



United International College

The introduction of SPSS

(Statistical Product and Service solutions)



Basic Knowledge
in Statistics and SPSS

Zhou Yongdao

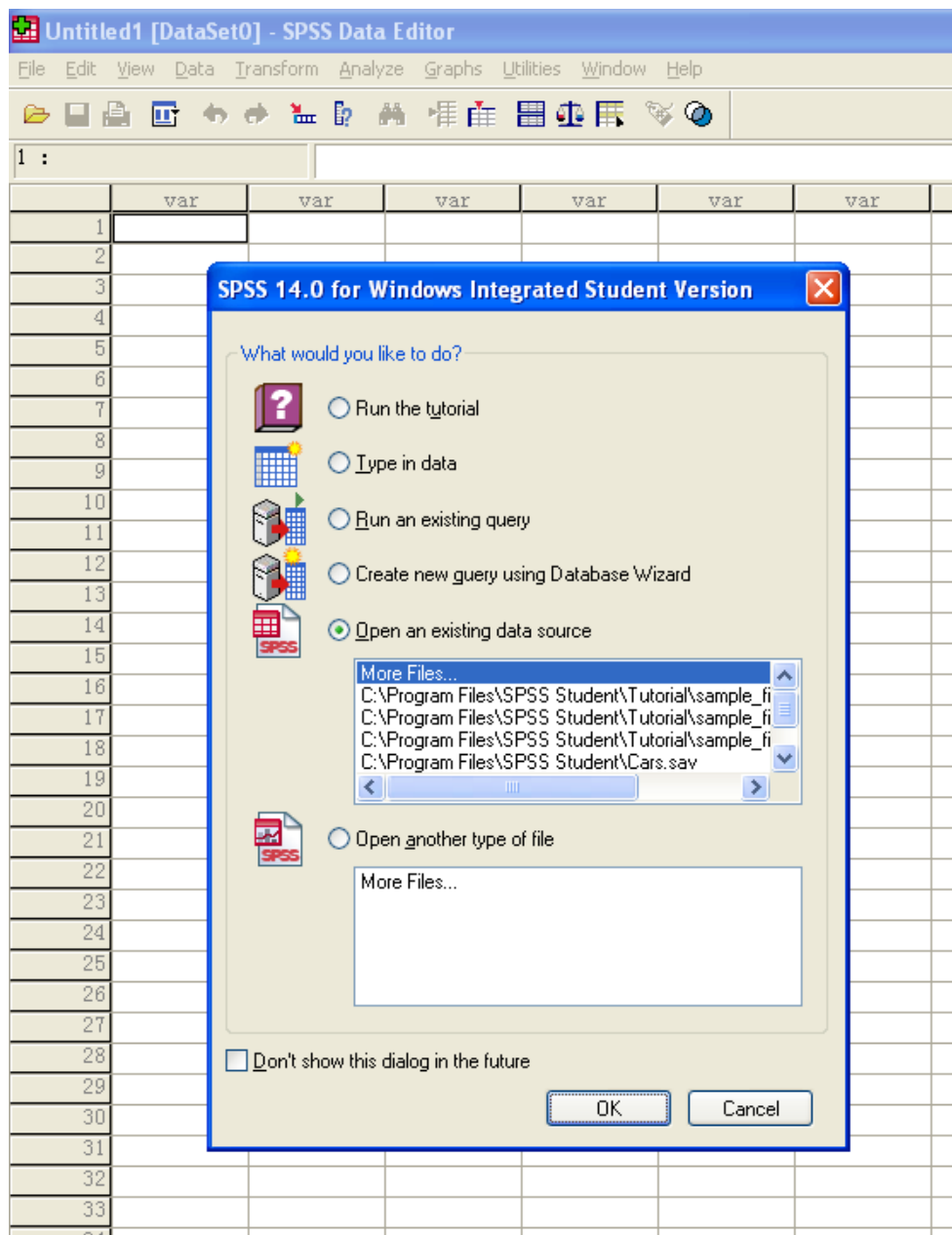
Contents

OVERVIEW:	3
DATA:	5
FREQUENCIES:	7
GRAPH:	9
• Simple bar chart	9
• Pie chart	12
• Boxplot	14
• Scatter/Dot	16
• Histogram	18
• Q-Q	20
REGRESSION ANALYSIS	22
• Method 1	22
• Method 2	26
• Method 3	35

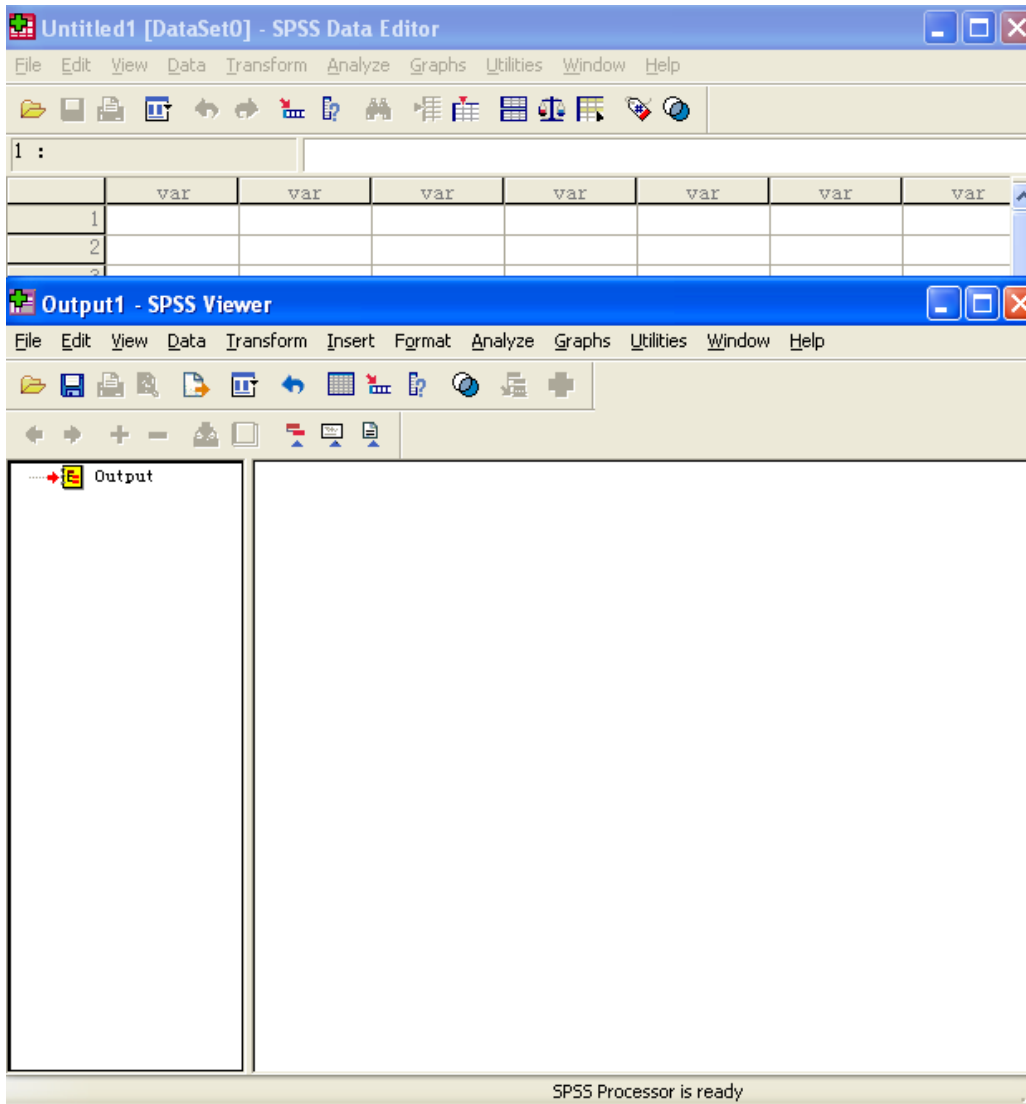
Overview:

SPSS for Windows provides a powerful statistical-analysis and data-management system in a graphical environment, using descriptive menus and simple dialog boxes to do most of the work for you. Most tasks can be accomplished simply by pointing and clicking the mouse.

Open the software; you can get this interface, click “cancel”.



When you start a session, you see the Data Editor window and Output Viewer window.



Many of the tasks that you want to perform with SPSS are available through menu selections. Each window in SPSS has its own menu bar with menu selections that are appropriate for that window type.

Data:

SPSS data files, which have a .sav file extension, contain your saved data. To open data, we can do:

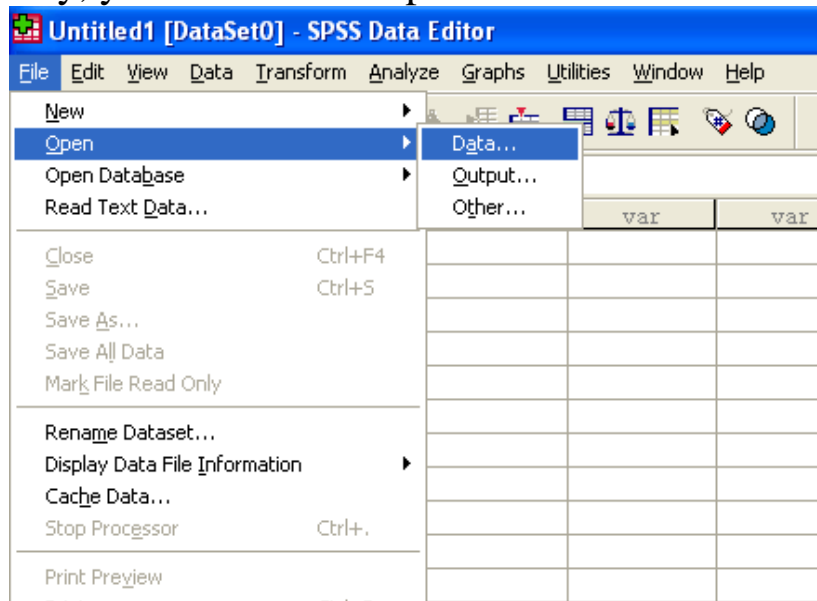
- From the menus choose:

File

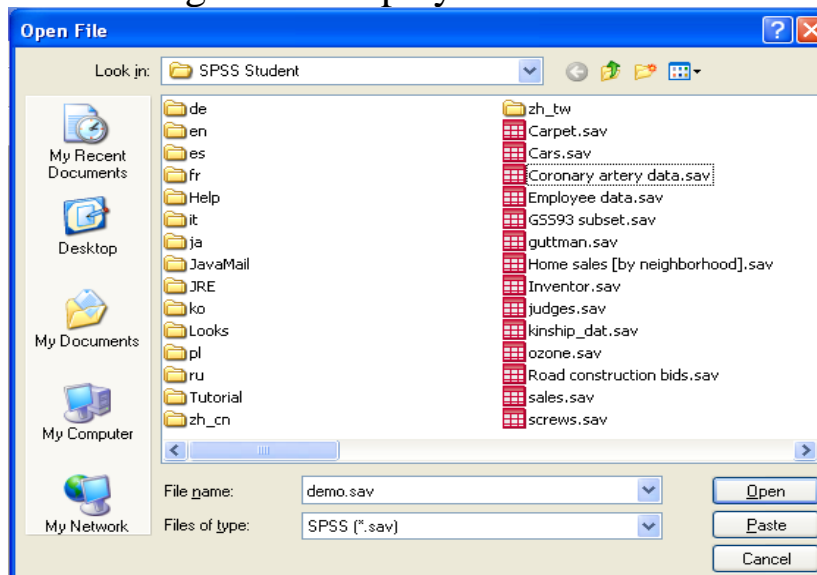
Open

Data...

Alternatively, you can use the Open File button on the toolbar.



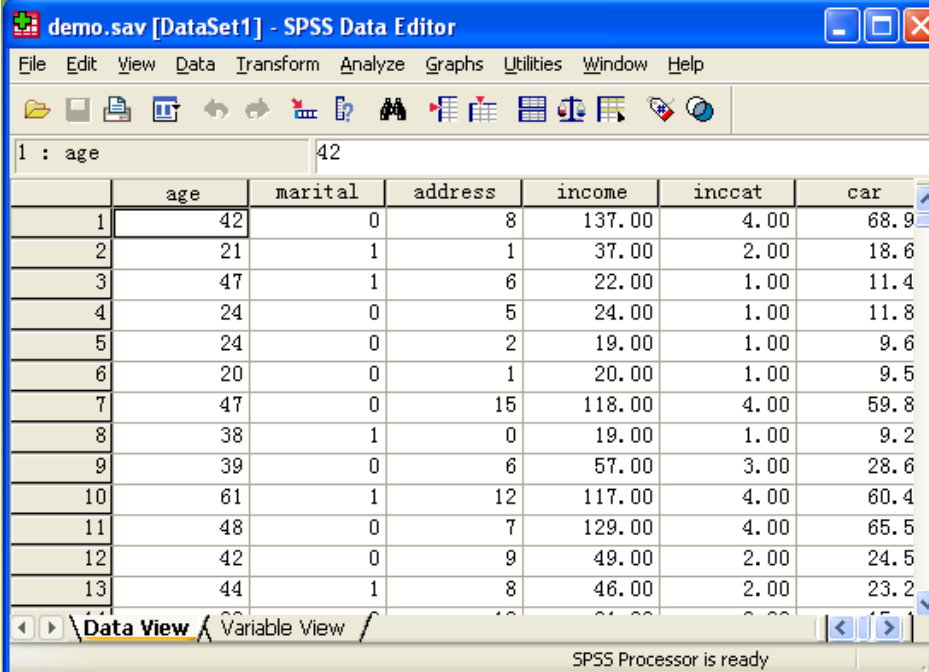
The Open File dialog box is displayed.



- Select a data file and click Open.

If you want to open the data file used in this example, demo.sav is located in tutorial\sample_files in the directory in which SPSS is installed.

The data file is displayed in the Data Editor.



demo.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : age 42

	age	marital	address	income	inccat	car
1	42	0	8	137.00	4.00	68.9
2	21	1	1	37.00	2.00	18.6
3	47	1	6	22.00	1.00	11.4
4	24	0	5	24.00	1.00	11.8
5	24	0	2	19.00	1.00	9.6
6	20	0	1	20.00	1.00	9.5
7	47	0	15	118.00	4.00	59.8
8	38	1	0	19.00	1.00	9.2
9	39	0	6	57.00	3.00	28.6
10	61	1	12	117.00	4.00	60.4
11	48	0	7	129.00	4.00	65.5
12	42	0	9	49.00	2.00	24.5
13	44	1	8	46.00	2.00	23.2

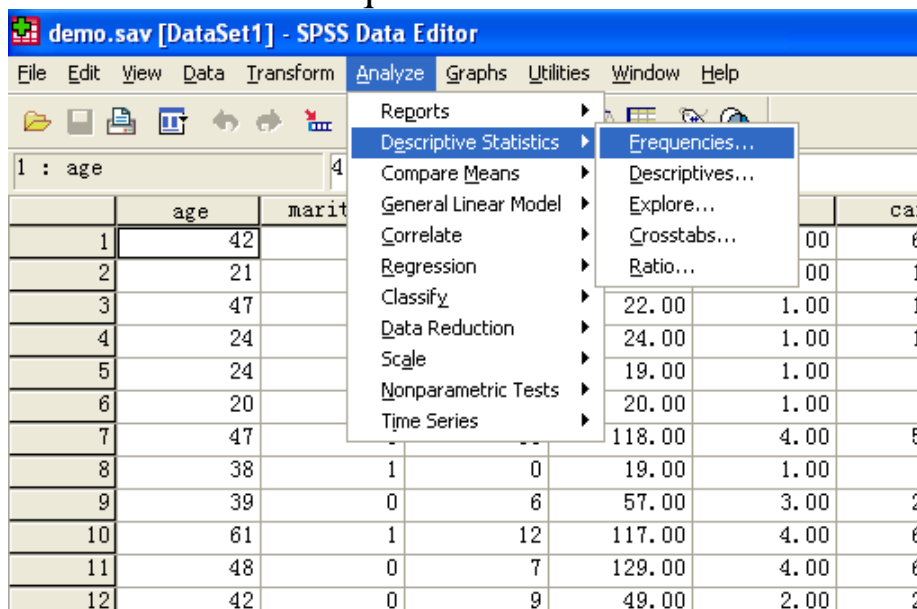
Data View Variable View

SPSS Processor is ready

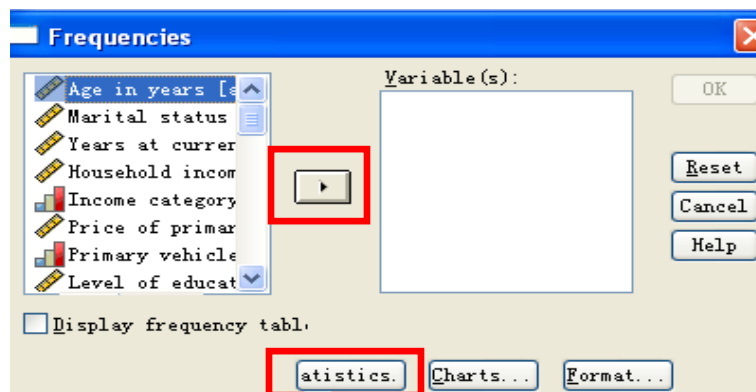
Frequencies:

The Frequencies procedure provides statistics and graphical displays that are useful for describing many types of variables. The Frequencies procedure is a good place to start looking at your data.

- From the menus choose:
Analyze
Descriptive Statistics
Frequencies...



- Select one or more categorical or quantitative variables.



statistics

- Click Statistics for descriptive statistics for quantitative variables. Click Charts for bar charts, pie charts, and histograms.

Frequencies: Statistics

Percentile Values

☐ Quartiles

☐ Cut points equal

☐ Percentile

Central Tendency

☐ Mean

☐ Median

☐ Mode

☐ Sum

☐ Values are group midpoi

Dispersion

☐ Std. deviat

☐ Minimum

☐ Variance

☐ Maximum

☐ Range

☐ S. E. mean

Distribution

☐ Skewness

☐ Kurtosis

Output1 - SPSS Viewer

File Edit View Data Transform Insert Format Analyze Graphs Utilities Window Help

Output

- Frequencies
 - Title
 - Notes
 - Active Data
 - Statistics

Frequencies

[DataSet1] C:\Program Files\SPSS Student\Tutorial\s

Statistics

Level of education		
N	Valid	1000
	Missing	0
Mean		2.52
Std. Error of Mean		.038
Median		2.00
Mode		2
Std. Deviation		1.200
Variance		1.439
Range		4
Minimum		1
Maximum		5

Graph:

You can create and edit a wide variety of chart types in SPSS. In these examples, we will create and edit six commonly used types of charts:

- Simple bar chart
- Pie chart
- Boxplot
- Scatter/Dot
- Histogram
- Q-Q

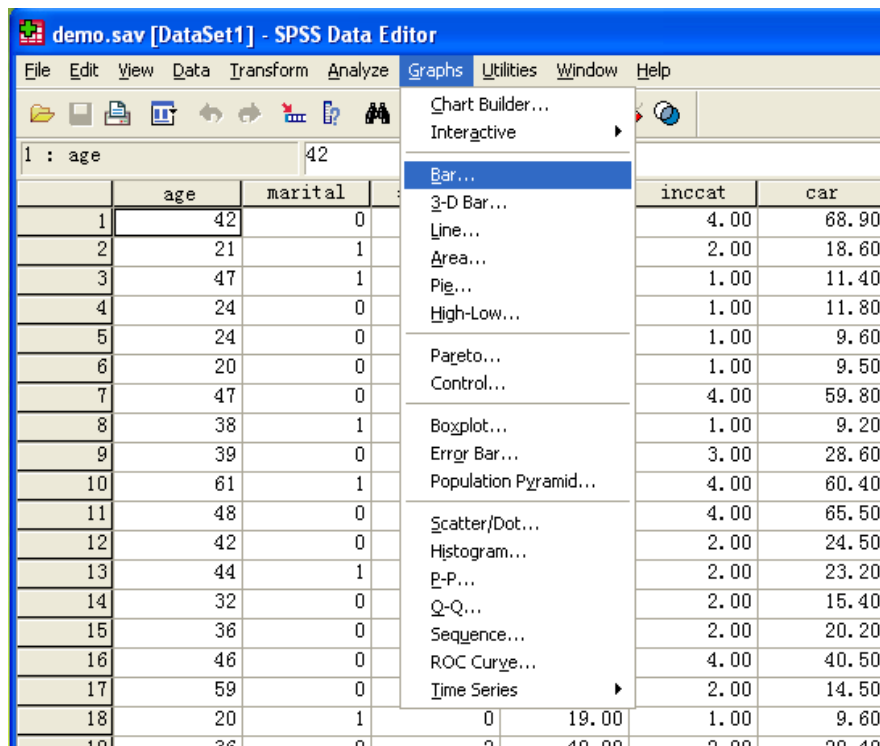
Simple bar chart

The bar chart will show a bar for each group (category) in a single categorical variable. The height of each bar will be determined by the result of a statistical calculation.

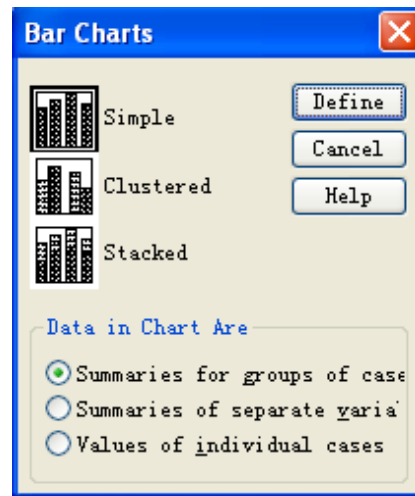
➤ From the menus, choose:

Graphs

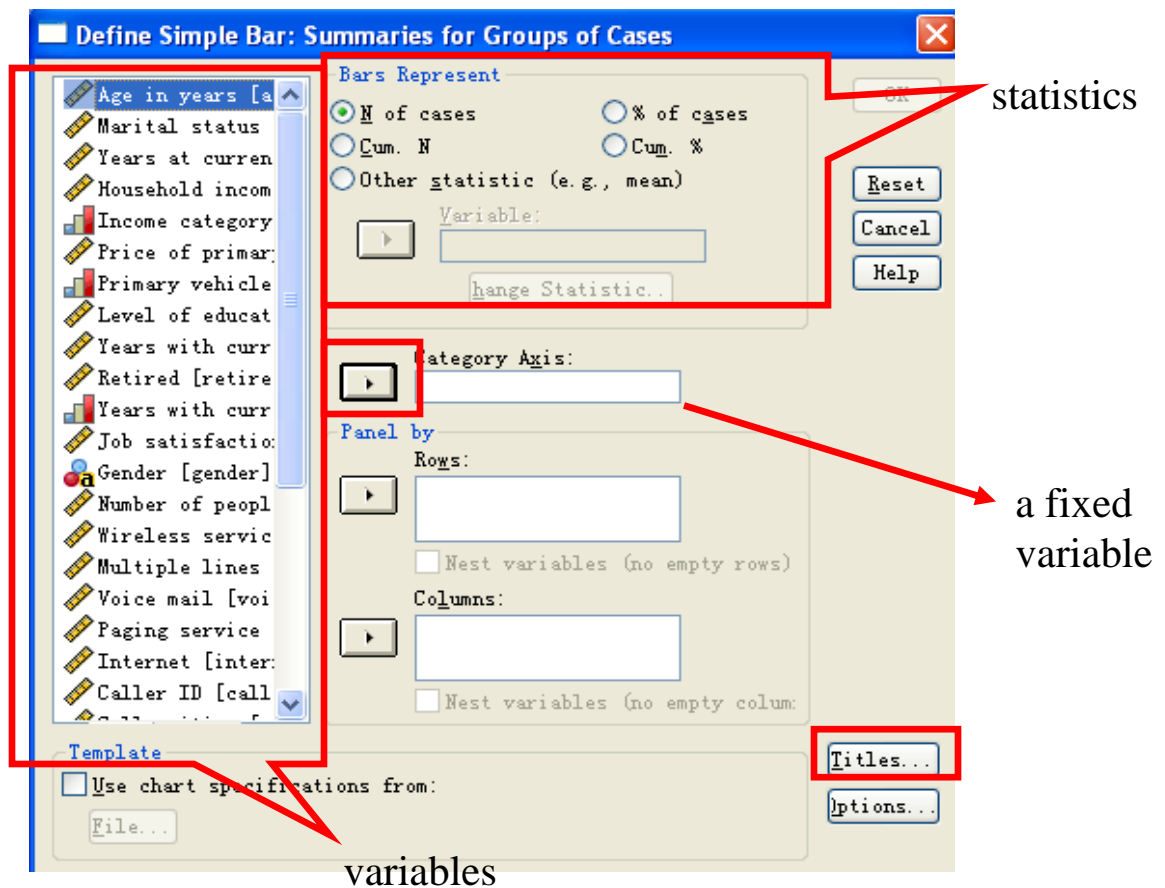
Bar

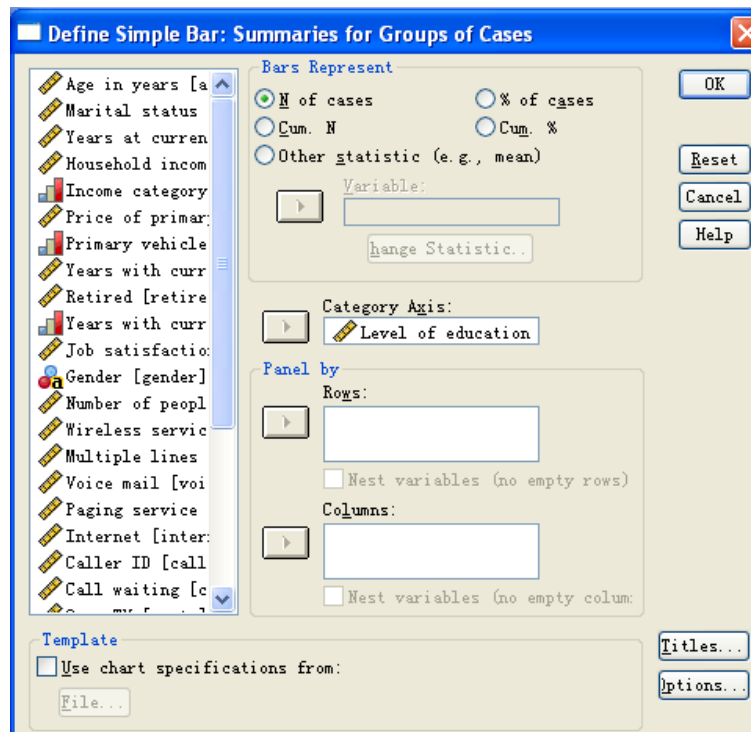


- In the Bar Charts initial dialog box, select the icon for simple, clustered, or stacked. Select the option under Data in Chart Are that best fits your data.

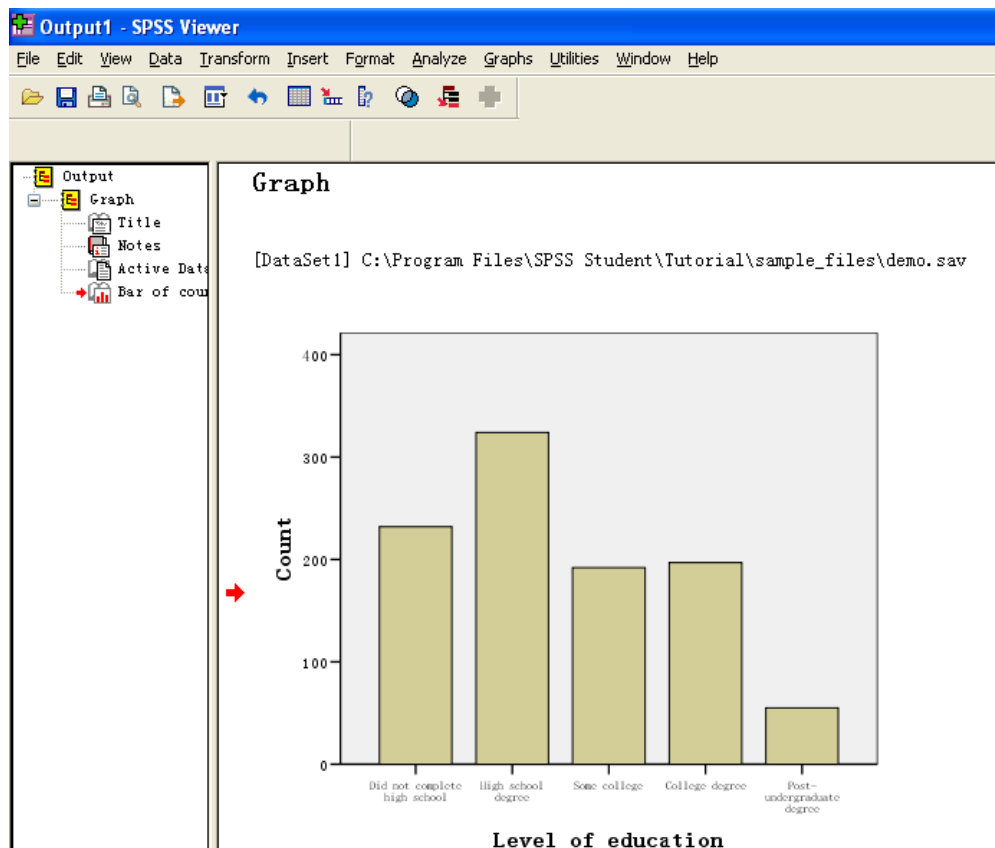


- Select Define, variables and options for the chart.

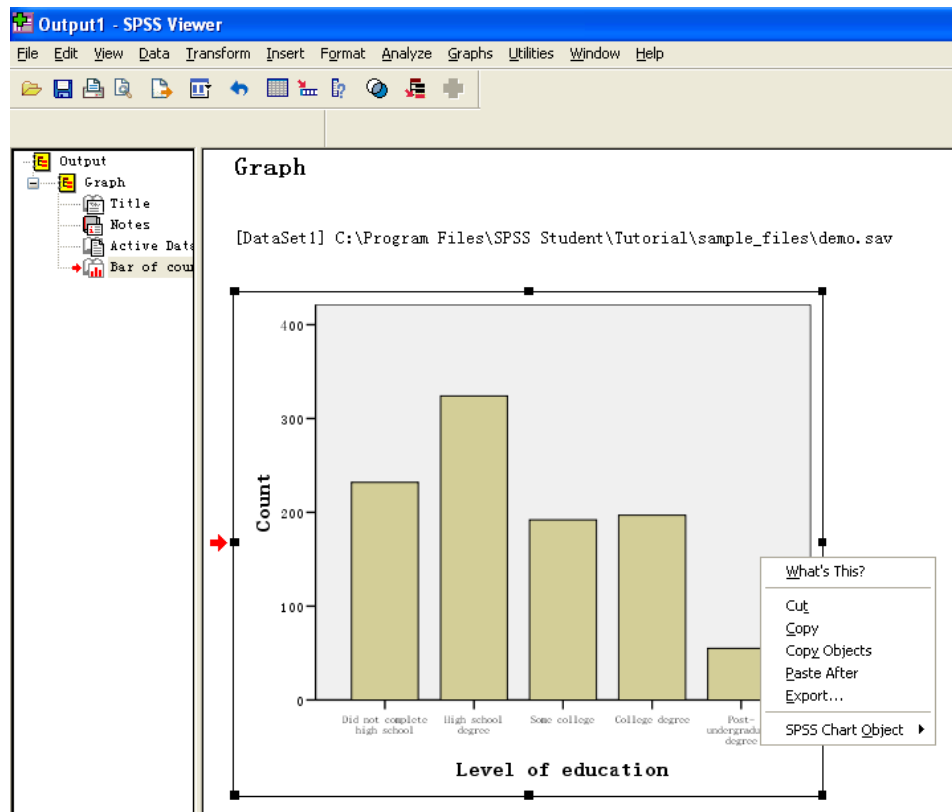




➤ Click OK to create the bar chart.



- Press the right arrow, you can “copy” or “export” the graph.



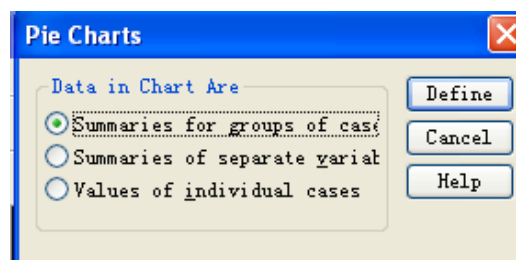
Pie chart

Pie Charts allows you to specify how data are represented in the chart.

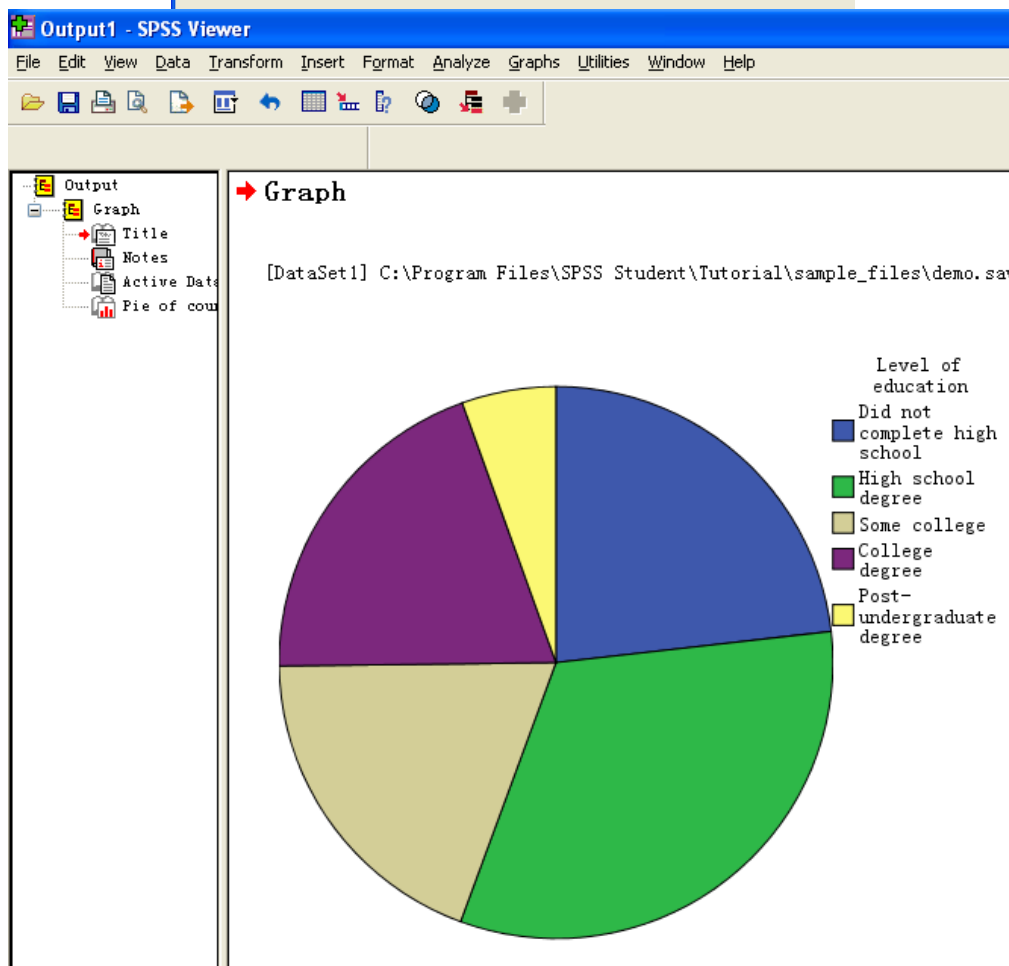
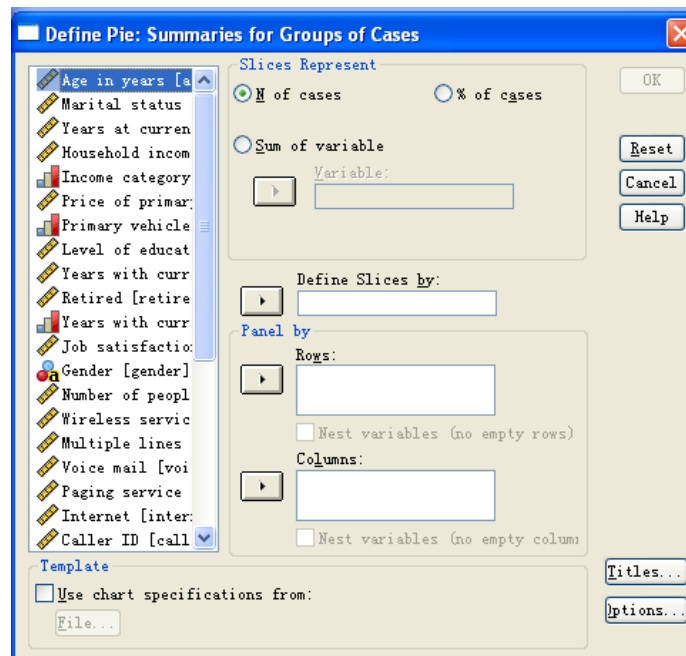
- From the menus, choose:

Graphs
Pie

- In the Pie Charts dialog, select an option under Data in Chart Are.



- Select Define, variables and options for the chart.



Boxplot

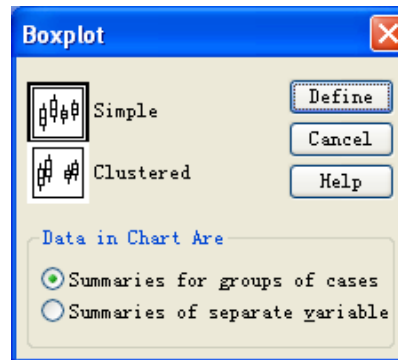
Boxplots show the median, interquartile range, outliers, and extreme cases of individual variables.

- From the menus, choose:

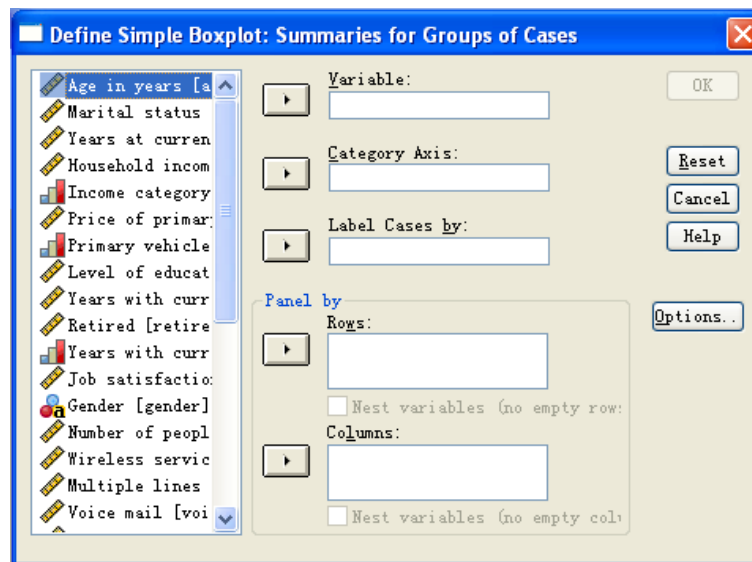
Graphs

Boxplot

- In the Boxplot initial dialog box, select the icon for simple or clustered. Select an option under Data in Chart Are.



- Select Define. Select variables and options for the chart.



Define Simple Boxplot: Summaries for Groups of Cases

Variable:

Category Axis:

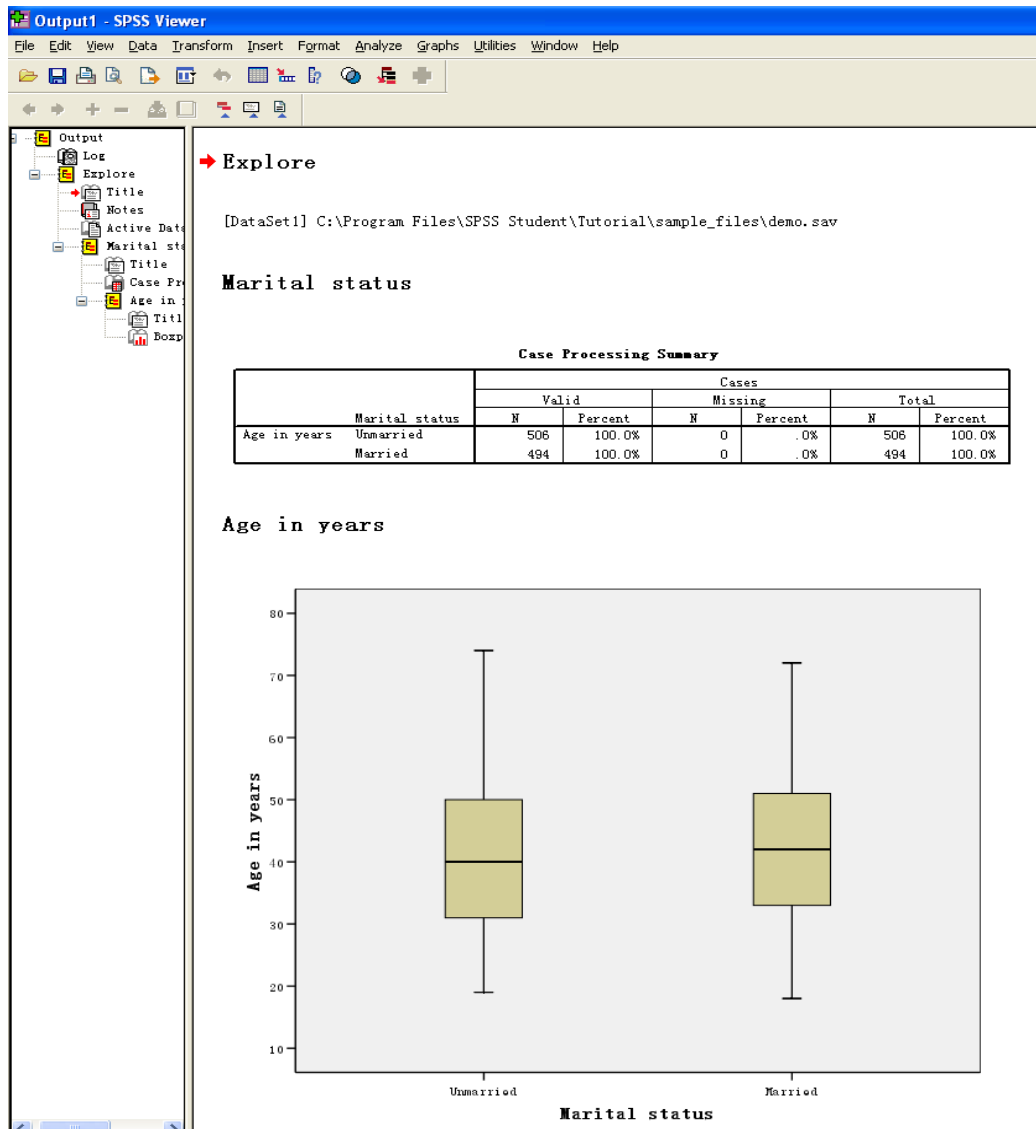
Label Cases by:

Panel by

Rows:

☐ Nest variables (no empty row)

Columns: ☐ Nest variables (no empty col)



Scatter/Dot

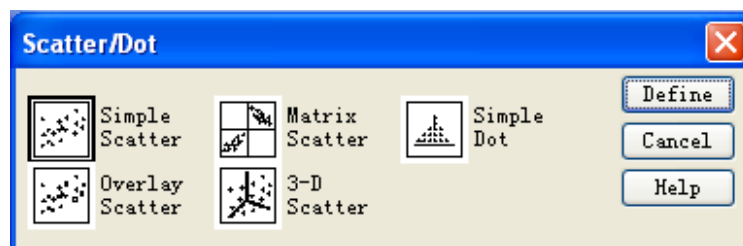
Scatter/Dot will show the relationship between two numeric variables.

- From the menus, choose:

Graphs

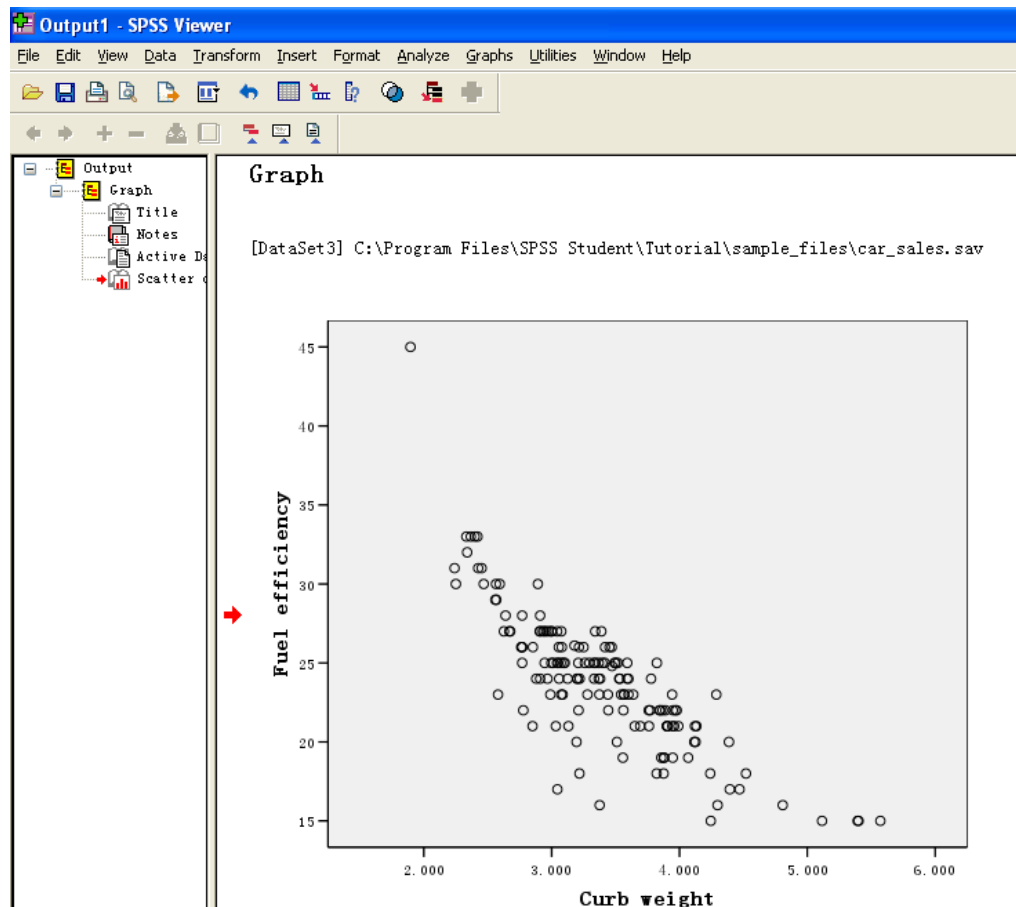
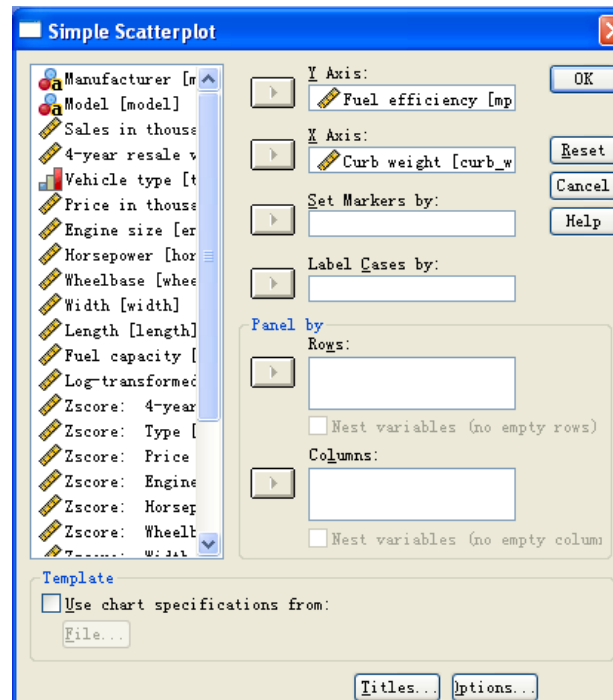
Scatter/Dot

- In the Scatter/Dot dialog, select the icon for simple, overlay, matrix, 3-D, or simple dot plot.



- Select Define. Select variables and options for the chart. For a scatter plot, you usually define a scale variable for each axis. Put the dependent variable on the y axis and the independent variable on the x axis.

In this example, we will create a scatterplot using the data file car_sales.sav



Histogram

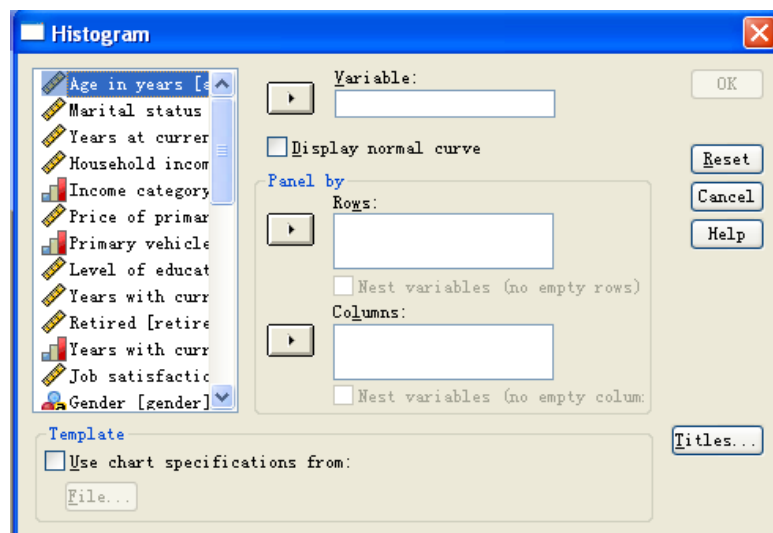
Creates a histogram showing the distribution of a single numeric variable.

- From the menus, choose:

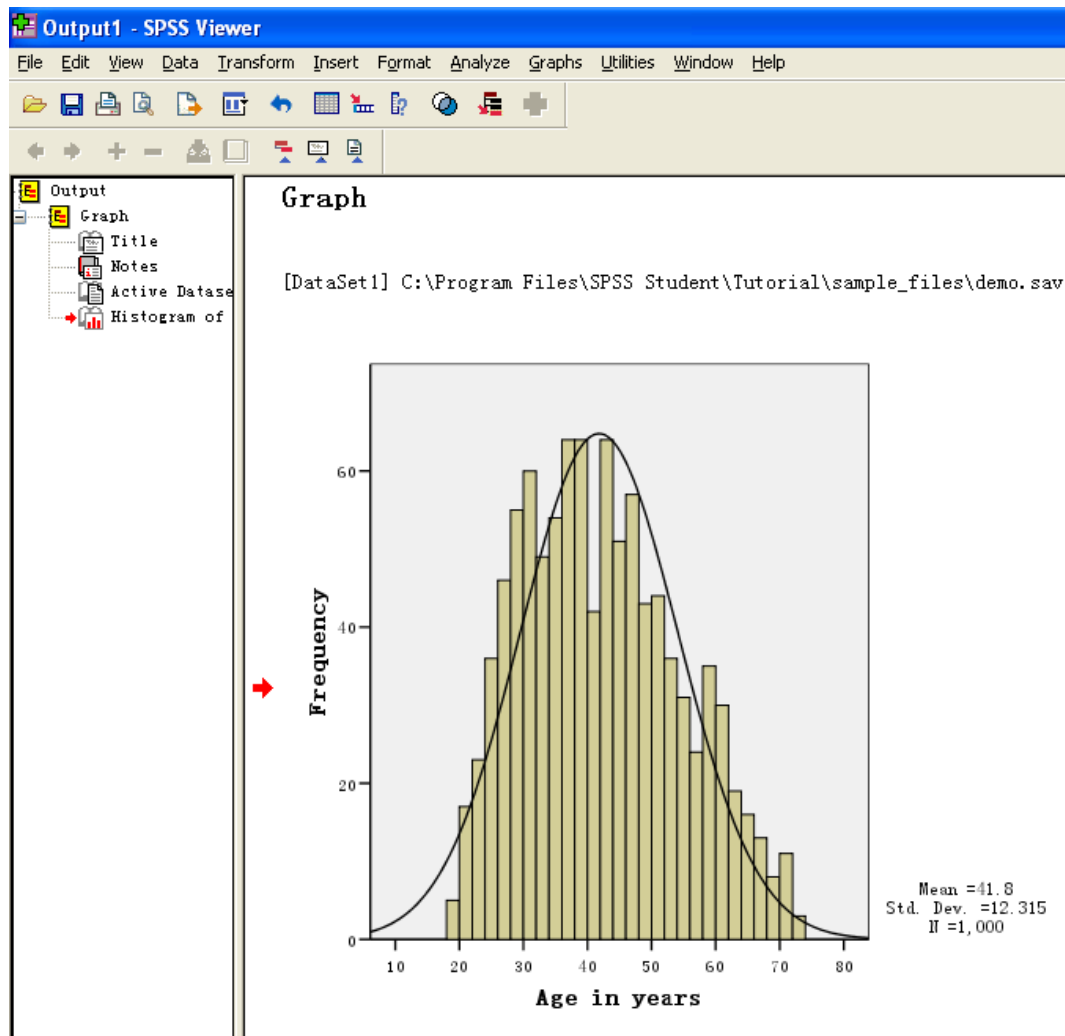
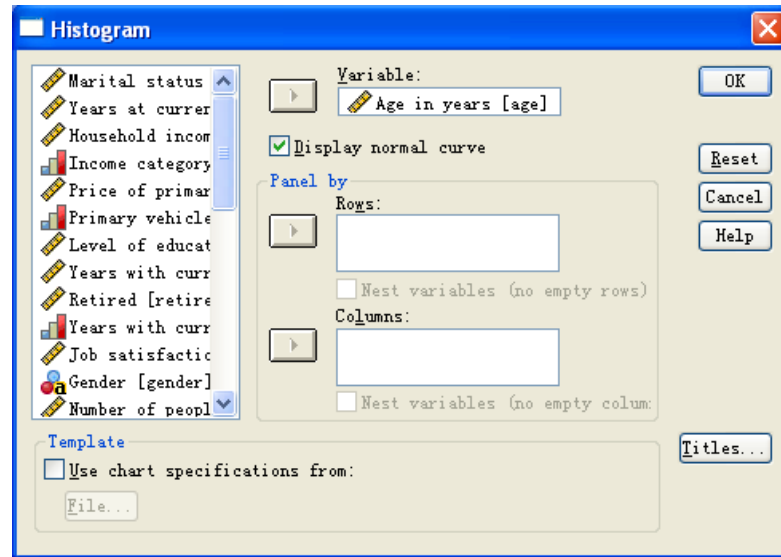
Graphs

Histogram

- Select a numeric variable for Variable in the Histogram dialog.



- Select Display normal curve to display a normal curve on the histogram.



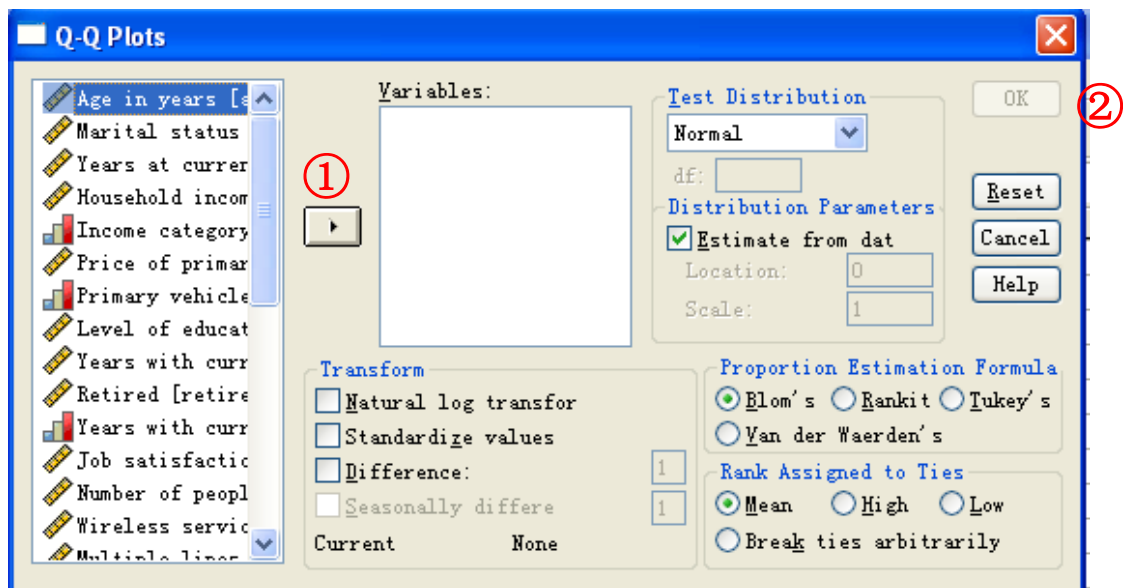
Q-Q

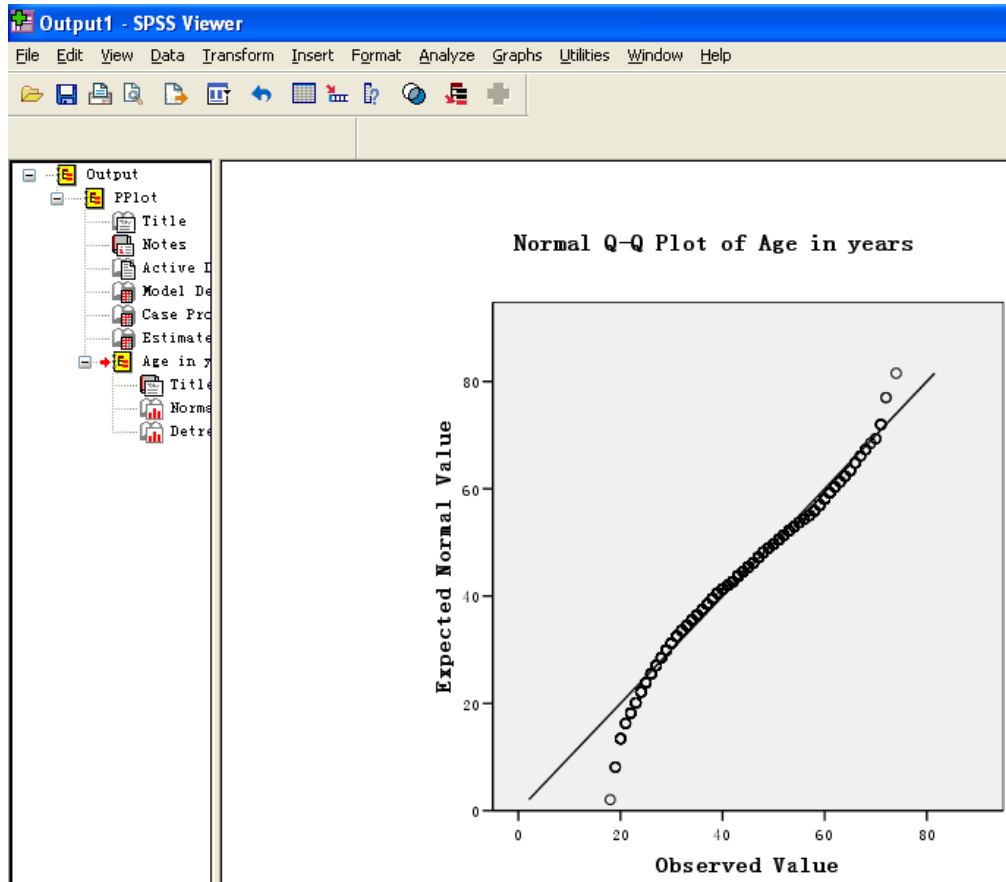
Plots the quintiles of a variable's distribution against the quintiles of any of a number of test distributions. Probability plots are generally used to determine whether the distribution of a variable matches a given distribution. If the selected variable matches the test distribution, the points cluster around a straight line.

- From the menus, choose:

Graphs
Q-Q

- Select one or more numeric variables and move them onto the Variables list.
- Select a test distribution.





Regression Analysis

Linear regression is used to model the value of a dependent scale variable based on its linear relationship to one or more predictors.

Given the data (the exercise 2.10, p74, textbook), we only use the data to introduce how to do the regression analysis.

1. Linear regression.

Method 1.

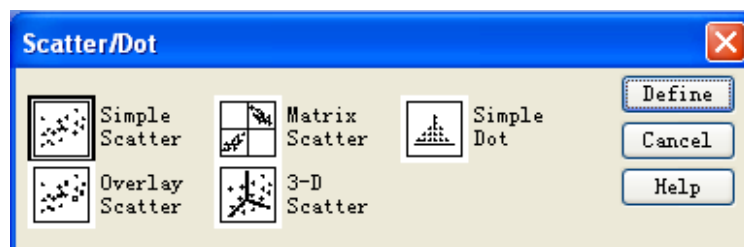
Scatter/Dot will show the relationship between two numeric variables and can be used to construct the linear regression line.

- From the menus, choose:

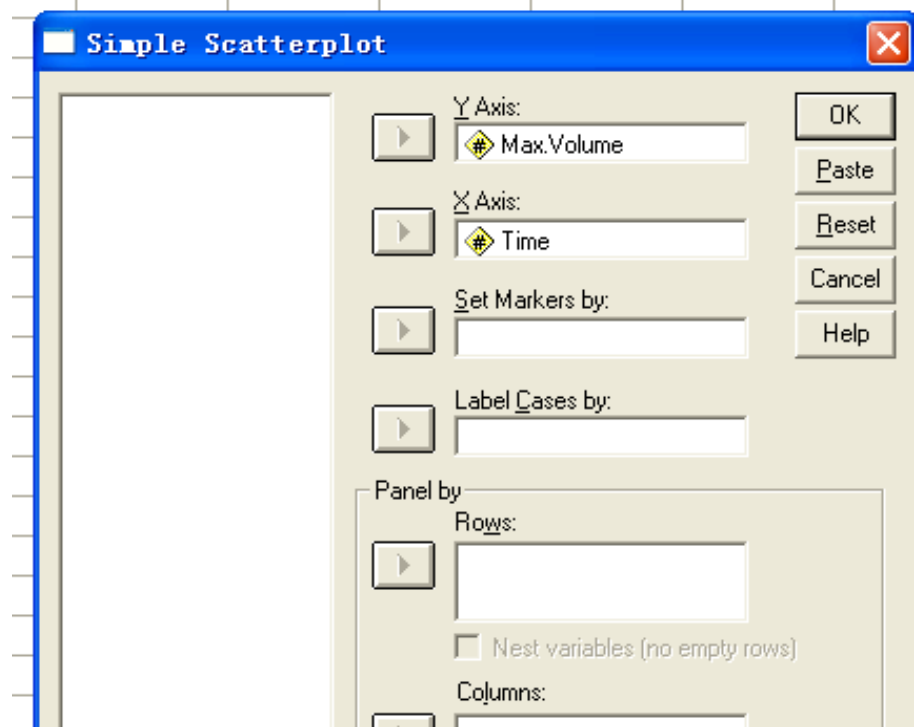
Graphs

Scatter/Dot

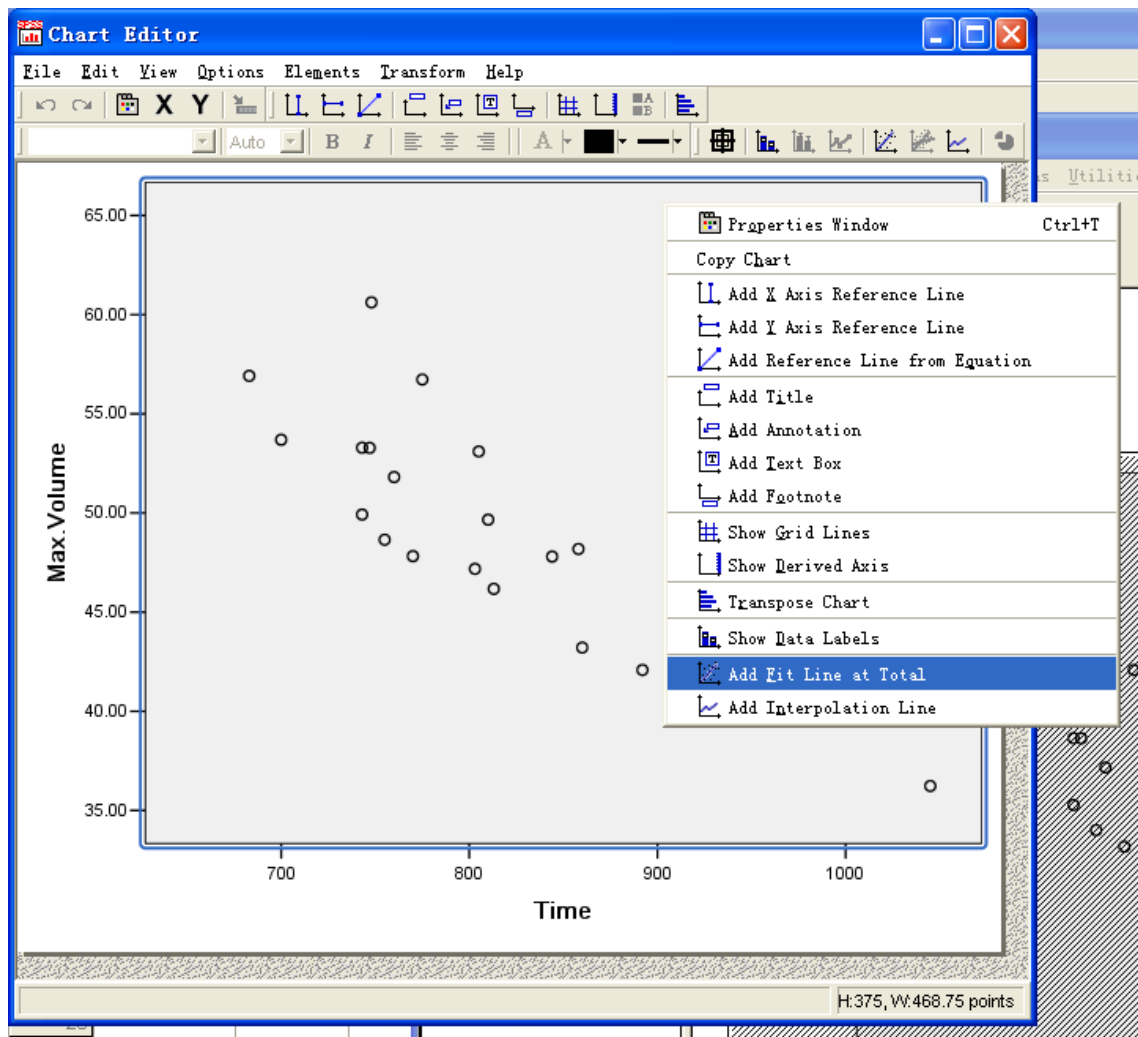
- In the Scatter/Dot dialog, click *Simple Scatter*



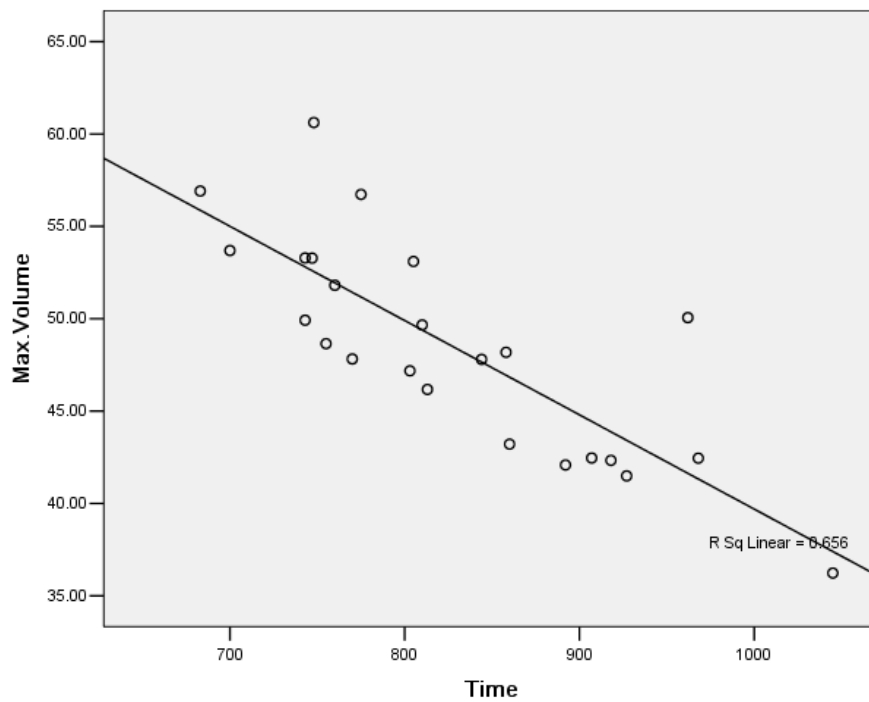
- Select Define. Select variables and options for the chart. For a scatter plot, you usually define a scale variable for each axis. Put the dependent variable on the y axis and the independent variable on the x axis.



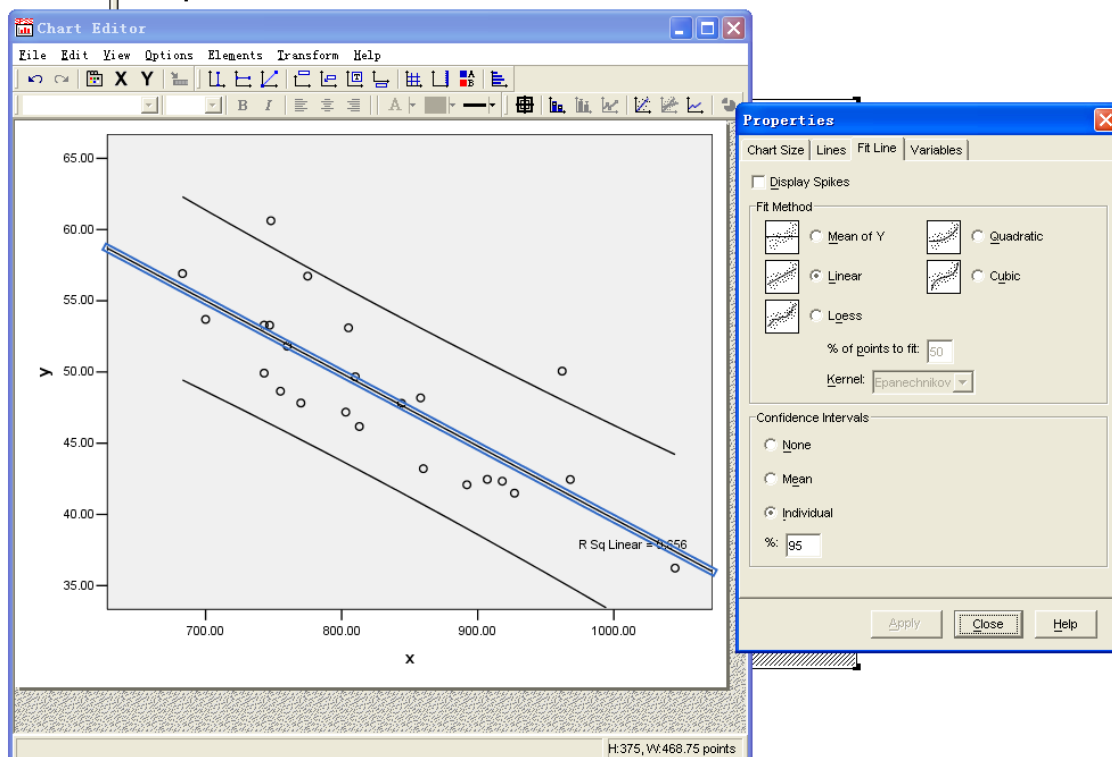
- click *Ok*
- In the output, double click the graph.



- To see a best-fit line overlaid on the points in the scatterplot, activate the graph by double-clicking on it.
- Select a point in the Chart Editor.
- Click the Add fit line tool, then close the Chart Editor.

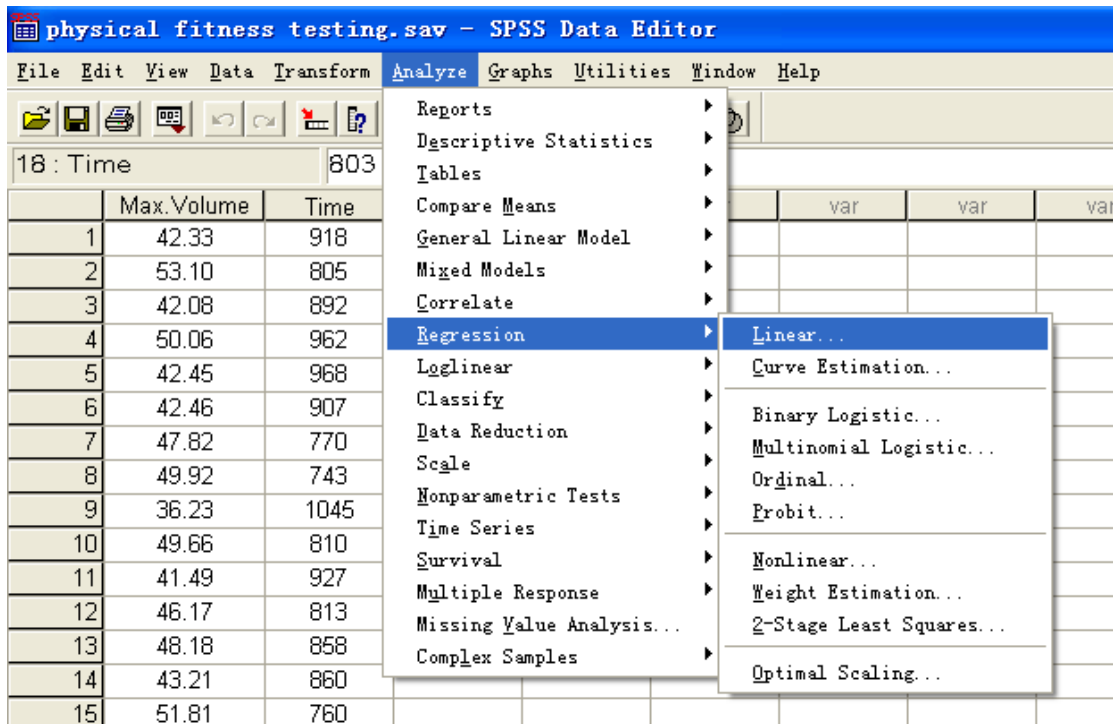


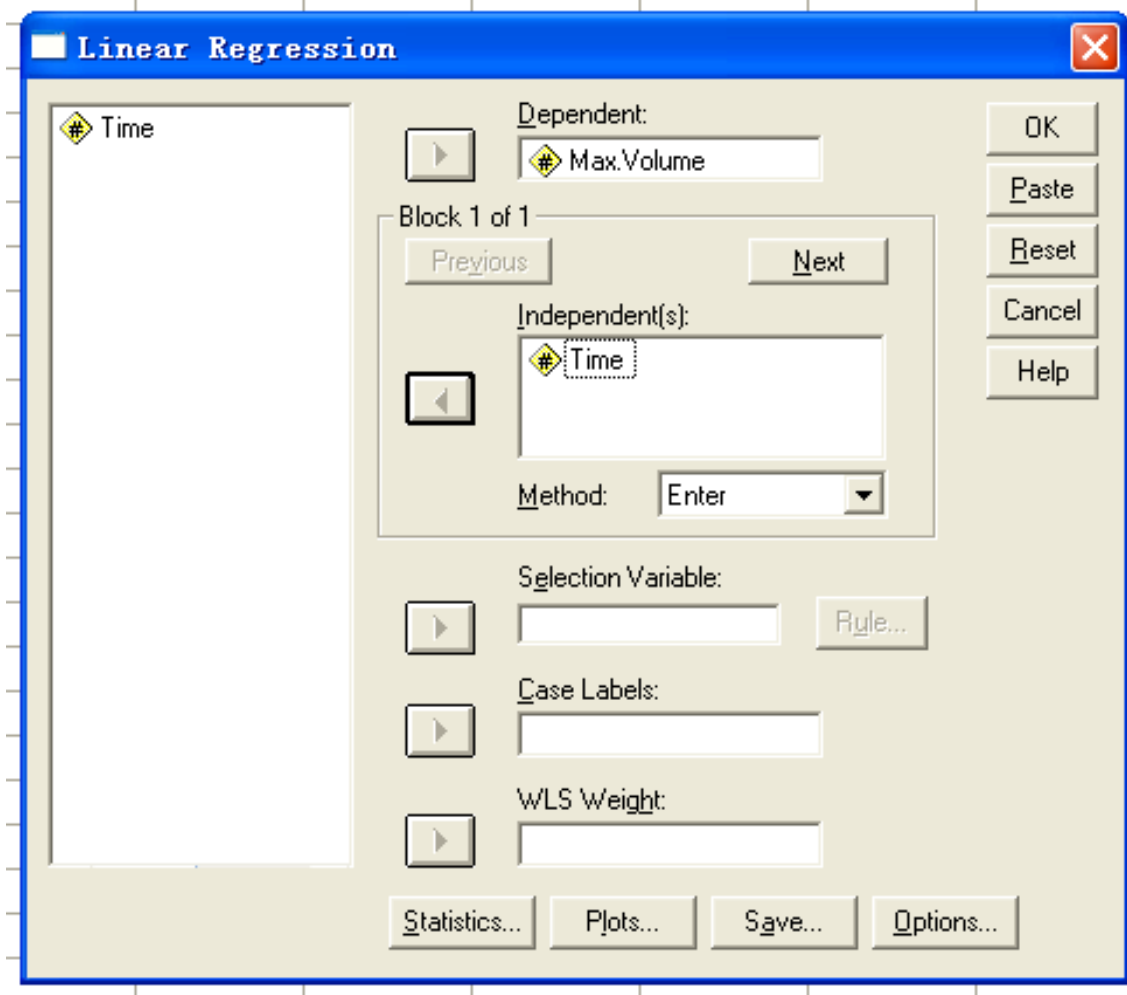
- Here you also can add the confidence interval 95% (can be used to detect the outliers)



Method 2

- (1). Insert the data.
- (2). Use the menu Analyze ==> Regression ==> Linear.



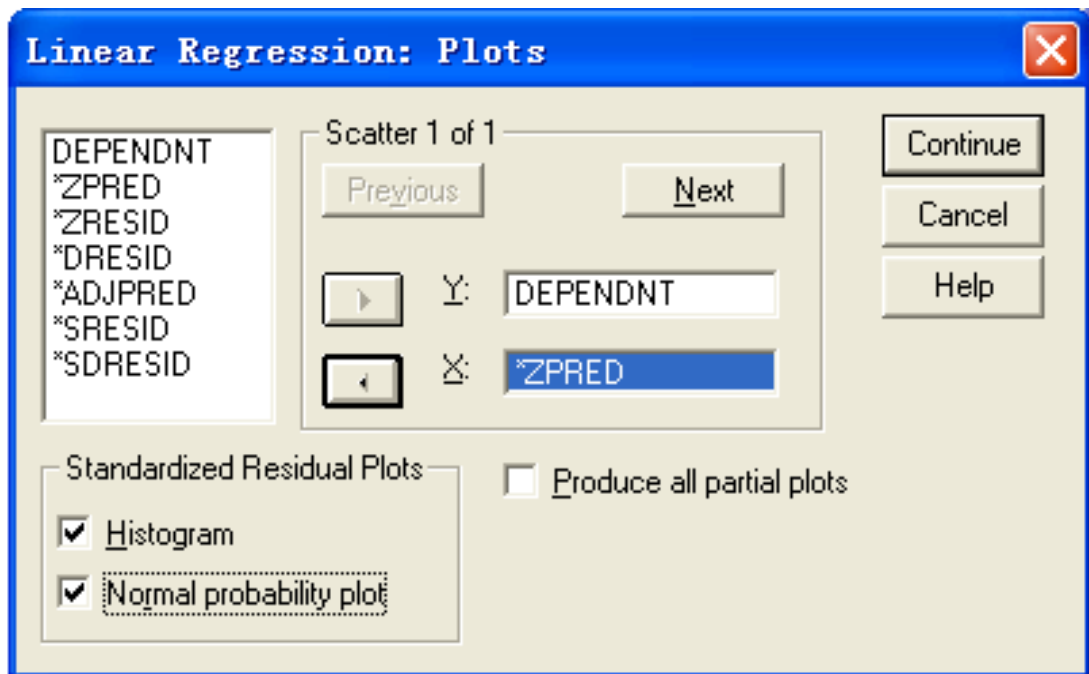


where

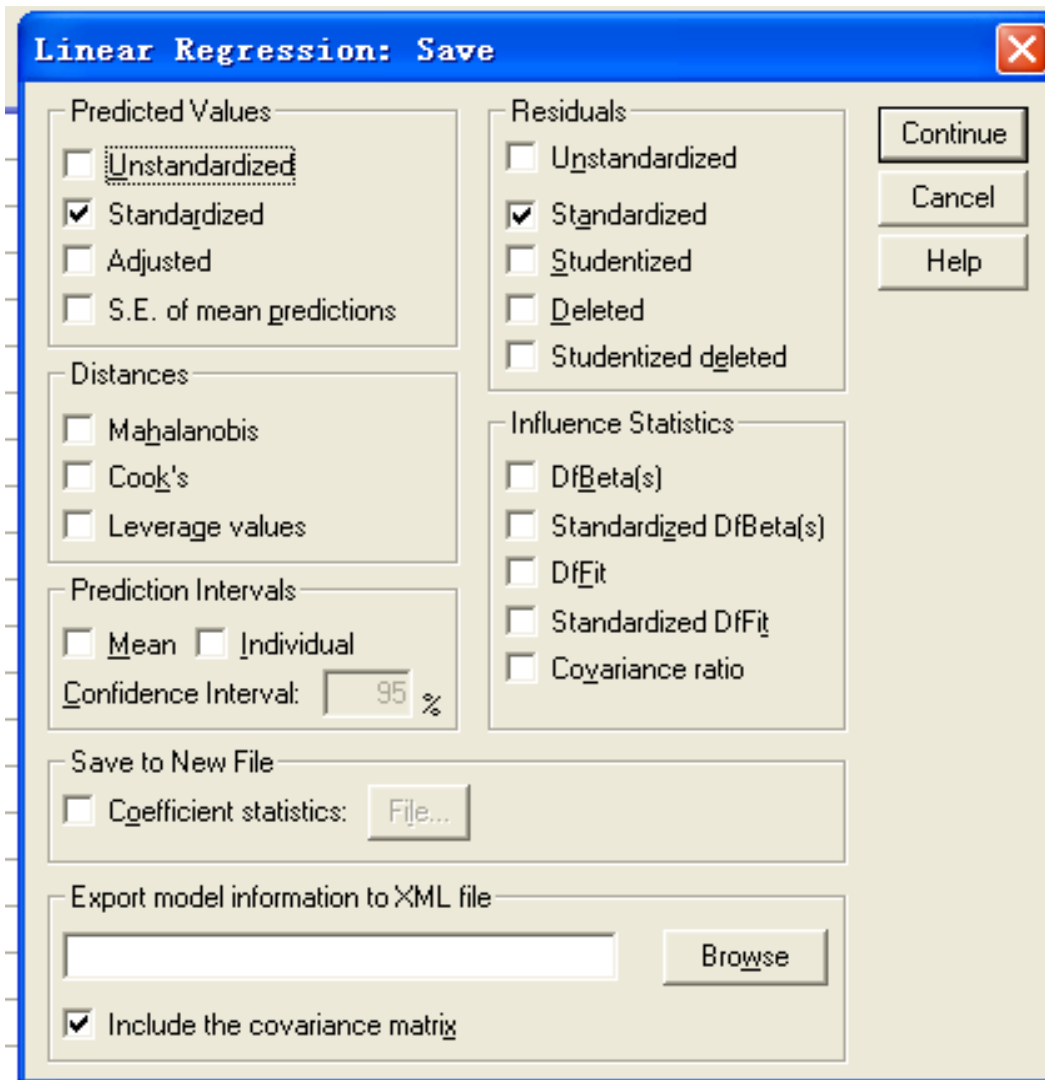
- Select Maximum Volume of O₂ (Max.Volume, for simplicity) as the dependent variable.
- Select Time in Seconds (Time, for simplicity) as the independent variable.
- Click **Plots**.

Method:

- Enter (all variables enter in)
- Stepwise
- Backward
- Forward (different methods to select the best models)



- Select DEPENDNT as the x variable and *ZPRED as the y variable.
- Select DEPENDNT as the x variable and *ZRESID as the y variable. (residual vs y)
- Select Histogram and Normal probability plot.
- Select “Produce all partial plots”, (partial plots in diagnostic plots)
- Click Continue
- Click Save in the Linear Regression dialog box



The image shows a 'Linear Regression: Save' dialog box with a blue title bar and a red close button. It contains several groups of options for saving regression results. The 'Predicted Values' group has 'Standardized' checked. The 'Residuals' group has 'Standardized' checked. The 'Distances' group has all options unchecked. The 'Prediction Intervals' group has 'Mean' and 'Individual' unchecked, with a 'Confidence Interval' of 95%. The 'Save to New File' group has 'Coefficient statistics' unchecked. The 'Export model information to XML file' group has a text box and a 'Browse' button, with 'Include the covariance matrix' checked. On the right, there are 'Continue', 'Cancel', and 'Help' buttons.

Group	Option	Selected
Predicted Values	Unstandardized	<input type="checkbox"/>
	Standardized	<input checked="" type="checkbox"/>
	Adjusted	<input type="checkbox"/>
	S.E. of mean predictions	<input type="checkbox"/>
Residuals	Unstandardized	<input type="checkbox"/>
	Standardized	<input checked="" type="checkbox"/>
	Studentized	<input type="checkbox"/>
	Deleted	<input type="checkbox"/>
	Studentized deleted	<input type="checkbox"/>
Distances	Mahalanobis	<input type="checkbox"/>
	Cook's	<input type="checkbox"/>
	Leverage values	<input type="checkbox"/>
Prediction Intervals	Mean	<input type="checkbox"/>
	Individual	<input type="checkbox"/>
Confidence Interval: 95 %		
Save to New File	Coefficient statistics: File...	<input type="checkbox"/>
	Export model information to XML file	
Text box		<input type="text"/>
Browse		<input type="button"/>
Include the covariance matrix		<input checked="" type="checkbox"/>

- Select Standardized in the Predicted Values group.
- Select Standardized in the Residuals group.
- Click Continue.
- Click OK in the Linear Regression dialog box.

These selections produce a linear regression model for polishing time based on diameter. Diagnostic plots of the Studentized residuals by the model-predicted values are requested, and various values are saved for further diagnostic testing.

Coefficients^a

Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.
	B	Std. Error	Beta		
1 (Constant)	90.722	6.546		13.858	.000
Time	-.051	.008	-.810	-6.481	.000

a. Dependent Variable: Max.Volume

This table shows the coefficients of the regression line.

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	514.683	1	514.683	42.006	.000 ^a
	Residual	269.555	22	12.253		
	Total	784.239	23			

a. Predictors: (Constant), Time

b. Dependent Variable: Max.Volume

- The Regression row displays information about the variation accounted for by your model.
- The Residual row displays information about the variation that is not accounted for by your model.
- The regression and residual sums of squares are approximately equal, which indicates that about half of the variation in polishing time is explained by the model.
- The significance value of the F statistic is less than 0.05, which means that the variation explained by the model is not due to chance.
-

While the ANOVA table is a useful test of the model's ability to explain any variation in the dependent variable, it does not directly address the strength of that relationship.

Model Summary^b

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.810 ^a	.656	.641	3.50036

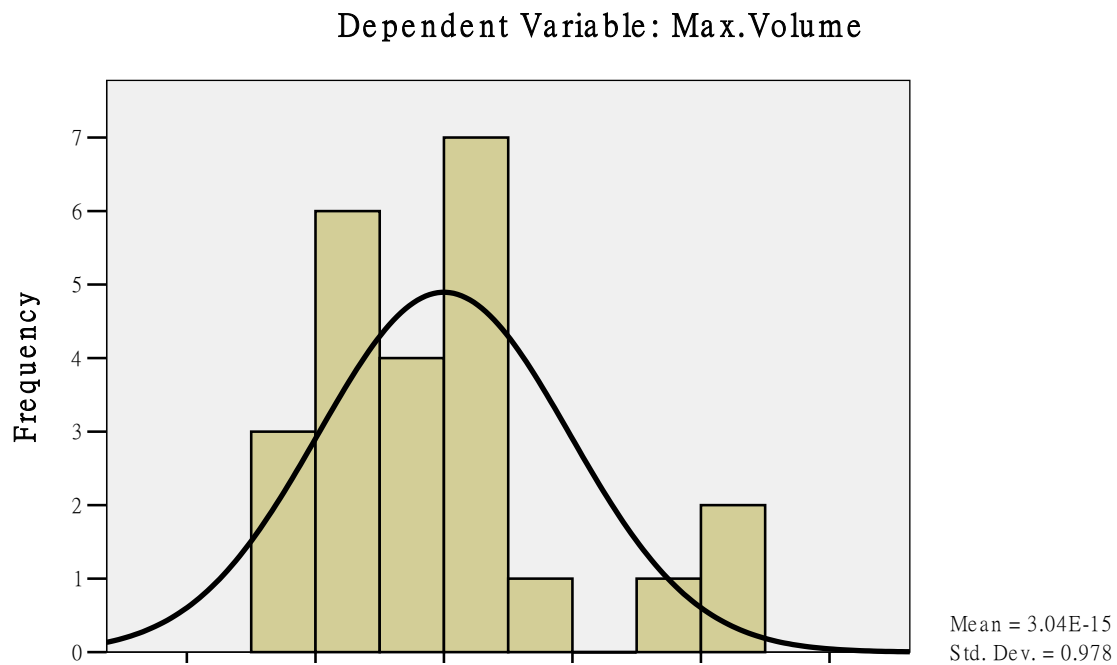
a. Predictors: (Constant), Time

b. Dependent Variable: Max. Volume

The model summary table reports the strength of the relationship between the model and the dependent variable.

- R, the multiple correlation coefficient, is the linear correlation between the observed and model-predicted values of the dependent variable. Its large value indicates a strong relationship.
- R Square, the coefficient of determination, is the squared value of the multiple correlation coefficient. It shows that about half the variation in time is explained by the model.
-

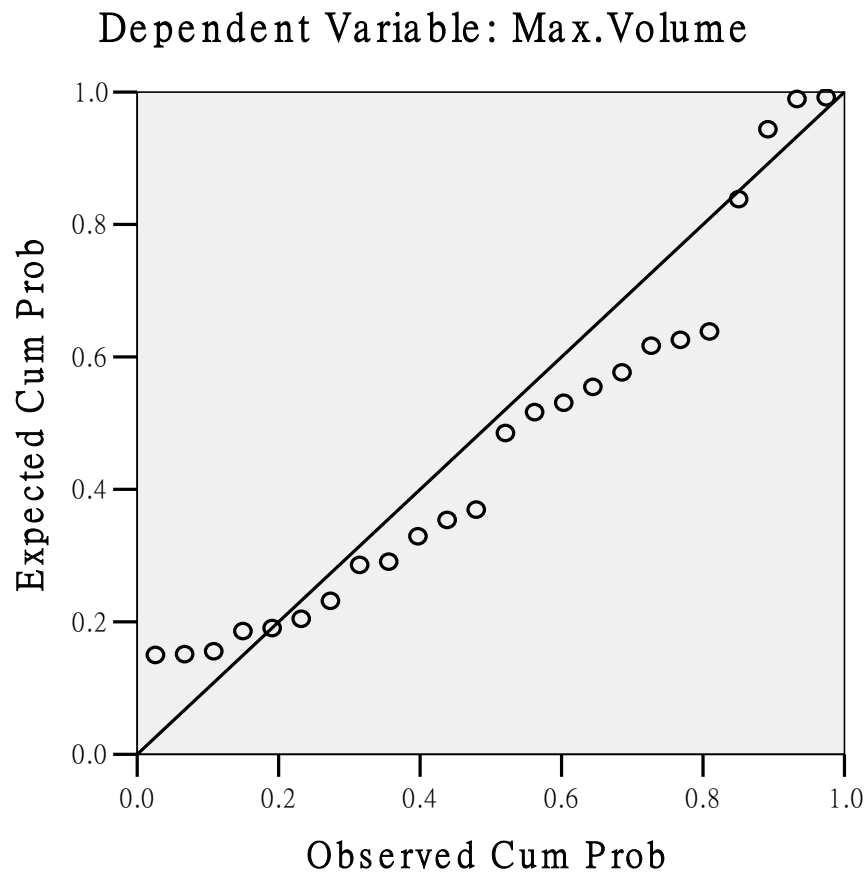
Histogram



A residual is the difference between the observed and model-predicted values of the dependent variable. The residual for a given product is the observed value of the error term for that product. A histogram or P-P plot of the residuals will help you to check the assumption of normality of the error term.

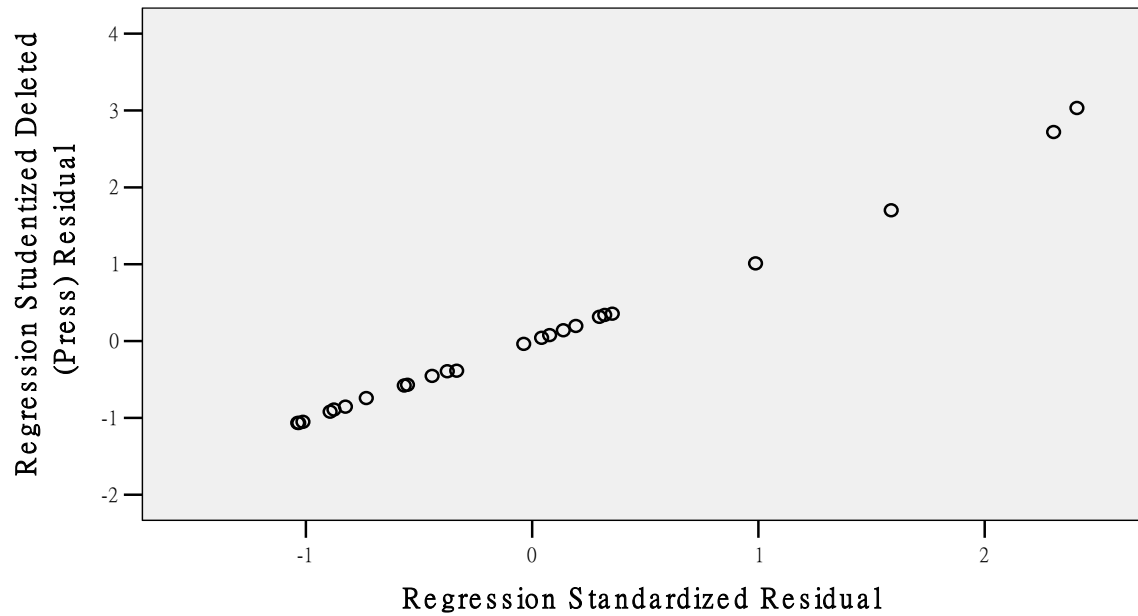
The shape of the histogram should approximately follow the shape of the normal curve. This histogram is unacceptably close to the normal curve.

Normal P-P Plot of Regression Standardized Residual

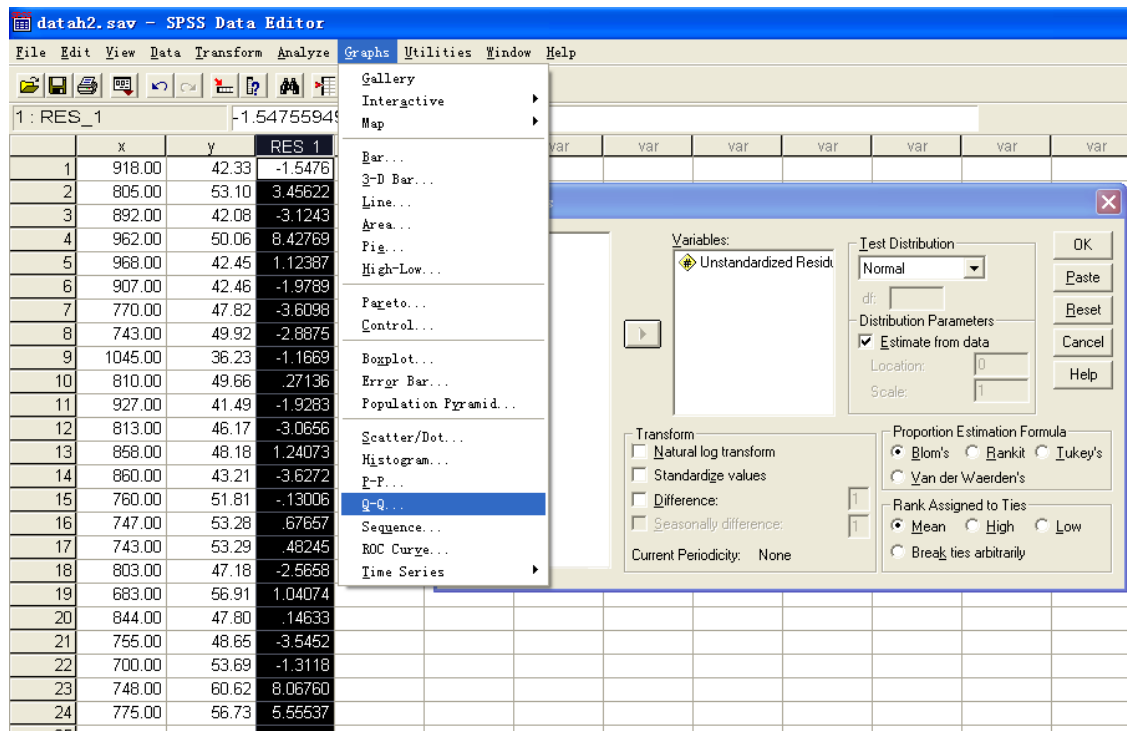


The P-P plotted residuals should follow the 45-degree line. Neither the histogram nor the P-P plot indicates that the normality assumption is violated.

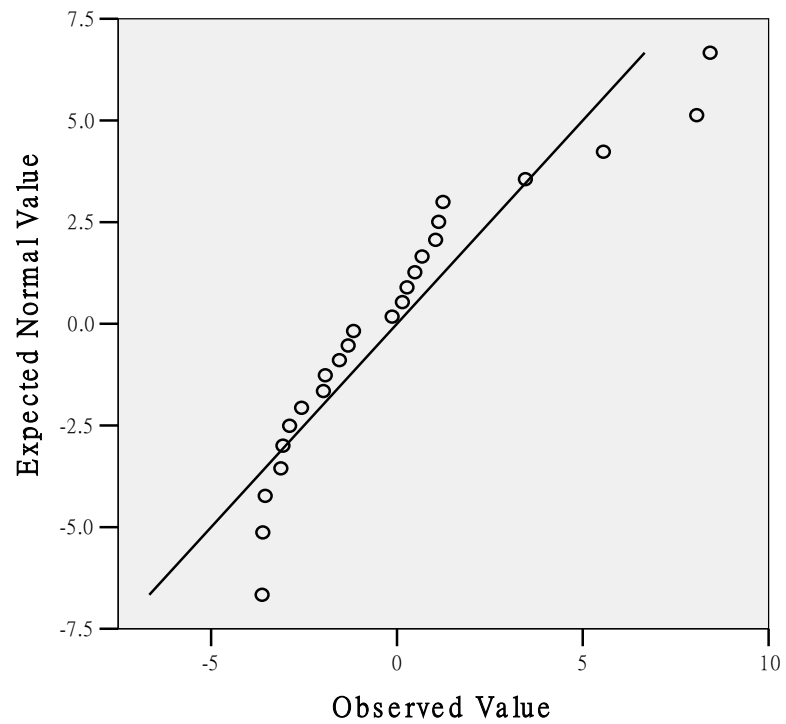
Dependent Variable: Max.Volume



Q-Q plot of residuals

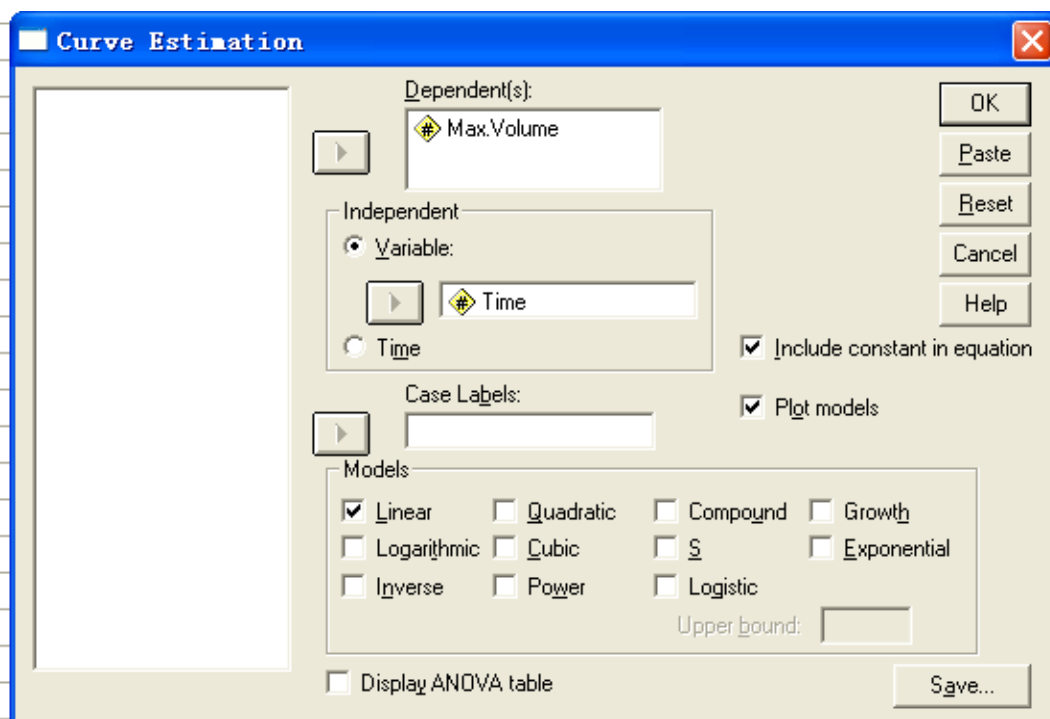
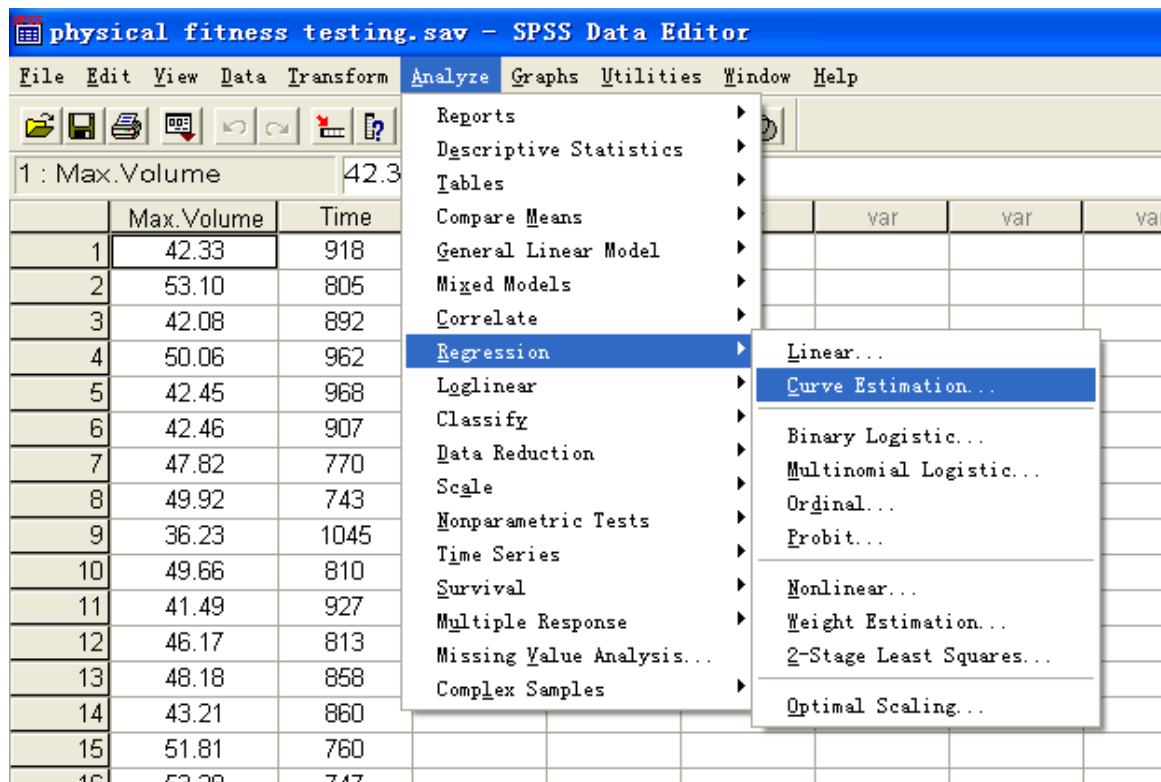


Normal Q-Q Plot of Unstandardized Residual



Method 3.

- (1). Insert the data.
- (2). Use the menu Analyze ==> Regression ==> Curve Estimation.



Max. Volume

