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.

© William D. Dupont, 2010, 2011
Use of this file is restricted by a Creative Commons Attribution Non-Commercial Share Alike license.
See http://creativecommons.org/about/licenses for details.

1. The Model

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

where

α, β_1, β_2, …, β_k are unknown parameters,

x_{i1}, x_{i2}, …, 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 ith patient.

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

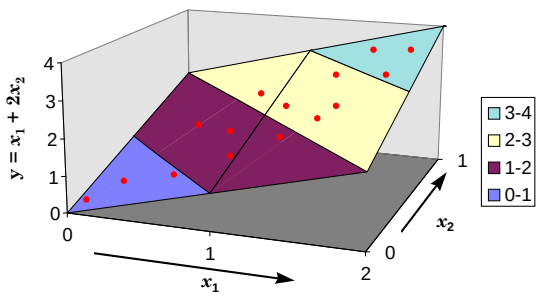
2. Reasons for Multiple Linear Regression

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 $x_1, x_2, \ldots, x_k$

3. Estimating Parameters

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

We estimate $a, b_1, \ldots, b_k$ 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

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

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

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

may be rewritten

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

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

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

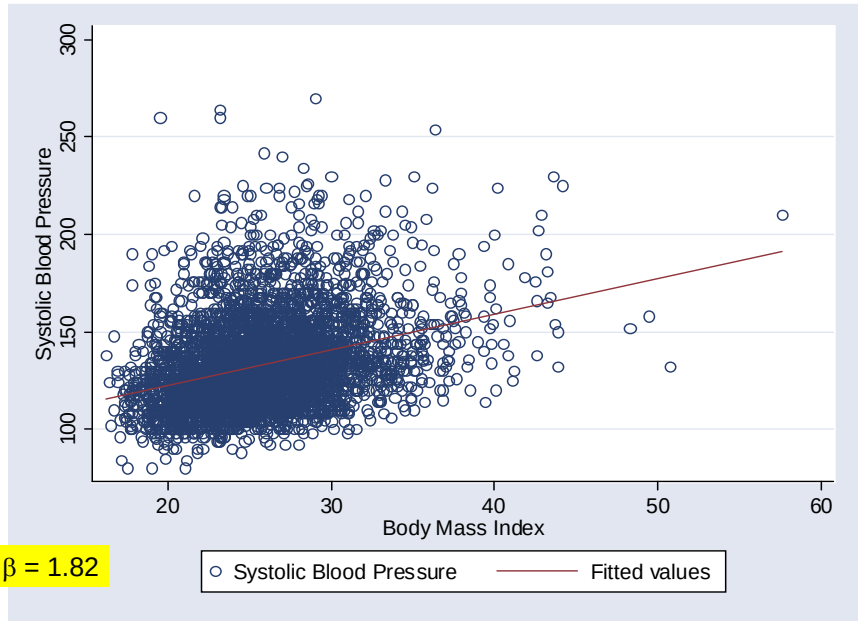
Follow-up information on coronary heart disease is also provided.

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.

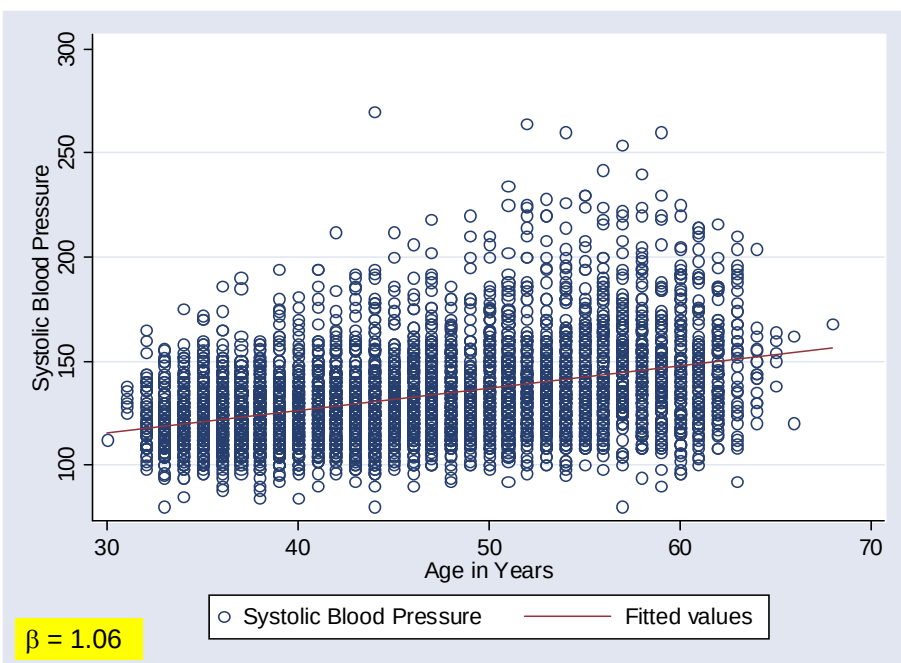
```
. * FramSBPbmiMulti.log
. *
. *  Framingham data set: Multiple regression analysis of the effect of bmi on
. *  sbp (Levy 1999).
. *
. use  "c:\WDDtext\2.20.Framingham.dta", clear

. 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)
```

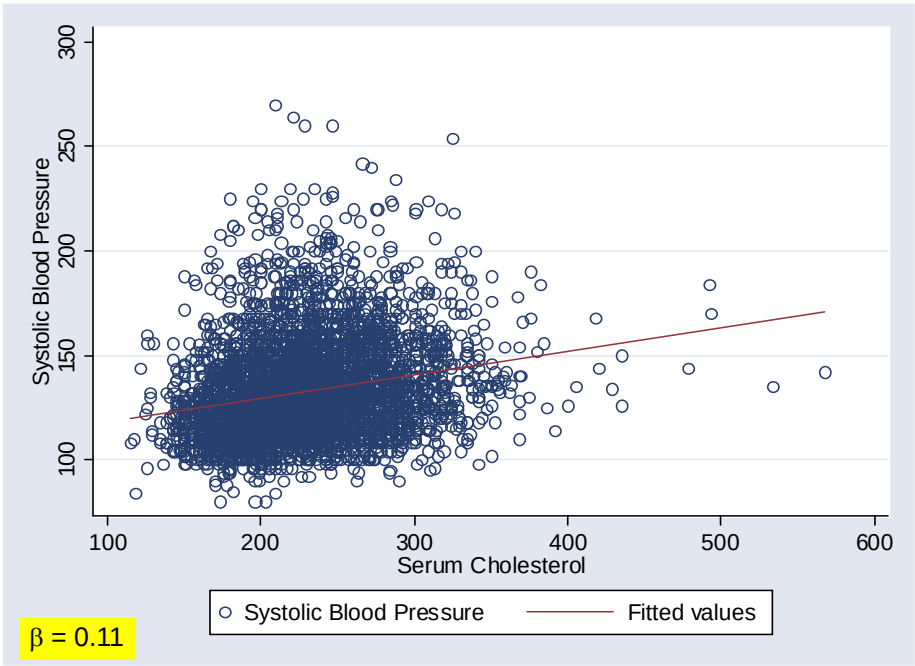


β = 1.82

```
. 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)
```

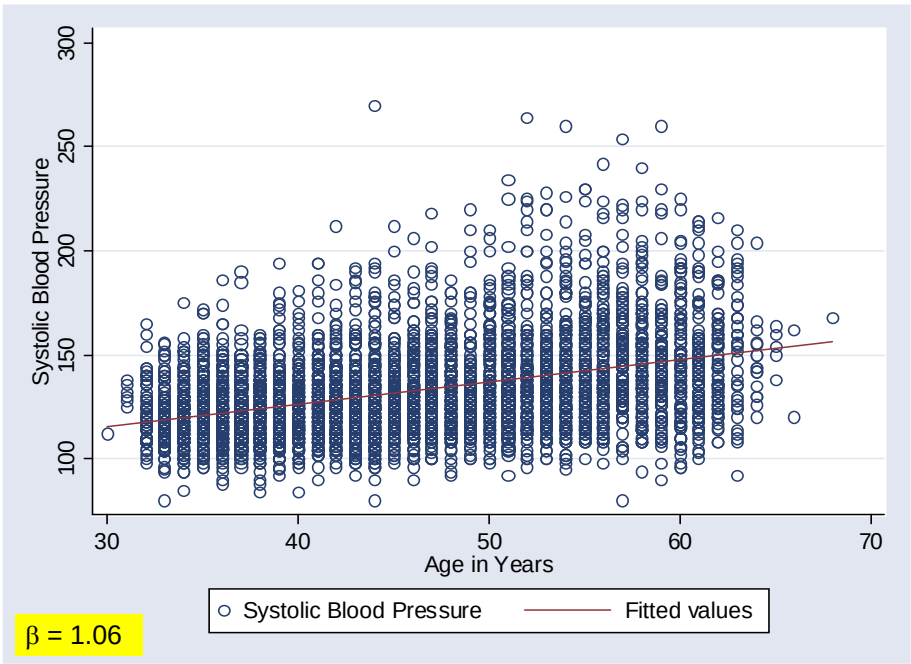


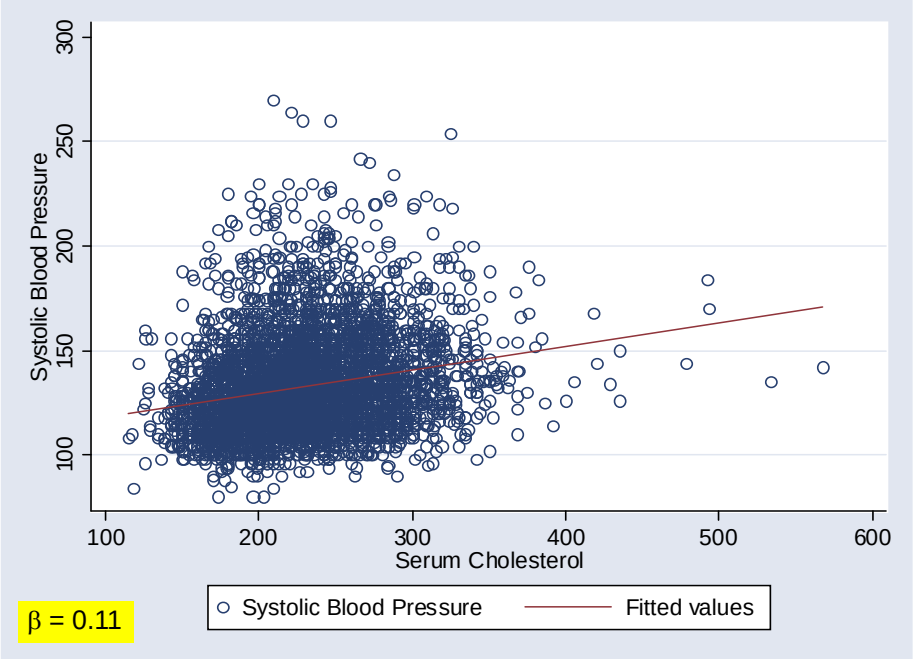
$\beta = 0.11$

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.

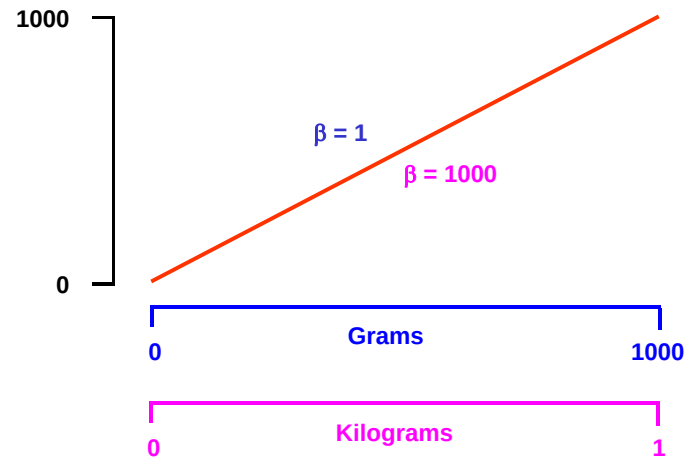
Note that the importance of a parameter depends not only on its magnitude but also on the range of the corresponding covariate. For example, the *scl* coefficient is only 0.11 as compared to 1.83 and 1.06 for *bmi* and *age*. However, the range of *scl* values is from 115 to 568 as compared to 16.2 - 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.

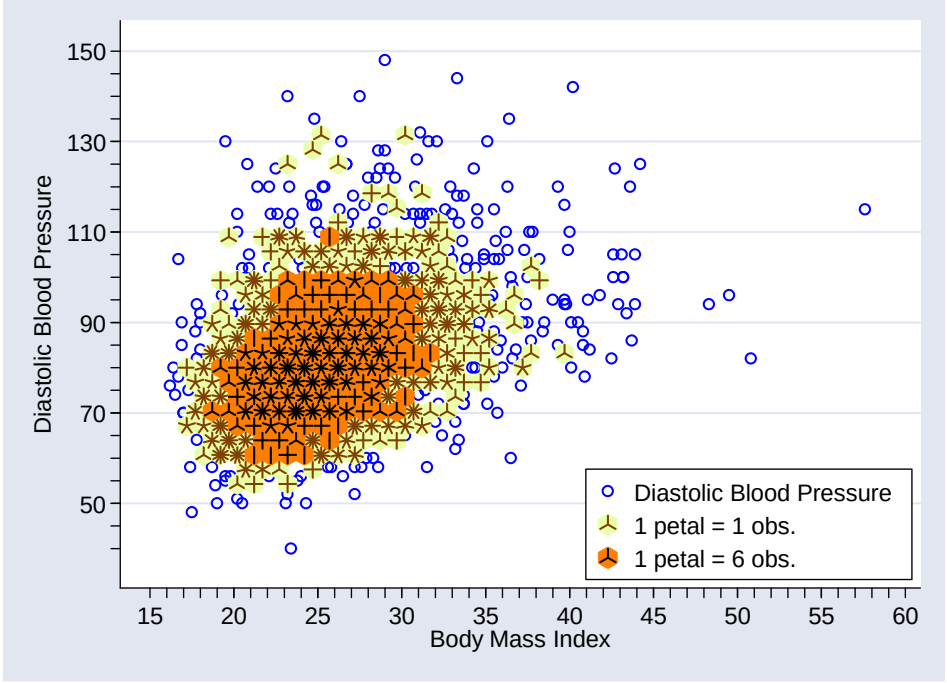
A **density distribution sunflower plot** is an attempt to provide a better sense of a bivariated distribution when observations are densely packed.

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.

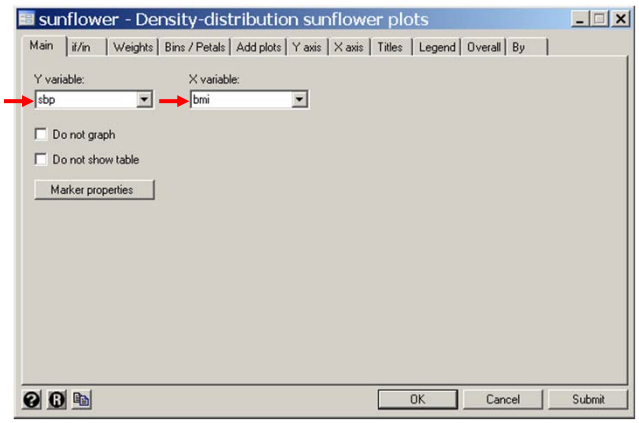
{3} The default bin height is given in units of y. It is chosen to make the bins regular hexagons on the graph.

{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

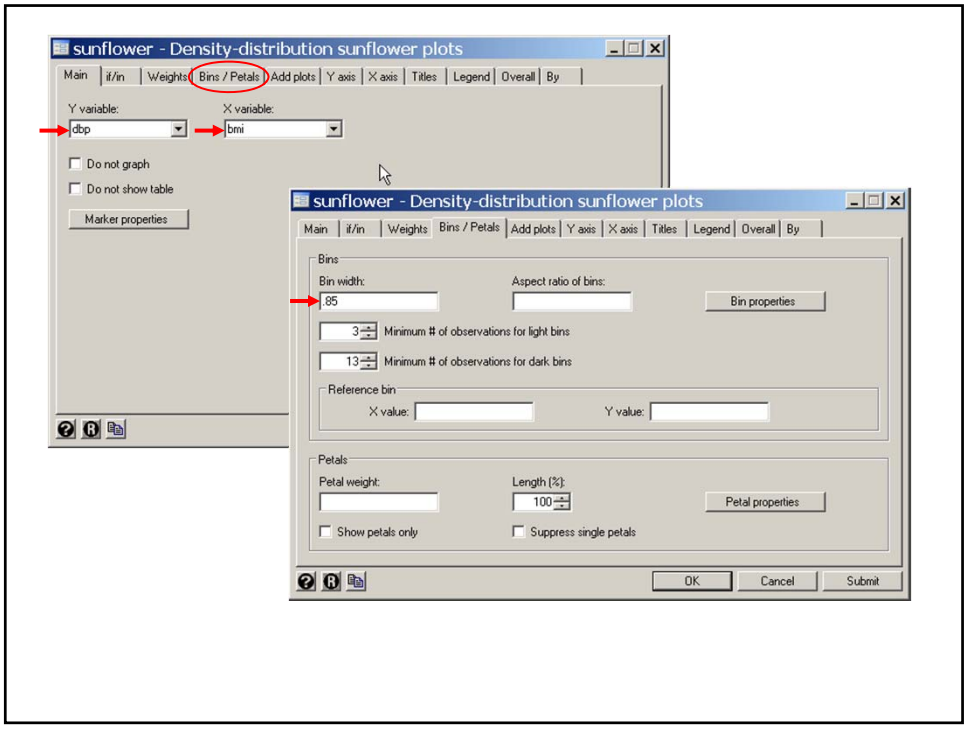

Systolic Blood Pressure
50
100
150
200
250
300
10
20
30
40
50
60
Body Mass Index
Systolic Blood Pressure
1 petal = 1 obs.
1 petal = 9 obs.

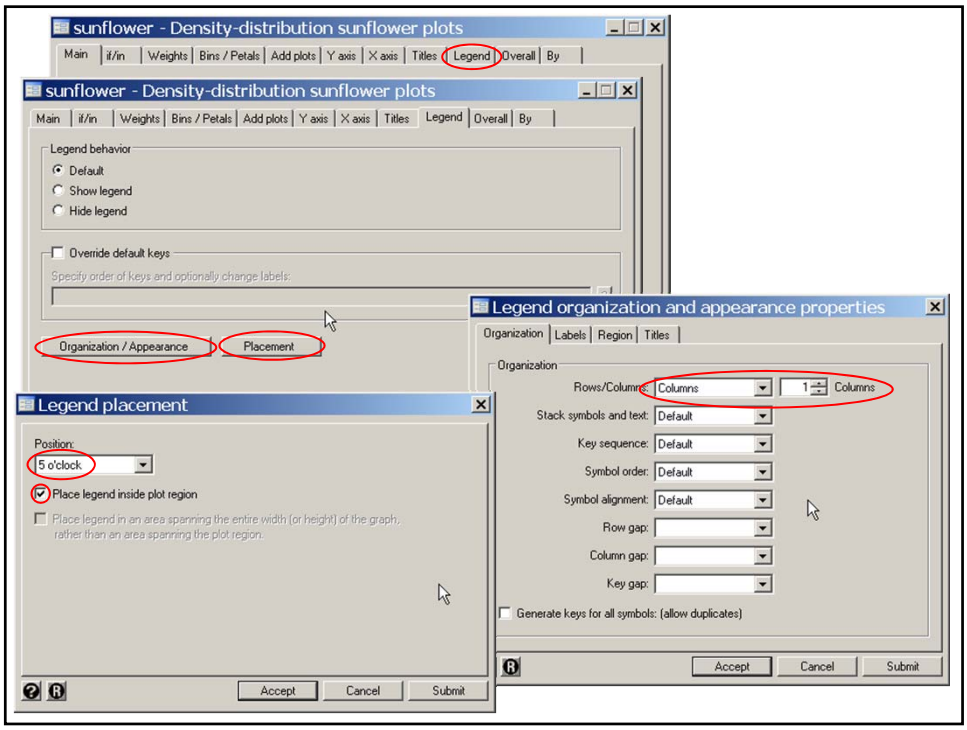
```
. more

. * 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
```

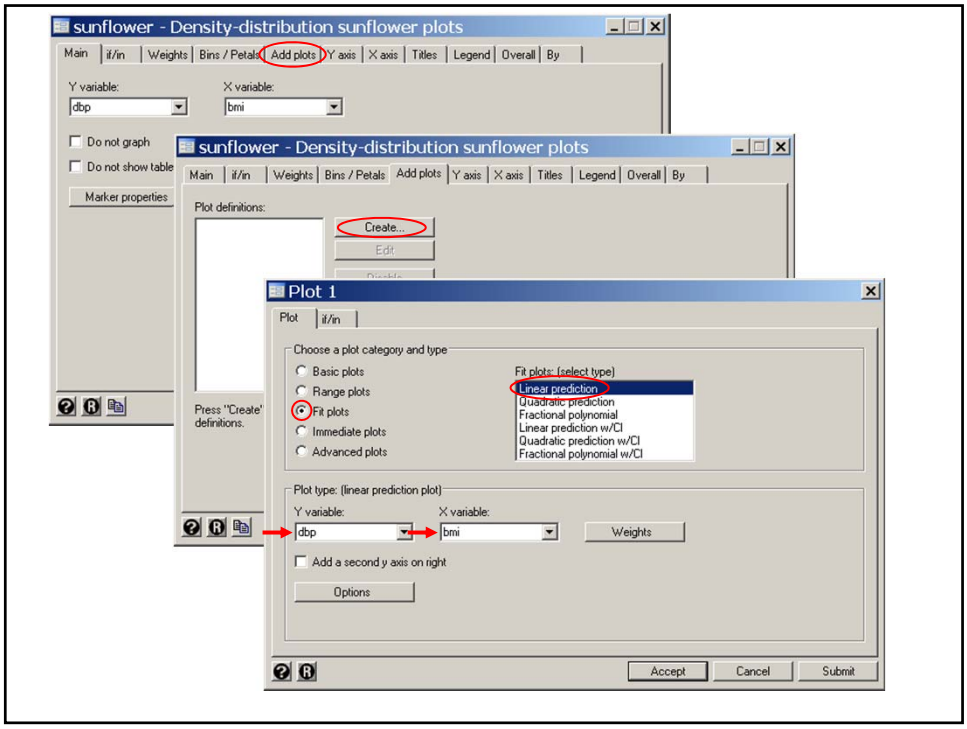
{1} *sunflower* accepts most standard graph options as well as special options that can control almost all aspects of the plot. Here *binwidth* specifies the bin width to be 0.85 kg/m^2.

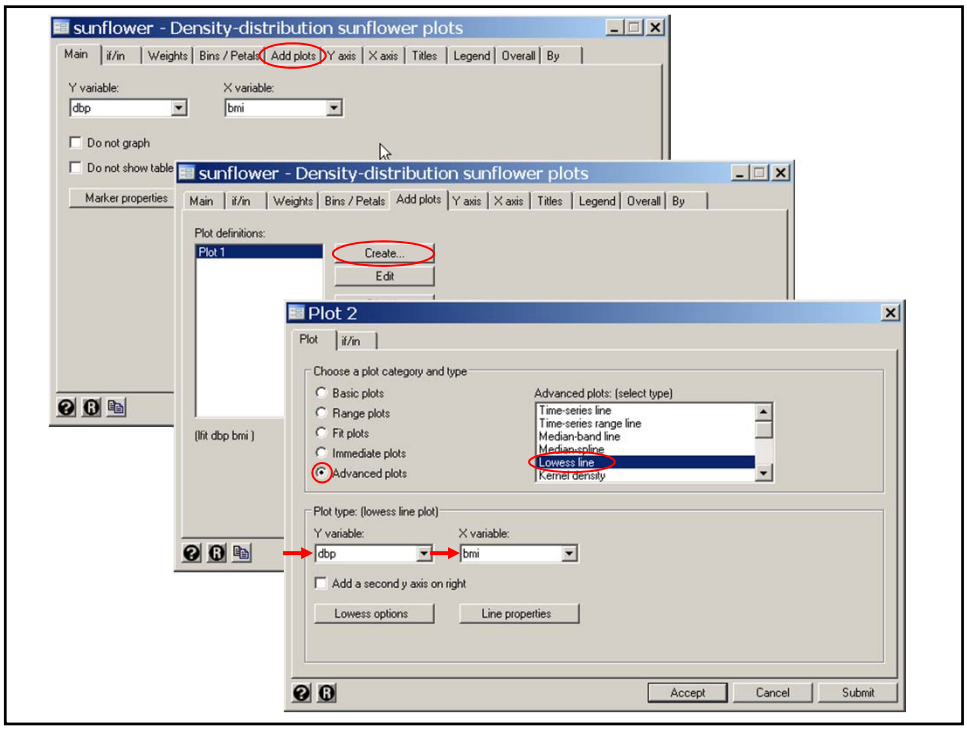
{2} The *position* sub-option of the legend option specifies that the legend will be located at 5 o'clock. *ring(0)* causes the legend to be drawn within the graph region. *cols(1)* requires that the legend keys be in a single column.

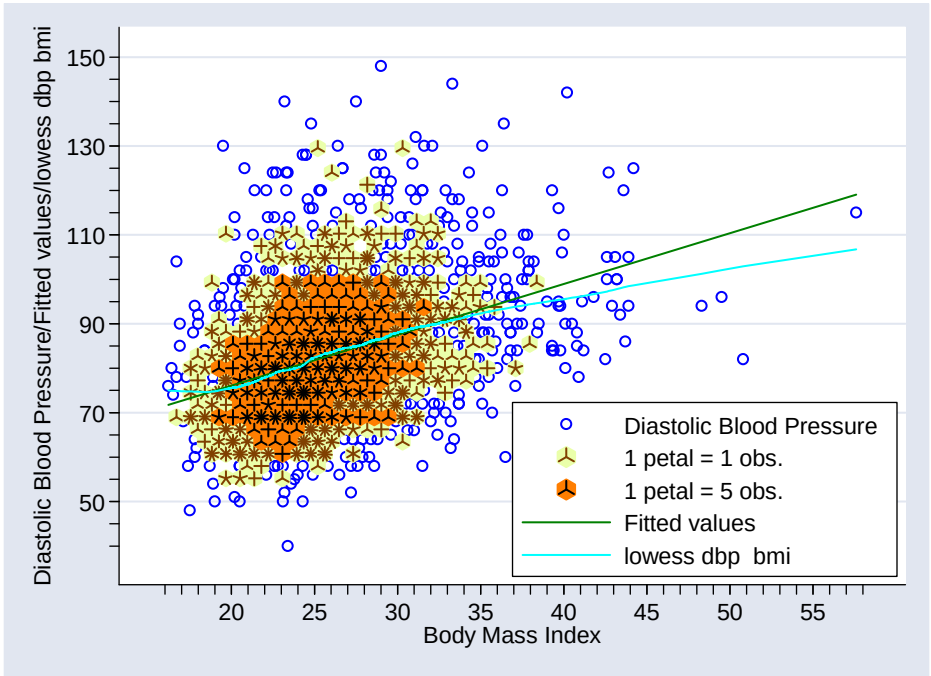
{3} The *addplot* option allows us to overlay other graphs on top of the sunflower plot. Here we draw the linear regression and lowess regression curves.


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 organization and appearance properties
Organization
Labels
Region
Titles
Rows/Columns:
Columns
1
Columns
Stack symbols and text:
Default
Key sequence:
Default
Symbol order:
Default
Symbol alignment:
Default
Row gap:
Column gap:
Key gap:
Generate keys for all symbols: (allow duplicates)
Accept
Cancel
Submit
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


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
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)
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
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)
Add a second y axis on right
Lowess options
Line properties
Accept
Cancel
Submit



8. Scatterplot matrix graphs

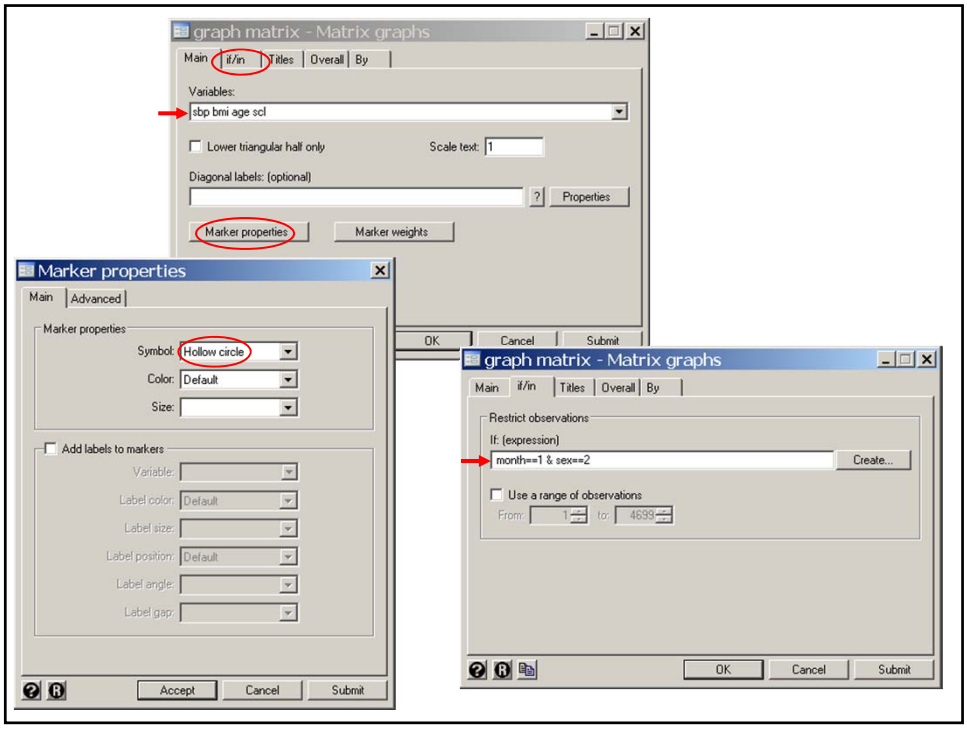
Another useful exploratory graphic is the scatter plot matrix. Here we look at the combined marginal effects of *sbp age bmi* and *scl*. The graph is restricted to women recruited in January to reduce the number of data points.

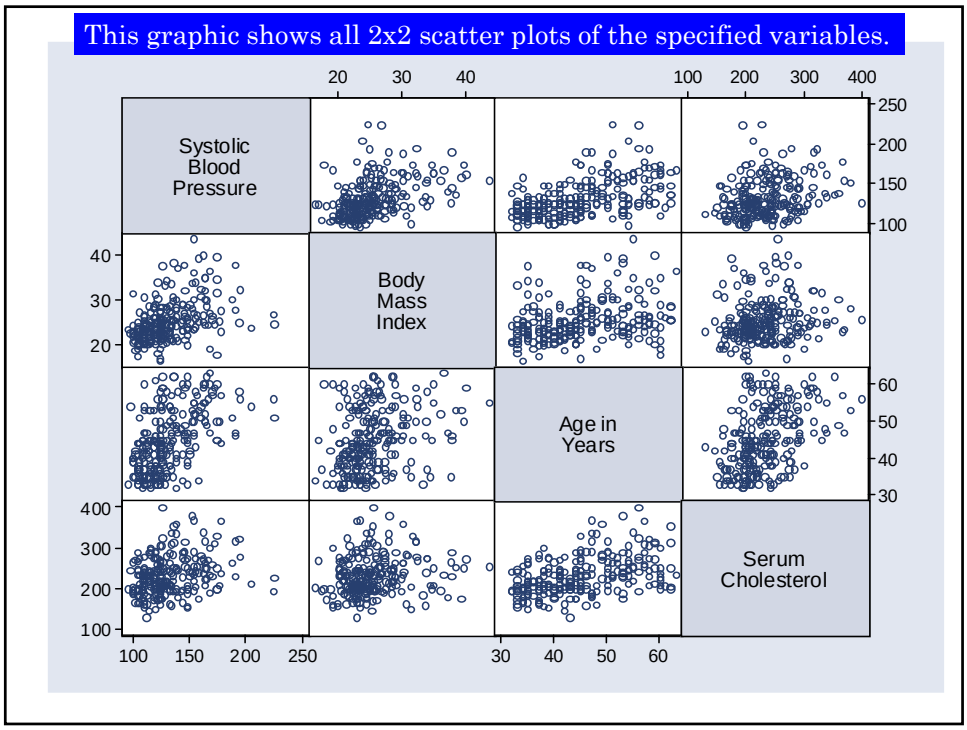
FramSBPbmiMulti.log continues as follows

```
. * 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
OK
Cancel
Submit
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
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
200
150
100
40
30
20
60
50
40
30
400
300
200
100
100
150
200
250
30
40
50
60

9. Modeling interaction in the Framingham baseline data

The first model that comes to mind is

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

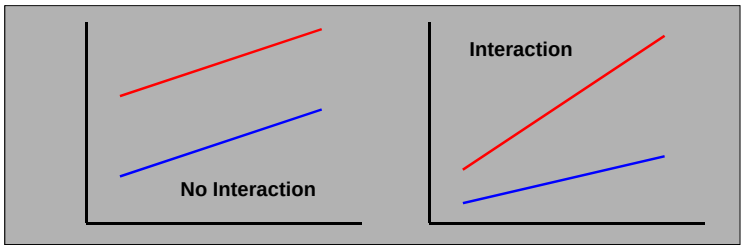
A potential weakness of this model is that it implies that the effects of the covariates on sbp_i are additive. To understand what this means, suppose we hold *age* and *scl* constant and look at *bmi* and *sex*. Then the model becomes

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.

$$\text{Then } women = \begin{cases} 1\text{: if subject is female} \\ 0\text{: if subject is male} \end{cases}$$

Consider the model

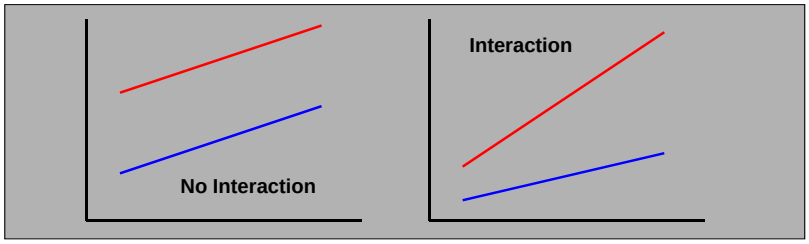
$sbp = \beta_1 + bmi \times \beta_2 + women \times \beta_3 + bmi \times women \times \beta_4$

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)
```

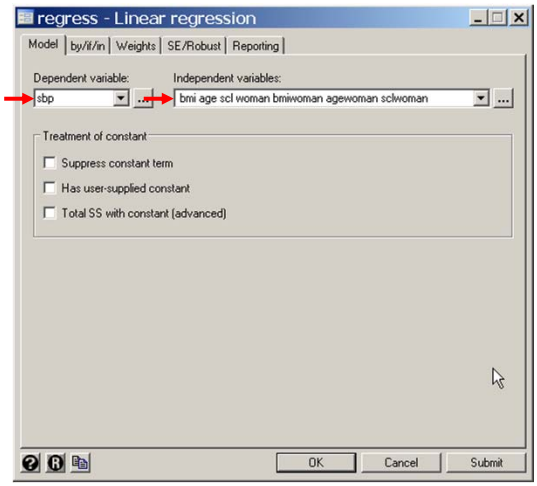
```
. regress sbp bmi age scl woman bmiwoman agewoman sclwoman

  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.

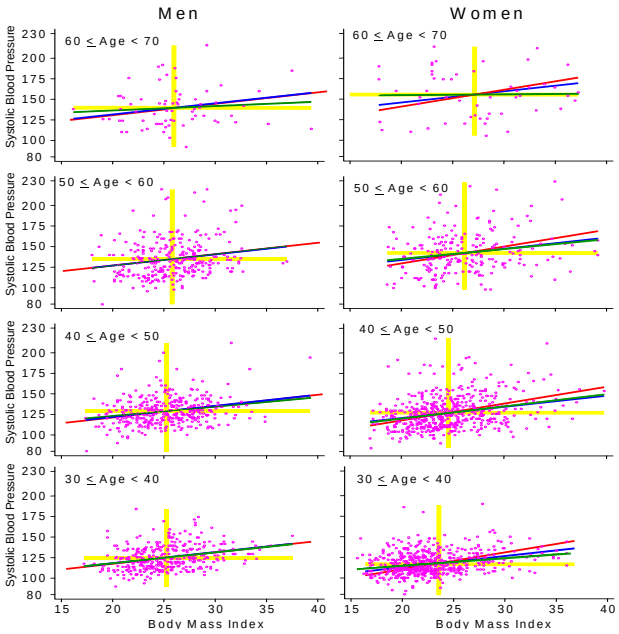
```
. regress sbp bmi age scl woman agewoman


  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.

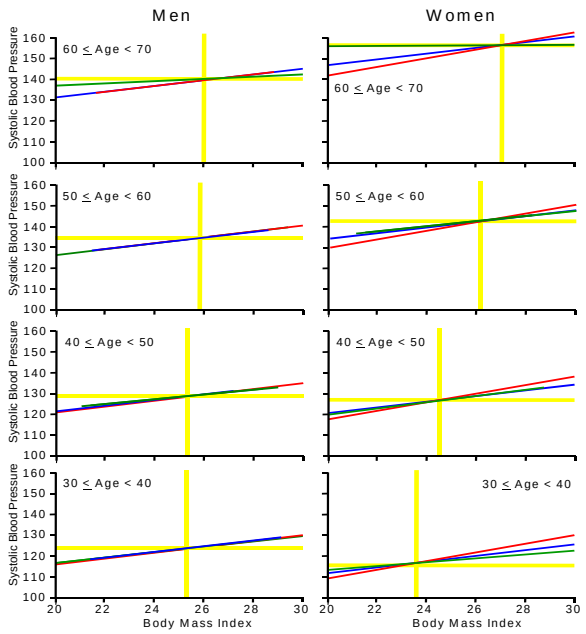


The **blue** lines have the slope from our **multiple regression model** of 1.36

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

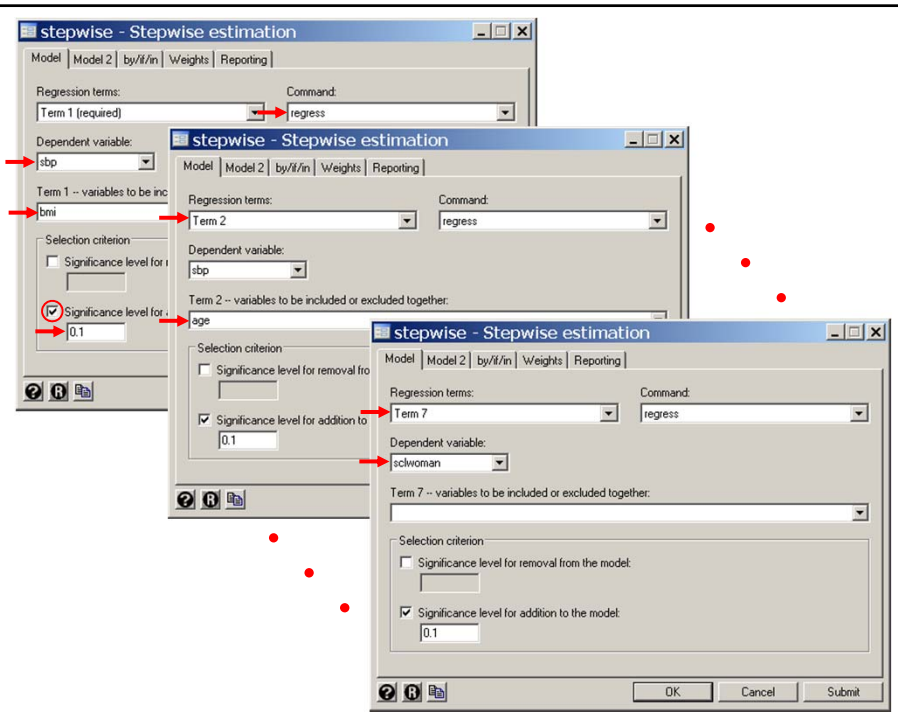
$sbp = \beta_1 + age \times \beta_2 + bmi \times \beta_3$

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

•

•

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



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

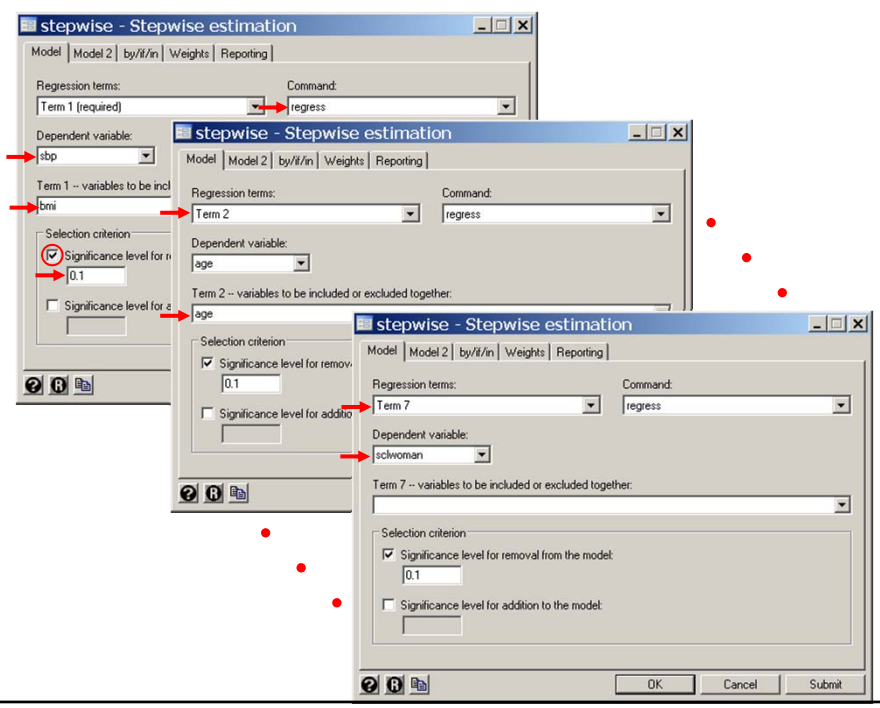
This method is similar to the forward method except that we start with all the variables and eliminate the variable with the least significance. The data is refit with the remaining variables and the process is repeated until all remaining variables have a significance level below some threshold.

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 ith 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 ith patient is a measure of her potential to influence the parameter estimates if the ith 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 $h_i = \text{var}\left(\hat{y}_i\right)/s^2$.

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

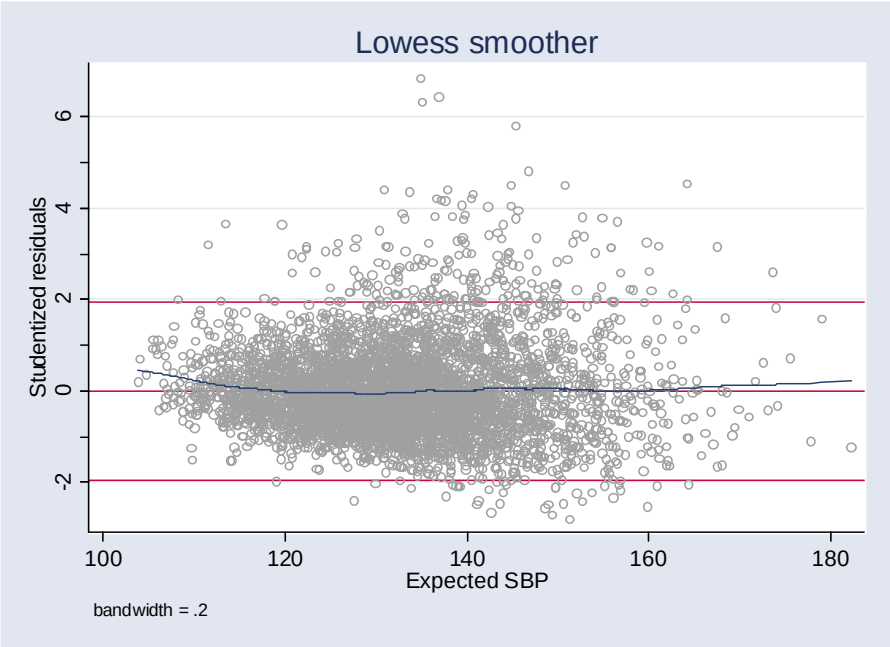
The standardized residual is $r_i = e_i / (s\sqrt{1-h_i})$ **{2.3}**

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 ith 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 $\pm$ 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.

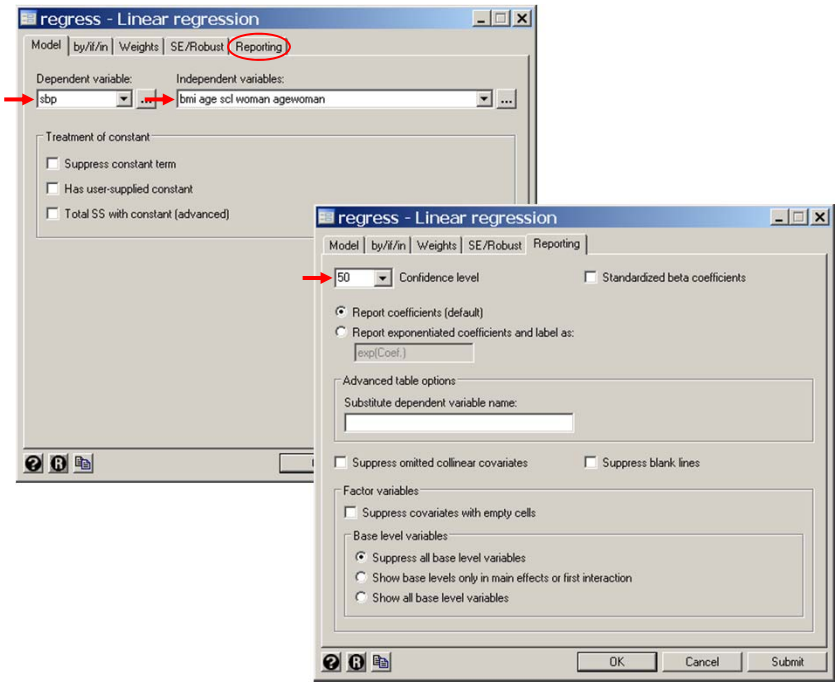
It is possible for a multivariate outlier to have a major effect on the parameter estimates but not be an obvious outlier on a 2×2 scatter plot.

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}
```

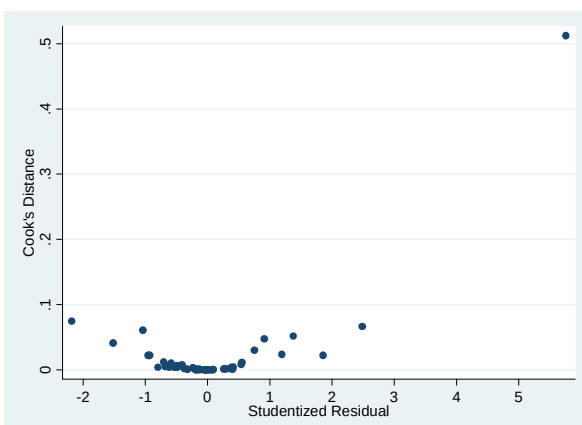
{1} The *level*(50) option specifies that 50% confidence intervals will be given for the parameter estimates.



```
  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 {3}
     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
------------------------------------------------------------------------------
```

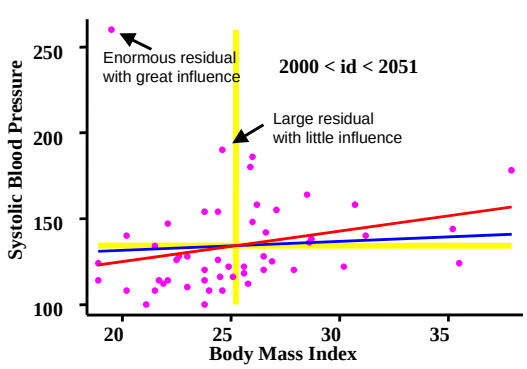
{1} The patient with the large Cook's D has ID 2049.

{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
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
------------------------------------------------------------------------------
```

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.

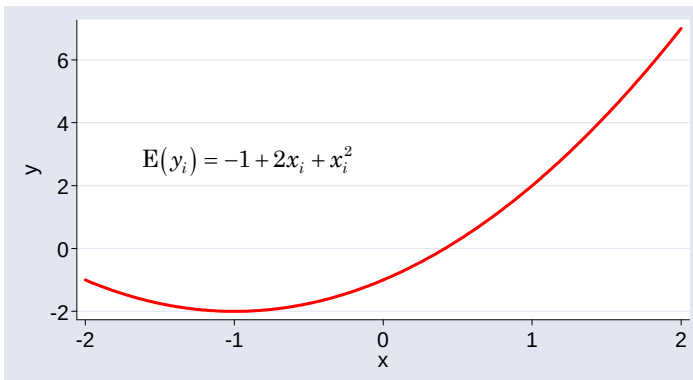
There are theoretical justifications for both methods. Neither is clearly better than the other.

18. Using Multiple Linear Regression for Non-linear Models

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

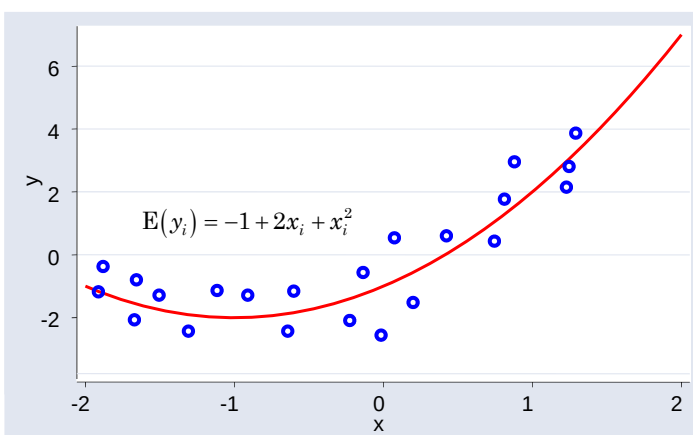
For example, suppose that there was a quadratic relationship between an independent variable x and the expected value of y. Then we could use the model

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



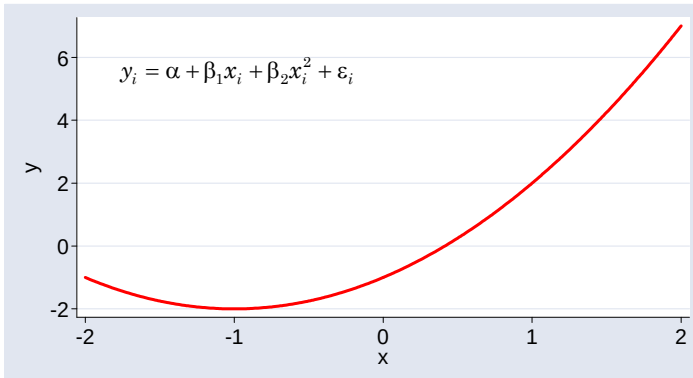
The preceding models $E(y_i)$ as a non-linear function of x_i. It is fine when correct but performs poorly for many non-linear models where the x-y relationship is not quadratic.

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

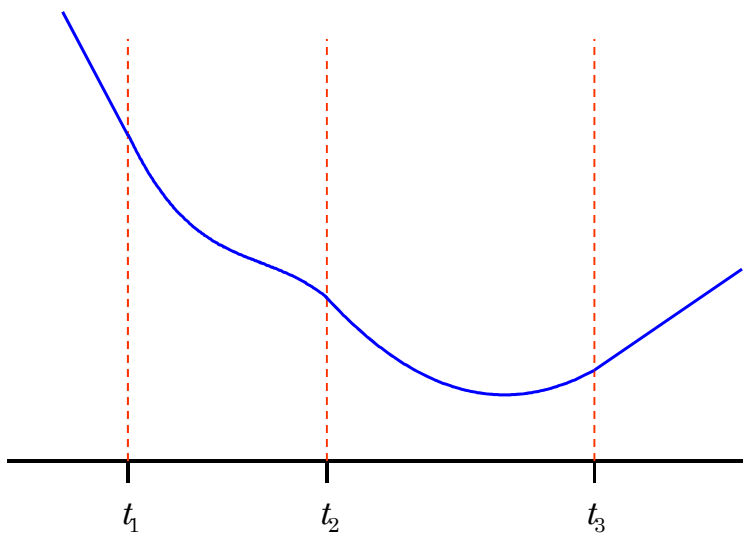
We wish to model y_i as a function of x_i using a flexible non-linear model. In a **restricted cubic spline model** we introduce k knots on the x-axis located at $t_1, t_2, \cdots, t_k$. We select a model of the expected value of y that

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.

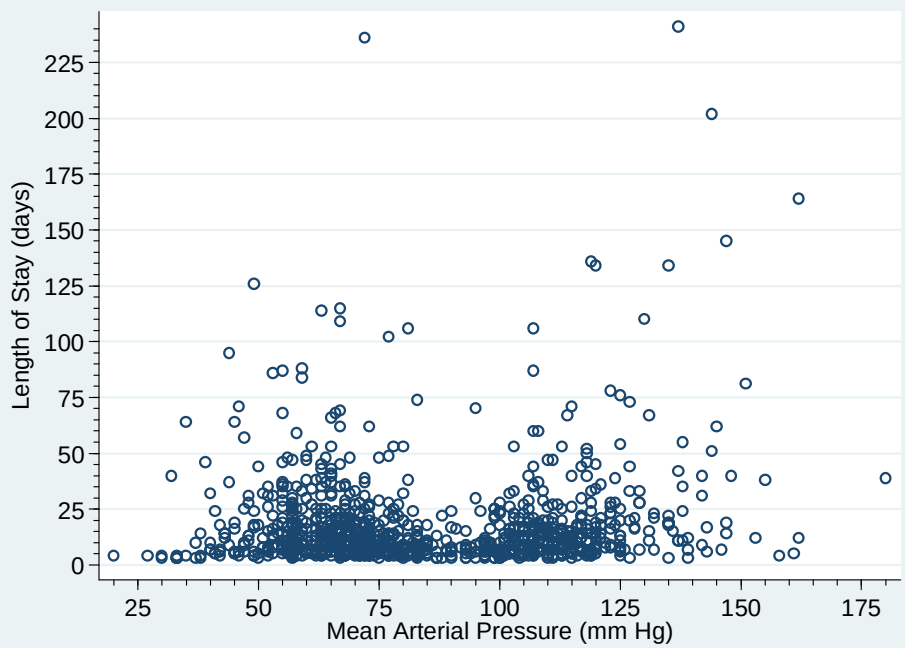
A random sample of data from 996 subjects in this study is available. See

3.25.2.SUPPORT.dta

`los` = length of stay in days.

`map` = baseline mean arterial pressure

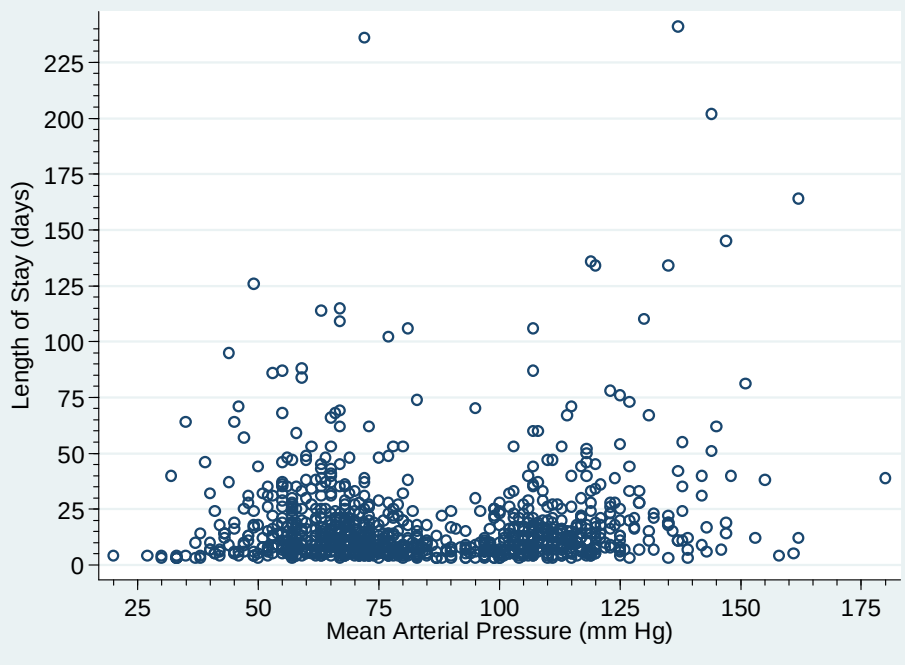
`fate` = $\begin{cases} 1: \text{ Patient died in hospital} \\ 0: \text{ Patient discharged alive} \end{cases}$



21. Fitting a Restricted Cubic Spline with Stata

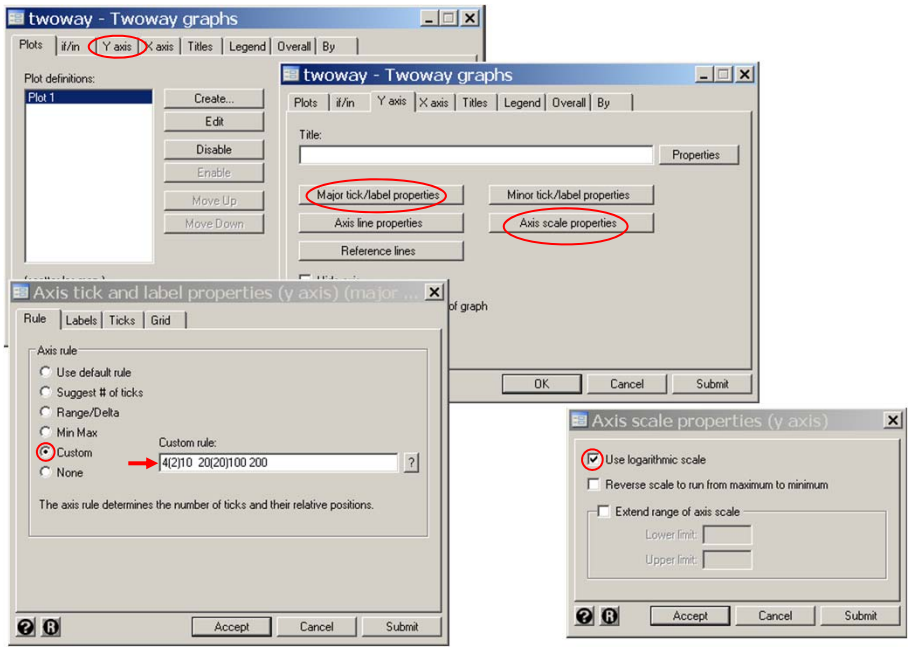
```
. *  SupportLinearRCS.log
. *
. *  Draw scatter plots of length-of-stay (LOS) by mean arterial
. *  pressure (MAP) and log LOS by MAP for the SUPPORT Study data
. *  (Lynn & Knauss, 1990).
. *
. use "C:\WDDtext\3.25.2.SUPPORT.dta" , replace

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

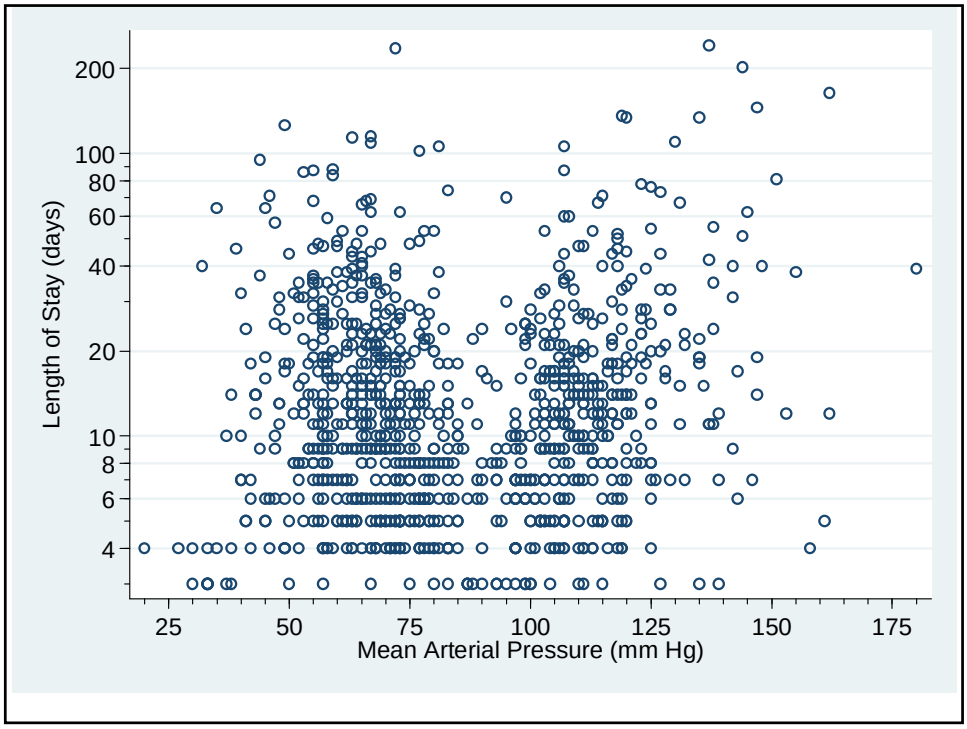
{1} Length of stay is highly skewed.



```
. 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.





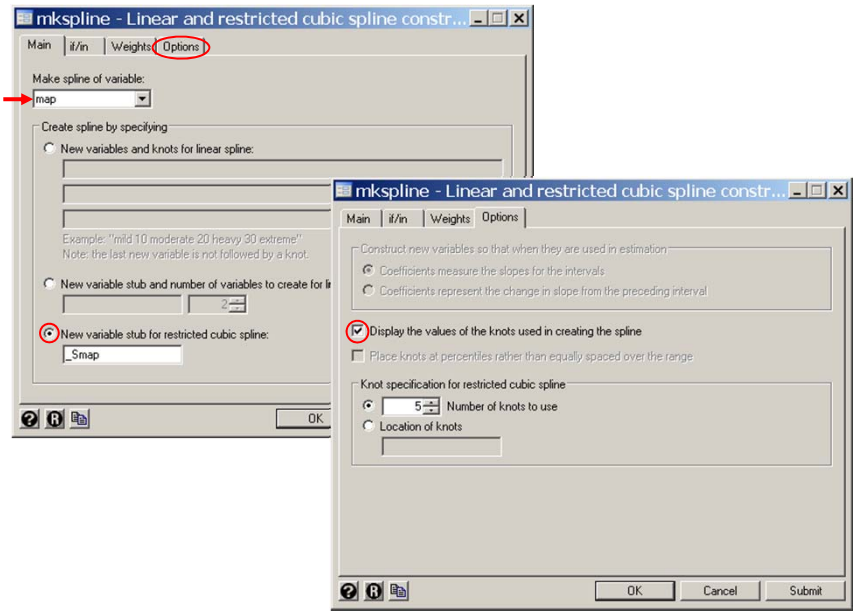

```
. *
. *  Regress log LOS against MAP using RCS models with
. *  5 knots at their default locations.  Overlay the expected
. *  log LOS from these models on a scatter plot of log LOS by MAP.
. *
. * Data > Create... > Other variable-creation... > linear and cubic...
mkspline _Smap = map, cubic displayknots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

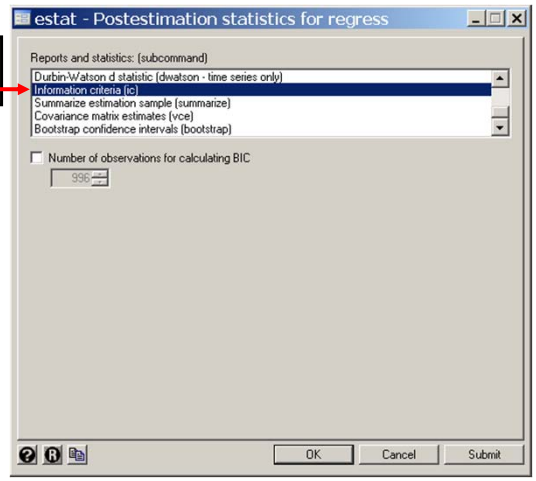
```
regress log_los _Smap1 _Smap2 _Smap3 _Smap4
```

```
. 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 $P < 0.00005$.

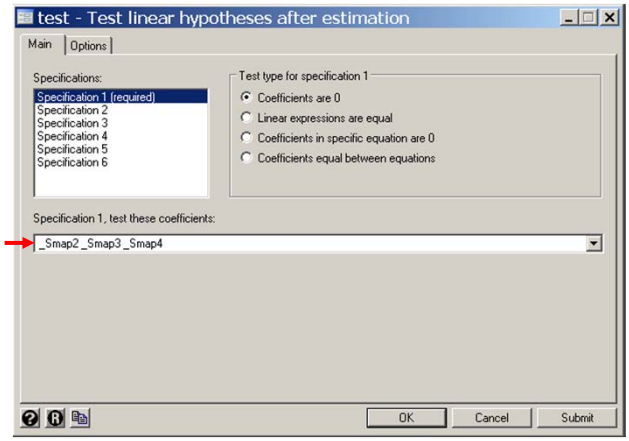
```
* Statistics > Postestimation > Reports and statistics
. estat ic                                                          {5}

-----------------------------------------------------------------------------
       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 $P < 0.00005$.



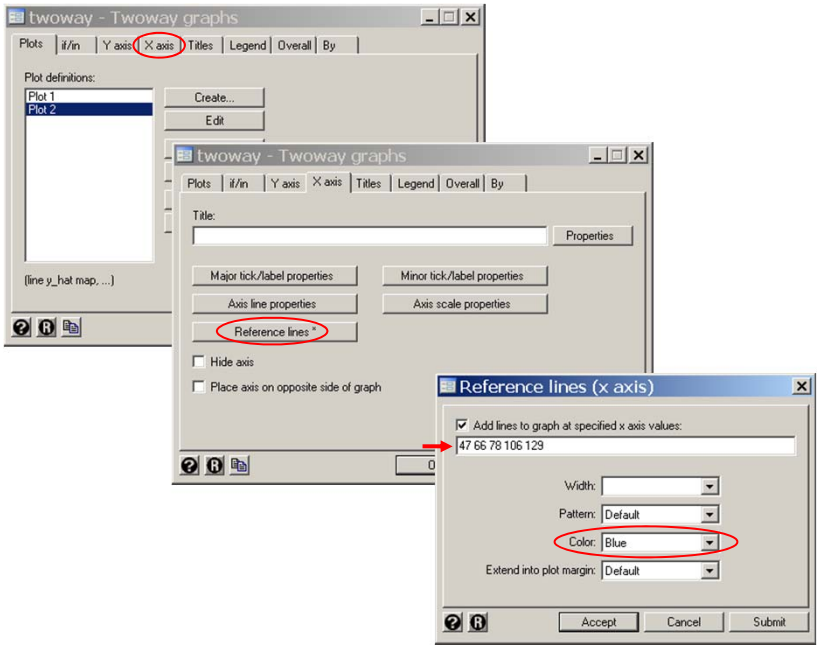
{7} **y_hat** is the estimated expected value of **log_los** under this model.

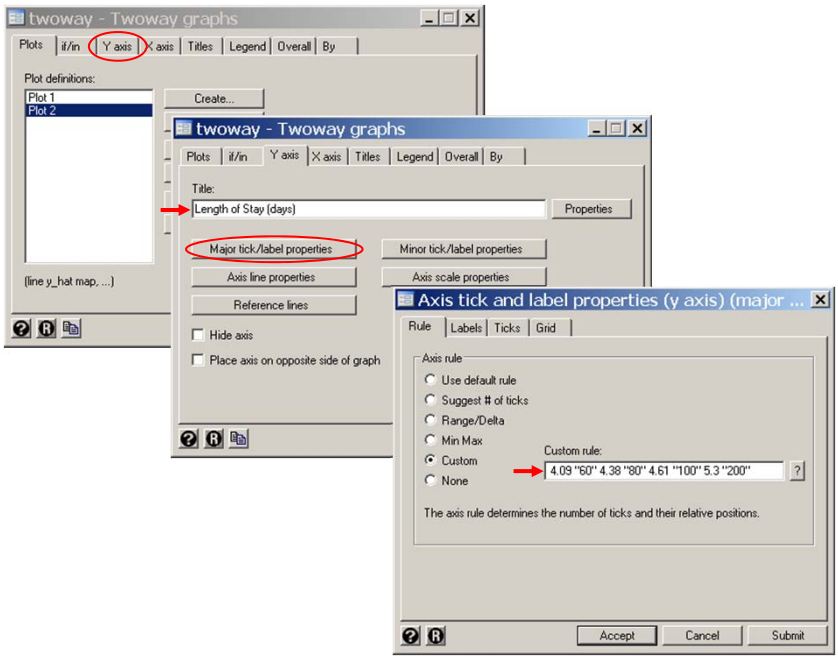
```
. predict y_hat5, xb                                                    {7}

. 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"))
```

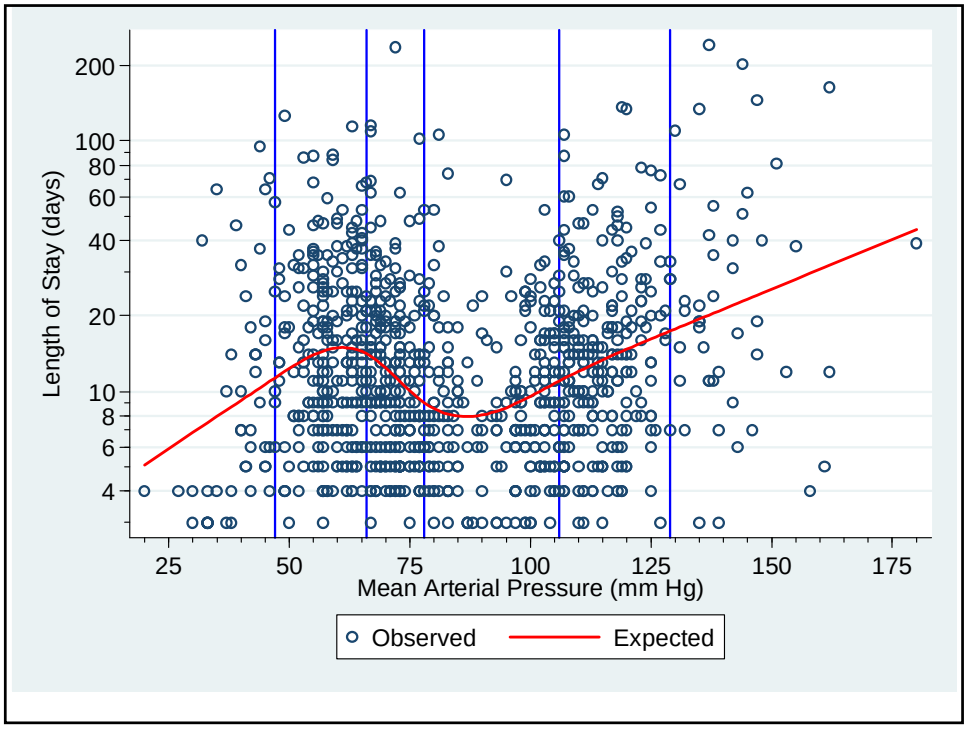
{8} Graph a scatterplot of **log_los** vs. **map** together with a line plot of the expected **log_los** vs. **map**.

{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.

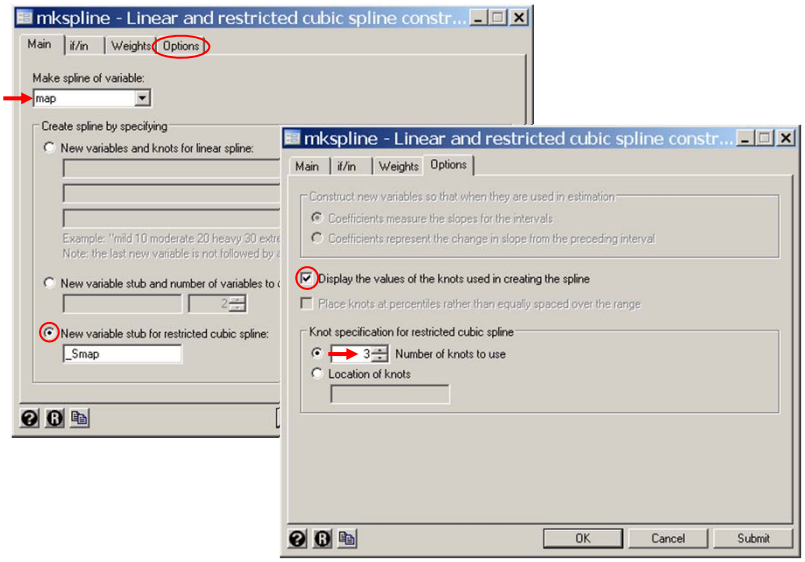

twoway - Twoway graphs
Plots
if/in
Y axis
X axis
Titles
Legend
Overall
By
Plot definitions:
Plot 1
Plot 2
Create...
Edit
(line y_hat map, ...)
Title:
Properties
Major tick/label properties
Minor tick/label properties
Axis line properties
Axis scale properties
Reference lines *
Hide axis
Place axis on 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
```

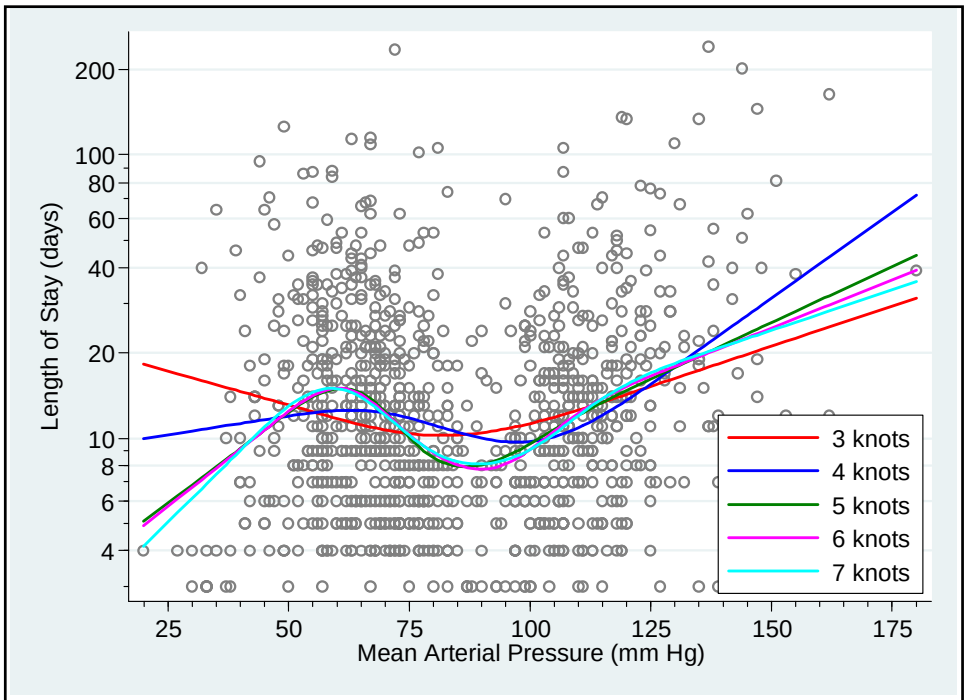
```
. drop _S*

. 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

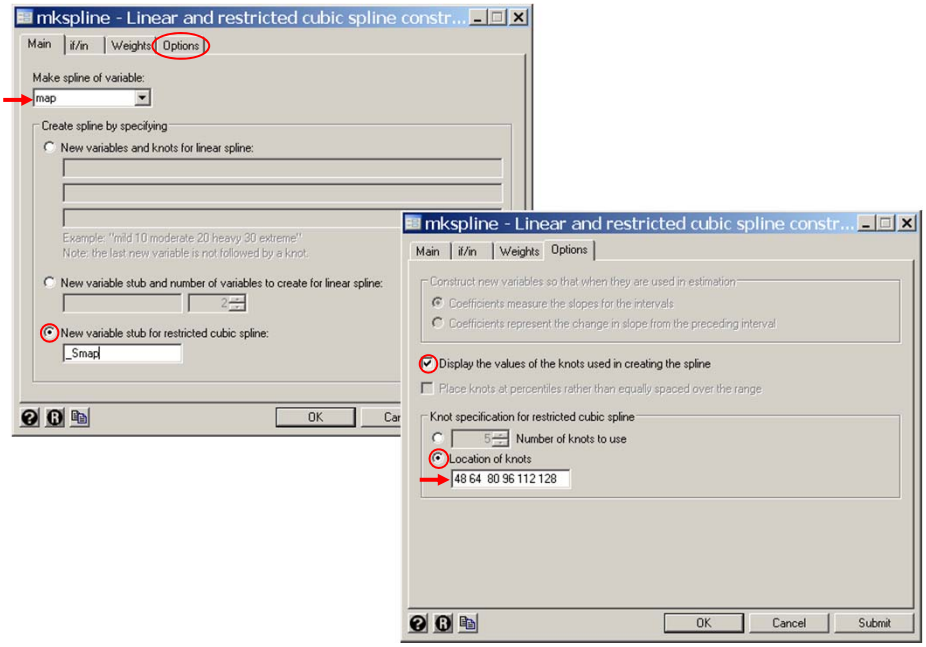
Knots	AIC	BIC
3	2,404.340	2,419.051
4	2,379.827	2,399.442
5	2,349.623	2,374.141
6	2,349.618	2,379.040
7	2,350.975	2,385.301

- 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



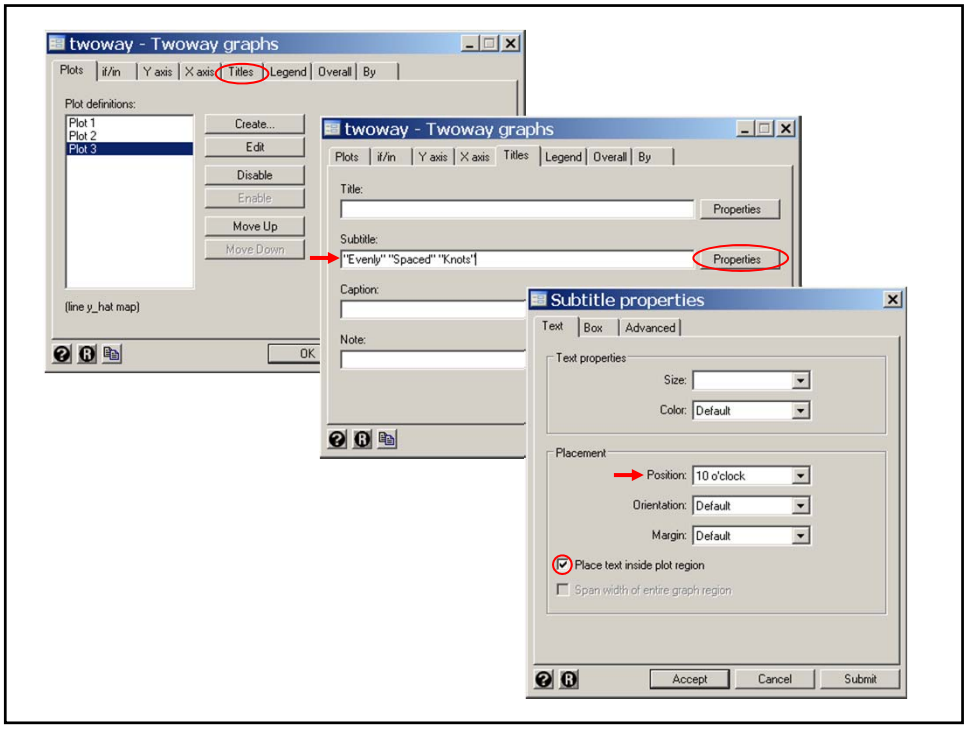

```
. regress log_los _S*

. predict y_hat, xb

. predict se, stdp

. generate lb = y_hat - invttail(_N-6, 0.025)*se                          {13}

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

. twoway rarea lb ub map , color(yellow)                                 /// {14}
>     || 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))        /// {15}
>       ytitle(Length of Stay (days)) legend(off)
```

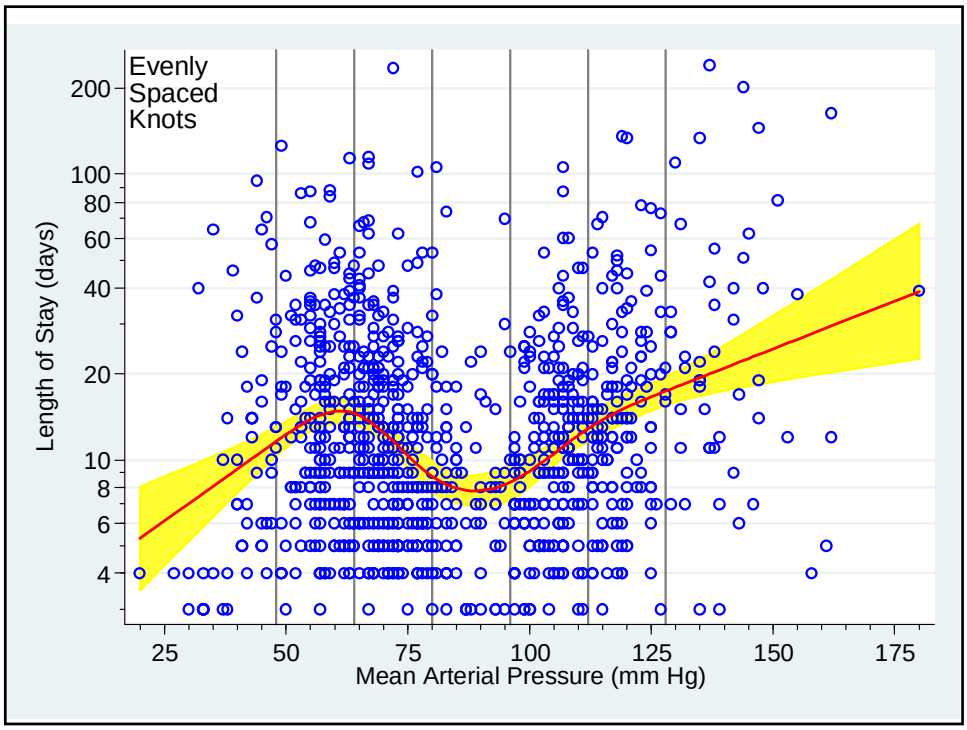
{Output Omitted}

{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.

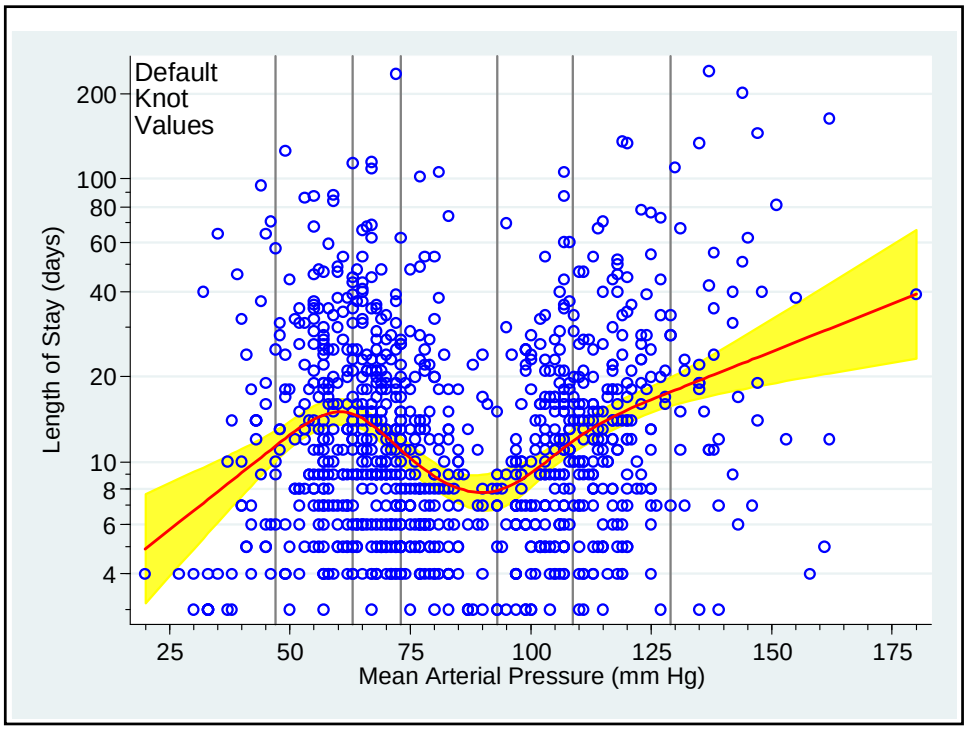
{13} _N-6 = the number of observations minus the number of parameters = 990 = the degrees of freedom of the MSE s^2. `lb` is the lower bound of the 95% confidence interval for `y_hat`.

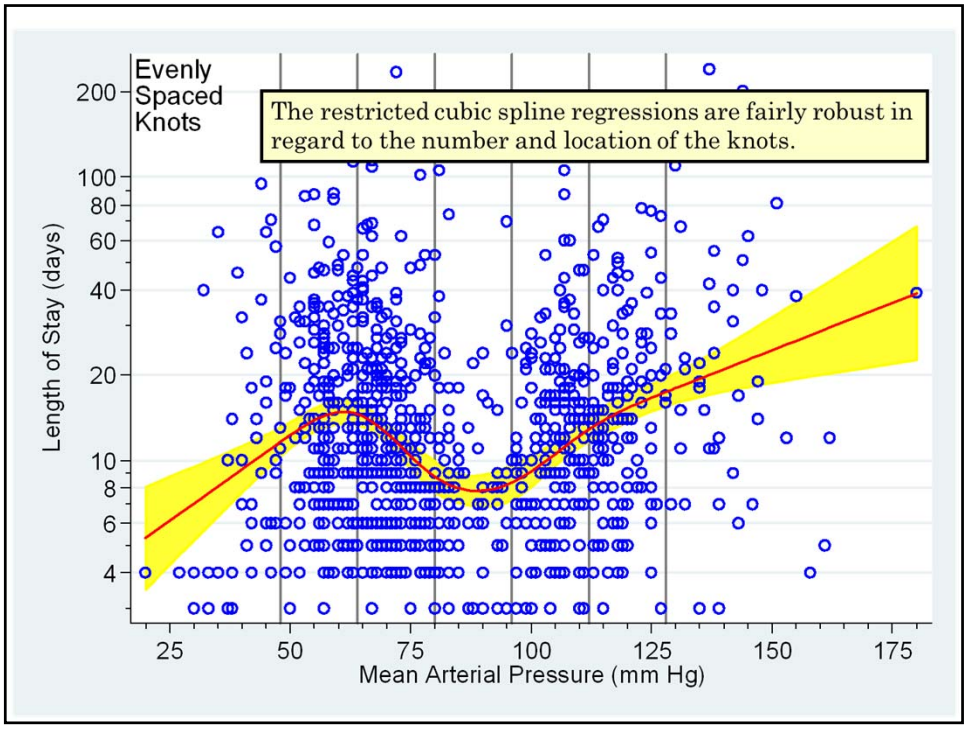
{14} This plot adds the 95% confidence region for the regression curve.


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)

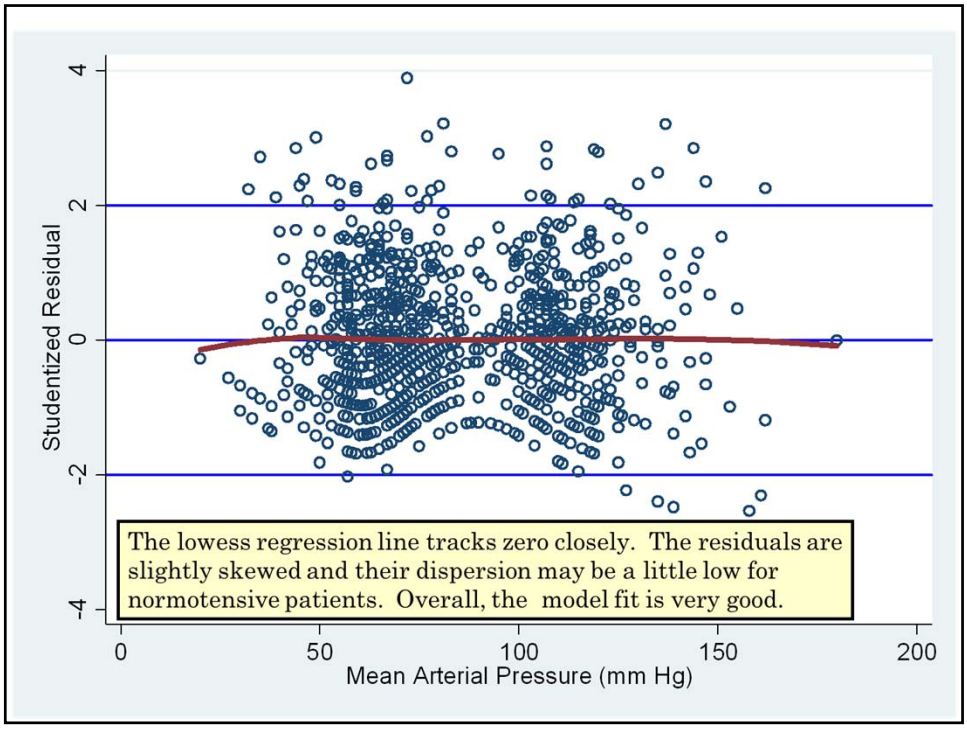
```
. *
. *  Replot 6 knot model with default knot spacing.
. *
. drop _S* y_hat se lb ub

. 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.

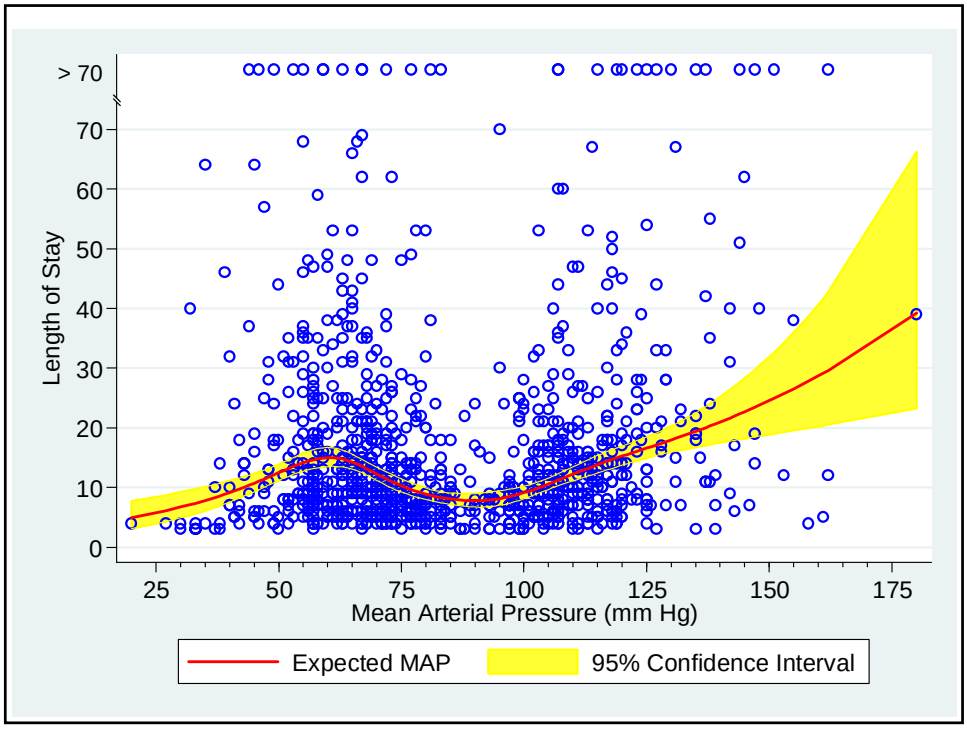


The lowess regression line tracks zero closely. The residuals are slightly skewed and their dispersion may be a little low for normotensive patients. Overall, the model fit is very good.

```
. *
. *  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.