                                    ///
>     || lfit bweight estriol                                      /// {1}
>     || lowess bweight estriol , bwidth(.8)                       /// {2}
>     ,  xlabel(10 (5) 25) xmtick(7 (1) 27) ytitle("Birth Weight (g/100)")
```

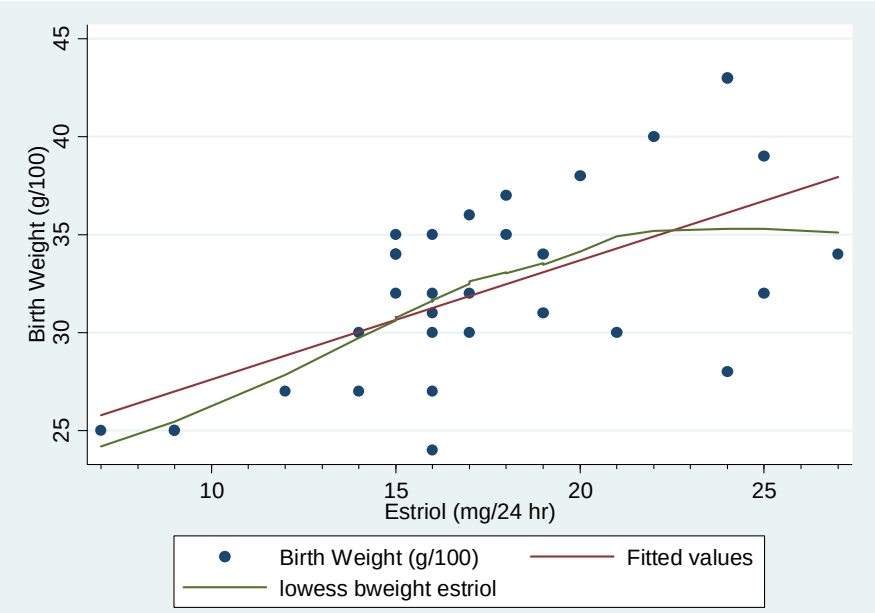
{1} This **lfit** command plots the linear regression line for *bweight* against *estriol*.

{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.

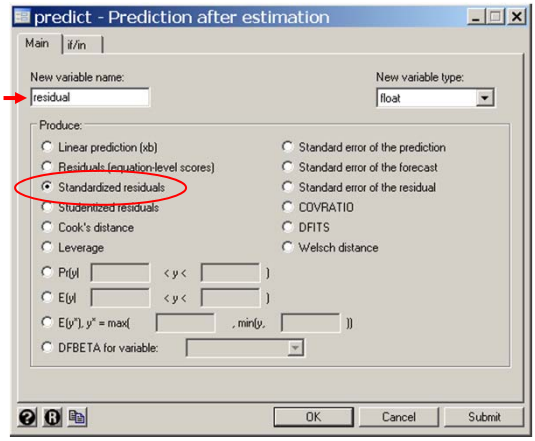
t_i is sometimes referred to as the jackknife residual

Plotting these studentized residuals against x_i assesses the homoscedasticity assumption.

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

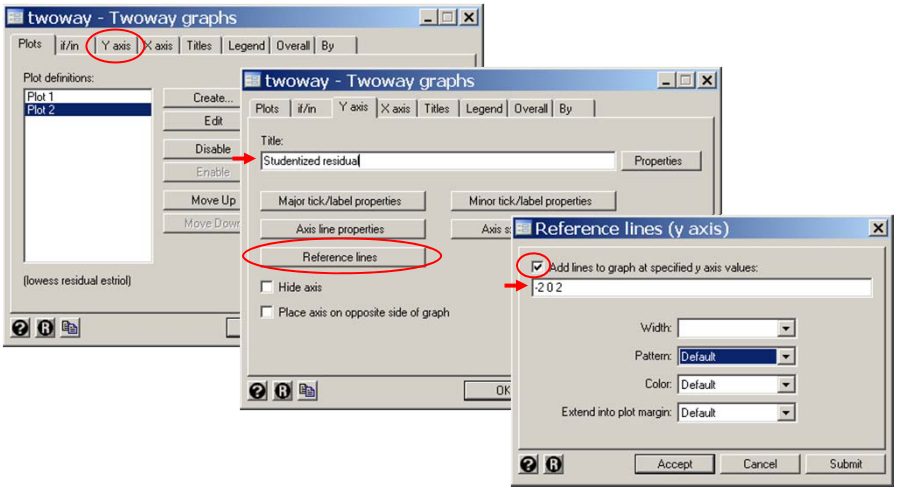
```
. predict residual, rstudent                                    {1}
```

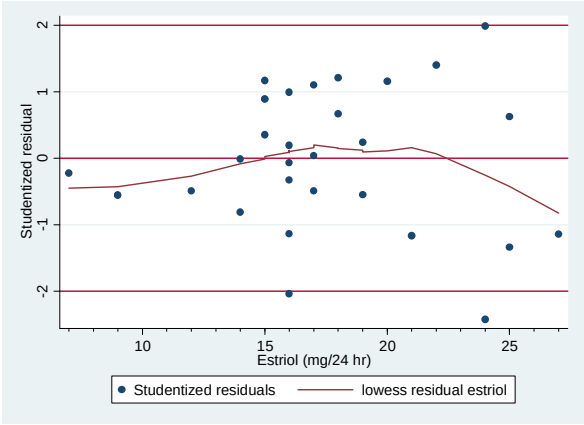
{1} Define *residual* to be a new variable that equals the studentized residual.



```
. 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.

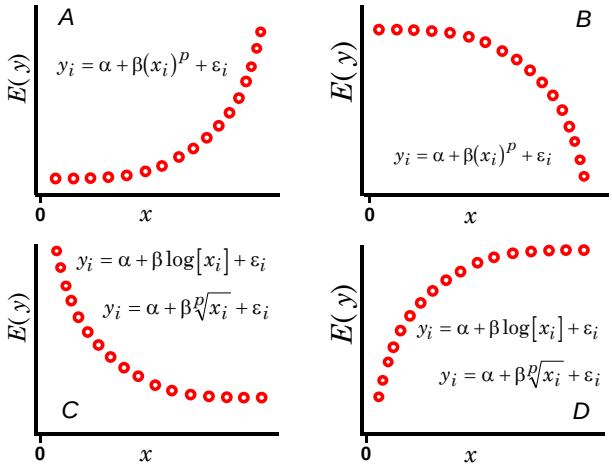
19. Normalizing the Data Distribution

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



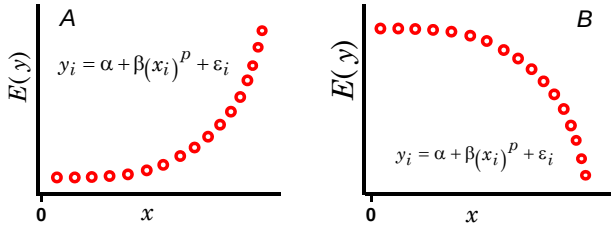
If x is positive then models of the form

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

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

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

should be considered for some $p > 1$.



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

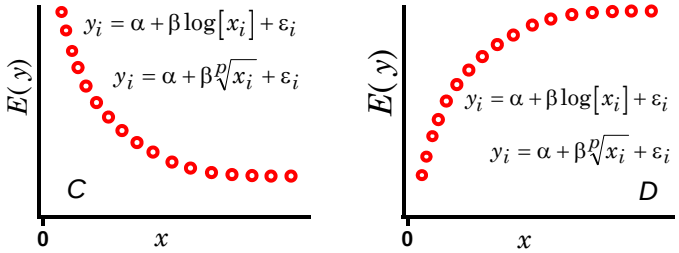
If x is positive then models of the form

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

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

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

should be considered for some $p > 1$.



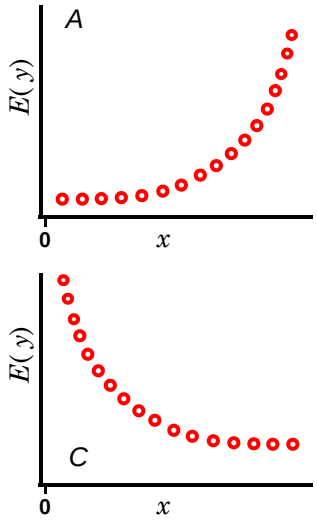
Data similar to panels C and D may be modeled with equations {1.14} and {1.15}.

The best value of p is found empirically.

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

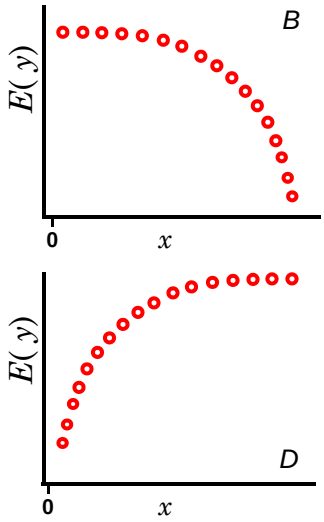
$$\log[y_i] = \alpha + \beta x_i + \varepsilon_i \qquad \{1.16\}$$

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.

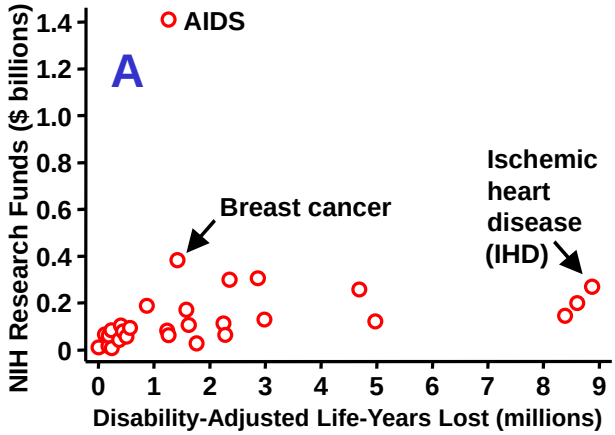


Figure A shows the untransformed scatter plot. Funding for AIDS is 3.7 times higher than for any other disease

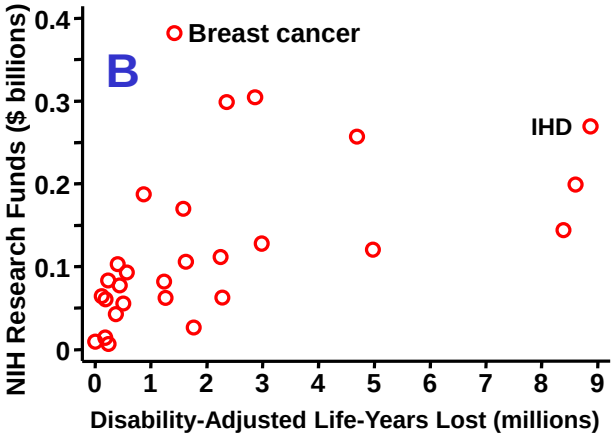


Figure B is similar to panel A except the AIDS data has been deleted and the *y*-axis has been re-scaled.

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

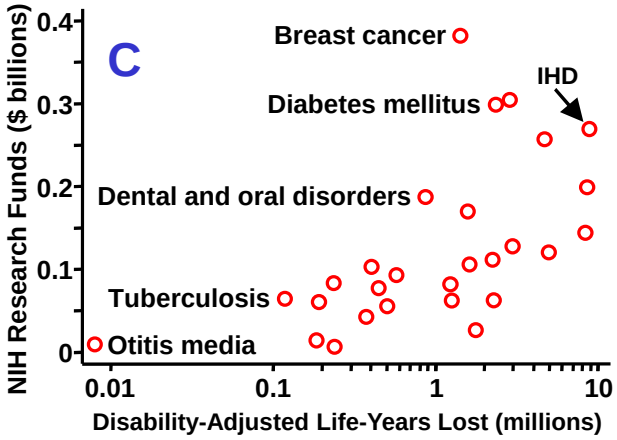
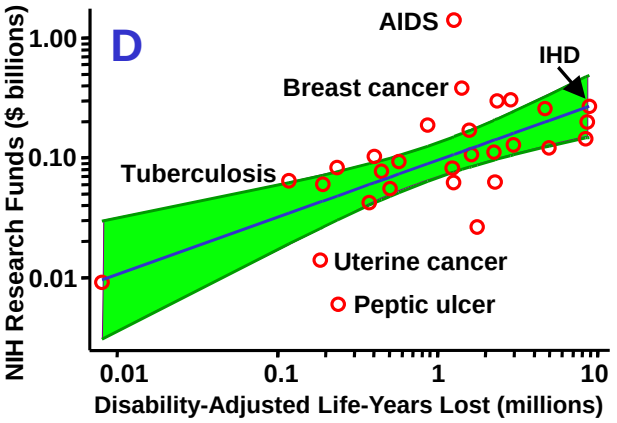


Figure C shows funding plotted against log disability-adjusted life-years.

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.

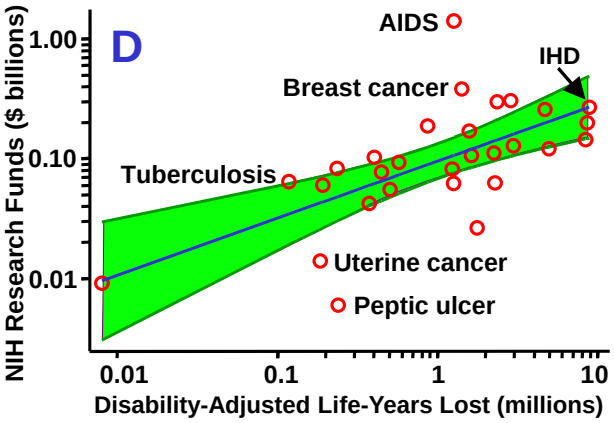


In **Figure D** we plot log funding against log disability.

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.

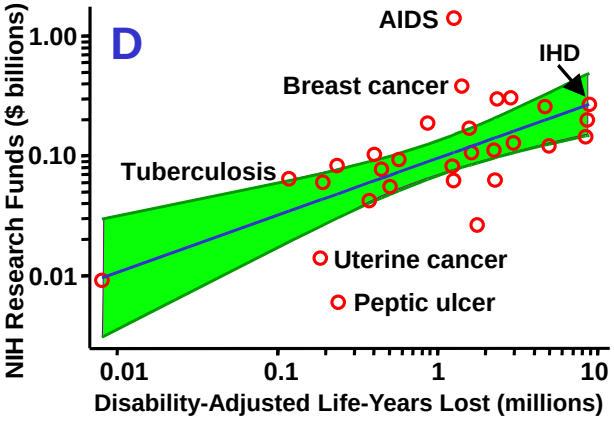
The linear regression line and associated 95% confidence intervals are shown in this panel.



The model for this linear regression is

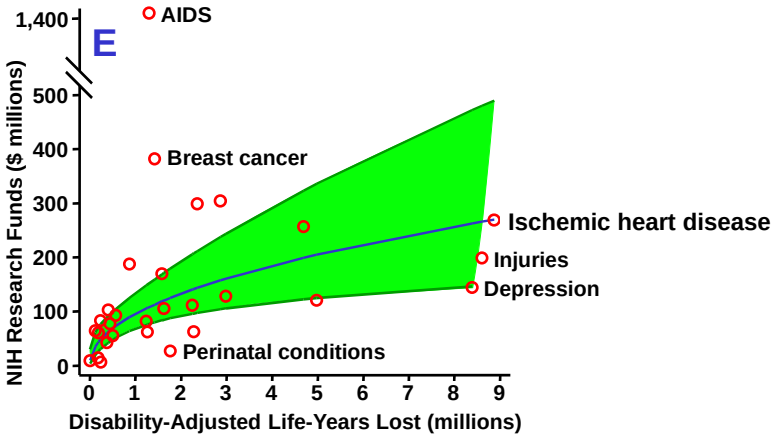
$$E[\log[y_i]]=\alpha+\beta\log[x_i] \qquad \{1.19\}$$

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.

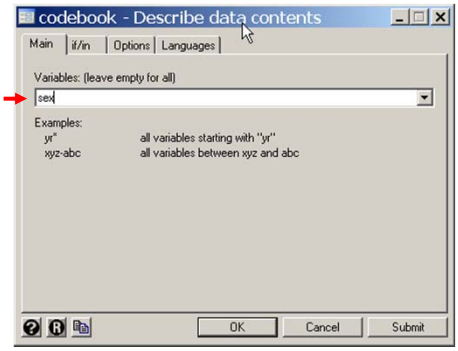
s_2 and $\text{var}(b_2)$ are similarly defined for group 2.

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
```

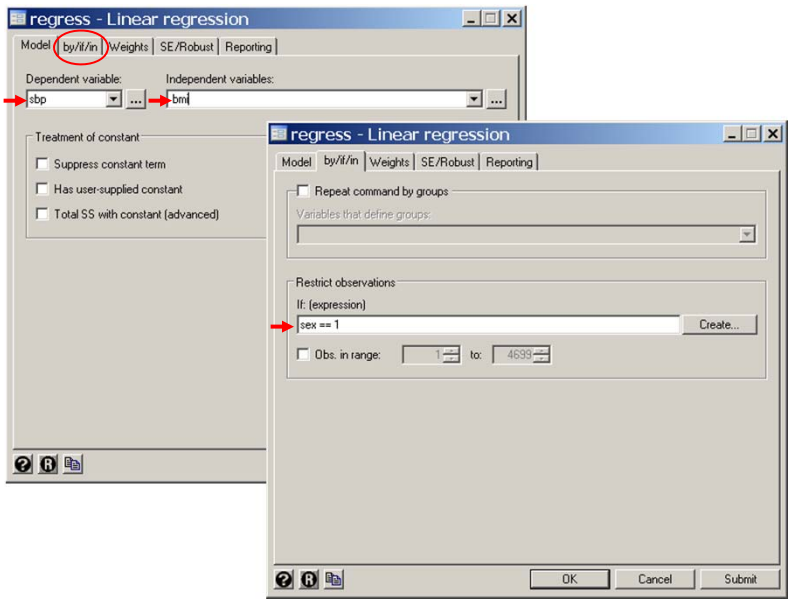
{1} The *codebook* command provides a summary of variables. *sex* takes numeric values 1 and 2 that denote men and women, respectively.



```
. 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.

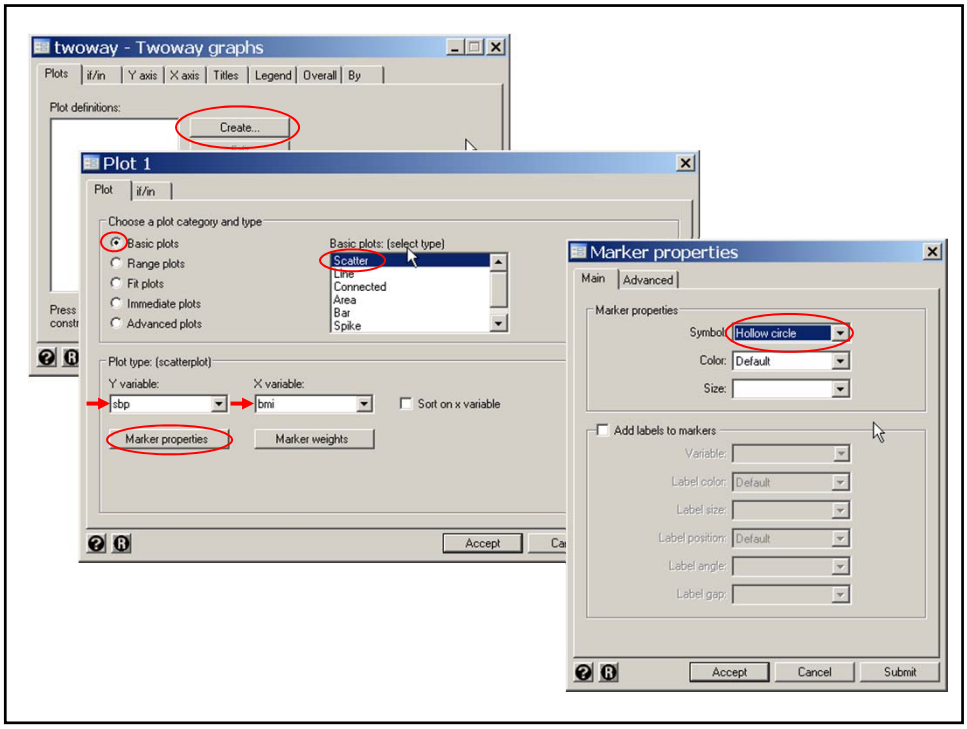


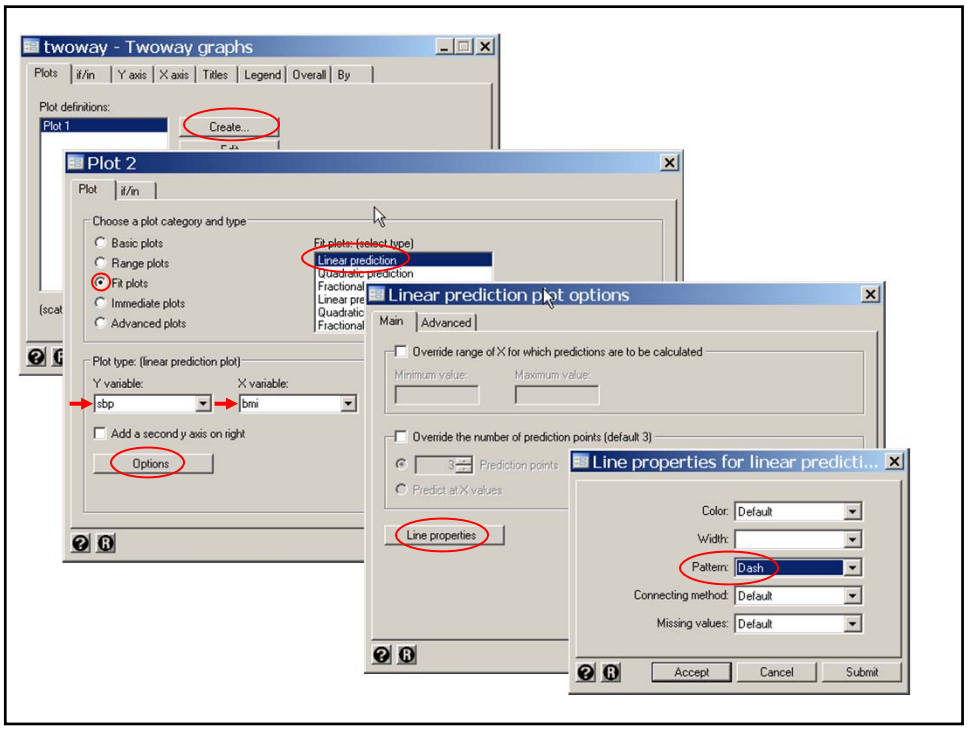
```
. regress sbp bmi if sex == 2

  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)
```

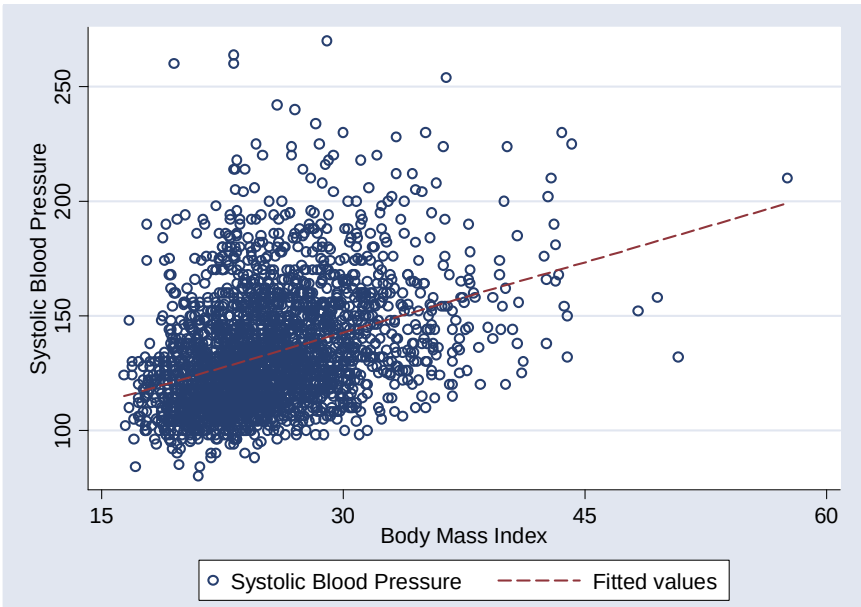
```
. sort bmi

. *
. *  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.

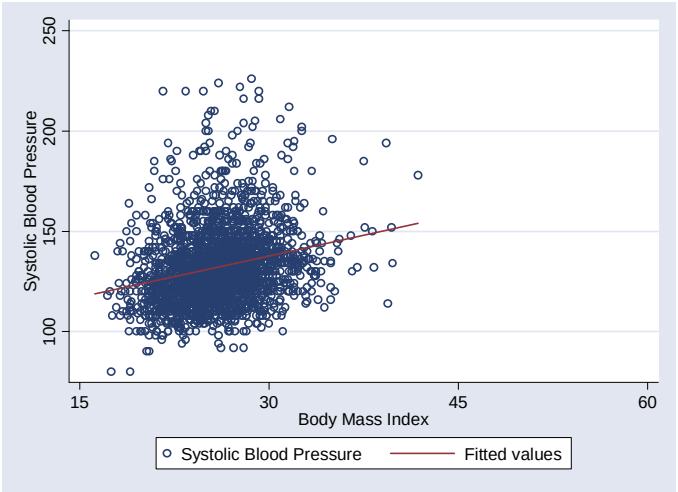
{2} **lpattern(dash)** specifies that a dashed line is to be drawn.


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}
```

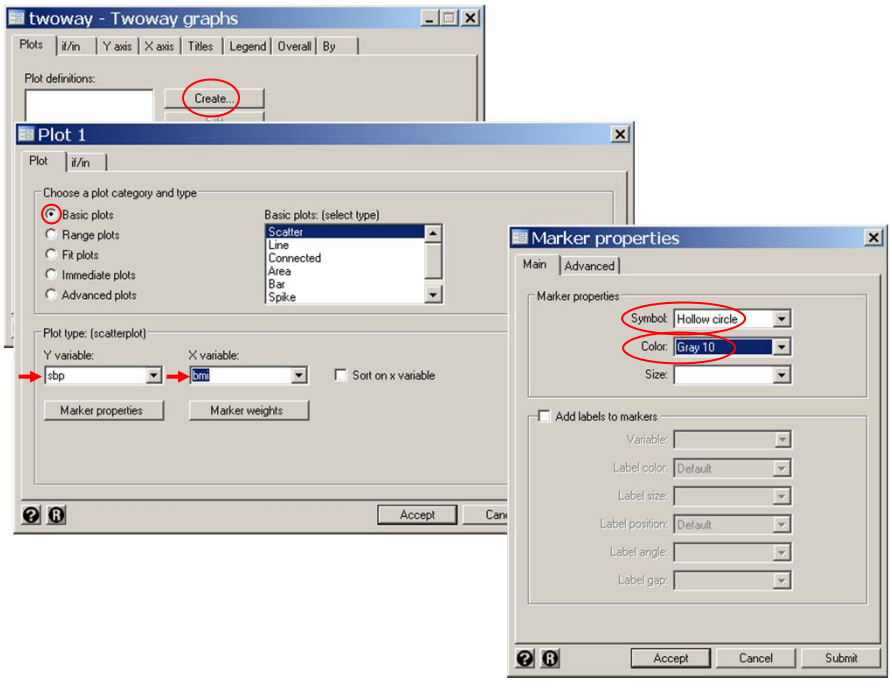
{1} **color** specifies the marker color used to specify individual observations. **gs10** indicates a gray is to be used. A gray scale of 0 is black; 16 is white.

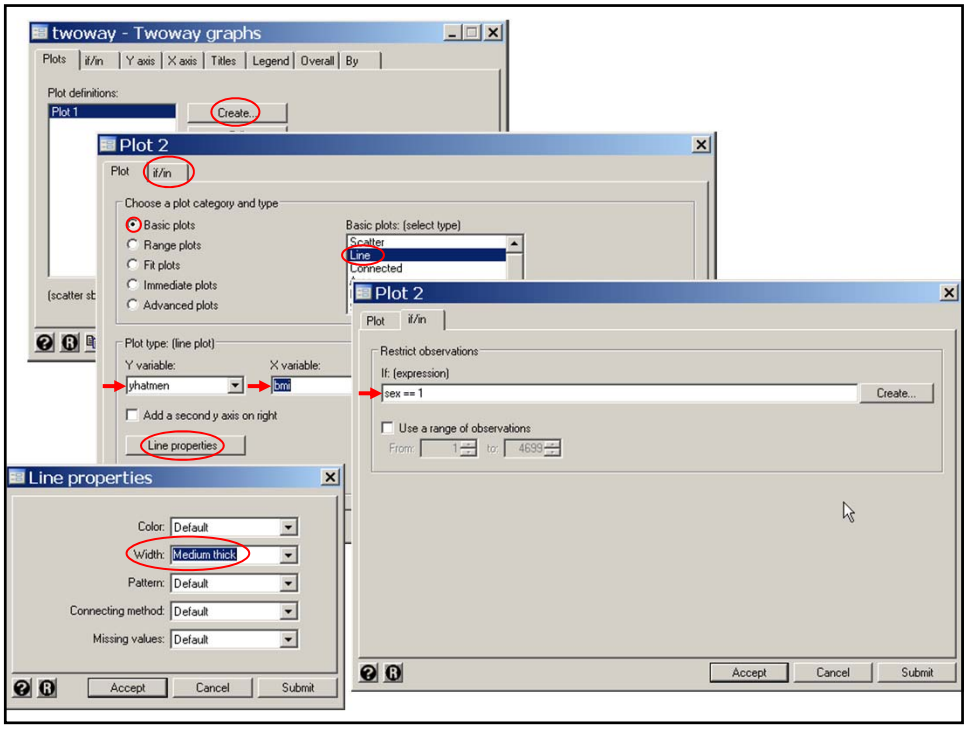
{2} **lwidth(medthick)** specifies width for the line

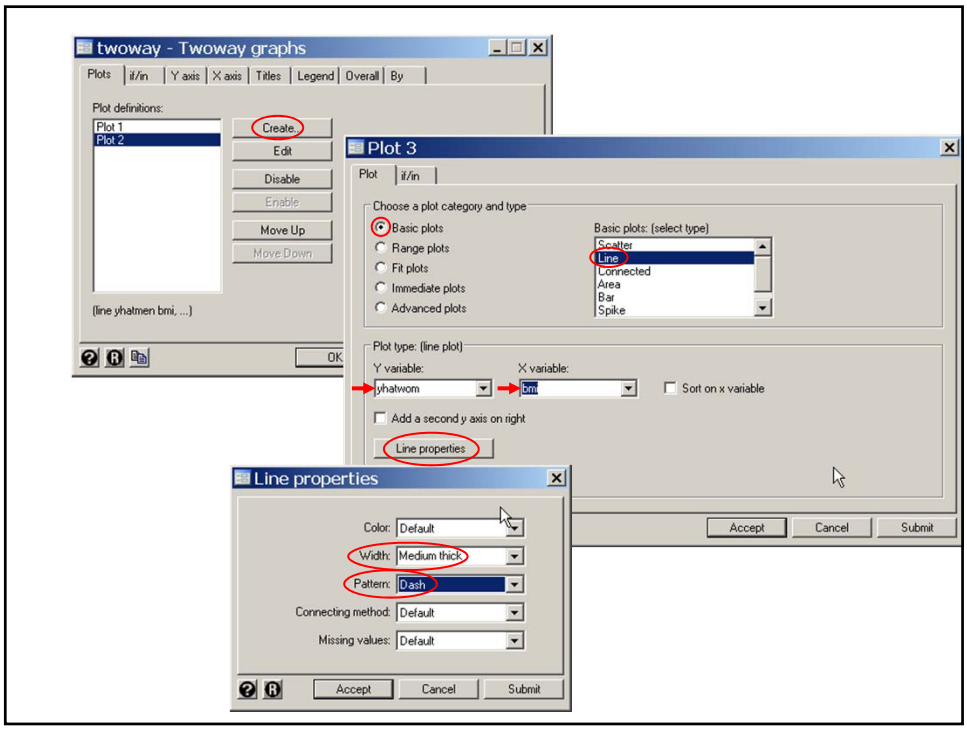
{3} **xsize** specifies the width of the graph in inches

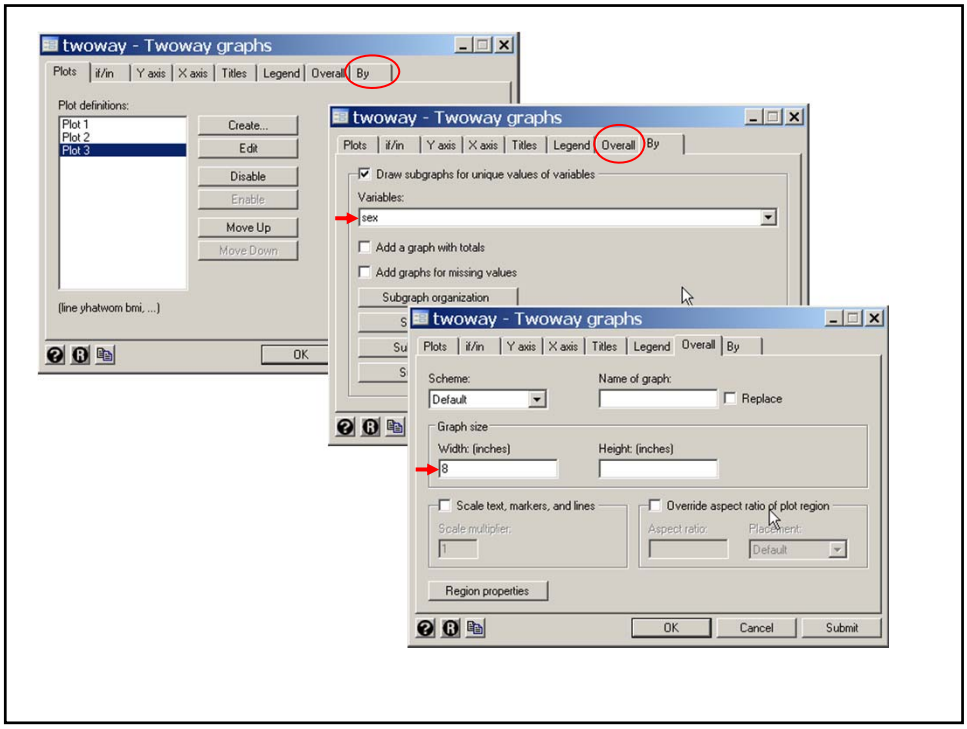
{4} **legend** controls the graph legend. **order** specifies the legend keys to be displayed and the labels assigned to these keys. The keys are indicated by the numbers 1, 2, 3, etc that are assigned in the order that they are defined.

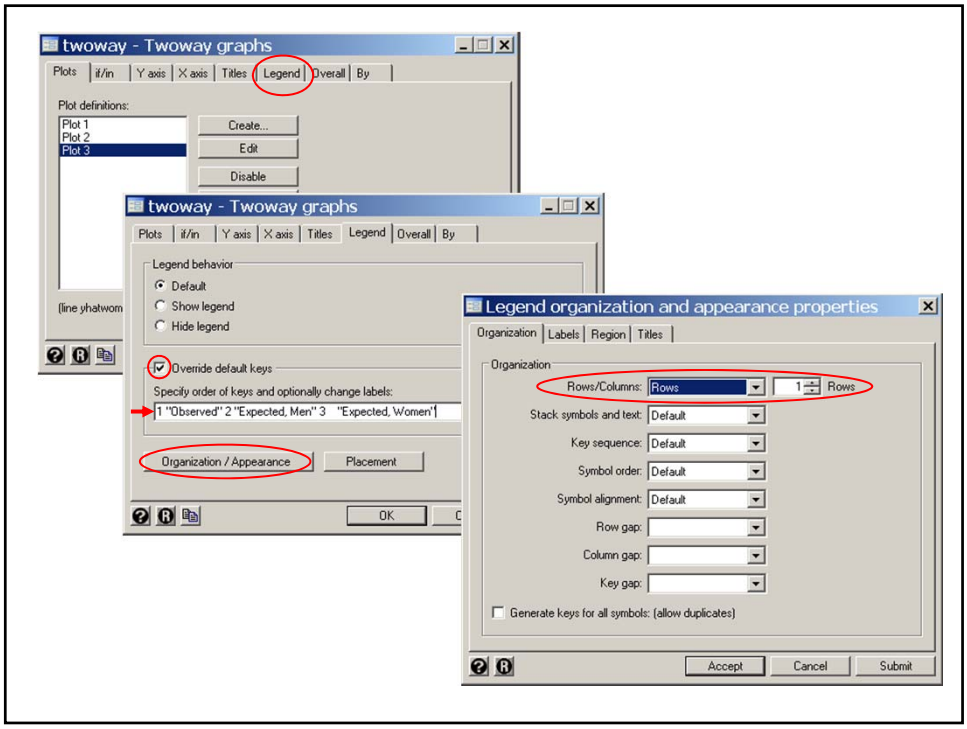
{5} **rows(1)** specifies that we want the legend displayed in a single row

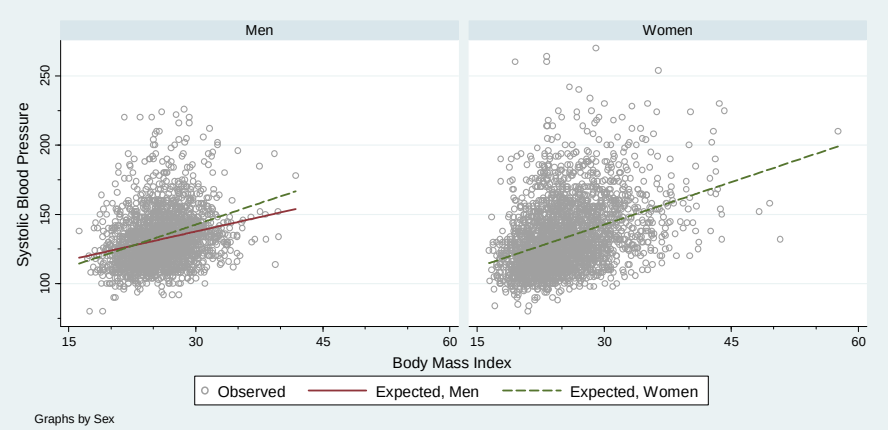



twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Create...
Plot 2
Plot
if/in
Choose a plot category and type
Basic plots
Range plots
Fit plots
Immediate plots
Advanced plots
Basic plots: (select type)
Scatter
Line
Connected
Plot type: (line plot)
Y variable:
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.