Mathematical modeling

第九讲 统计回归模型(3)

周毓明
zhouyuming@nju.edu.cn

南京大学计算机科学与技术系

课程内容

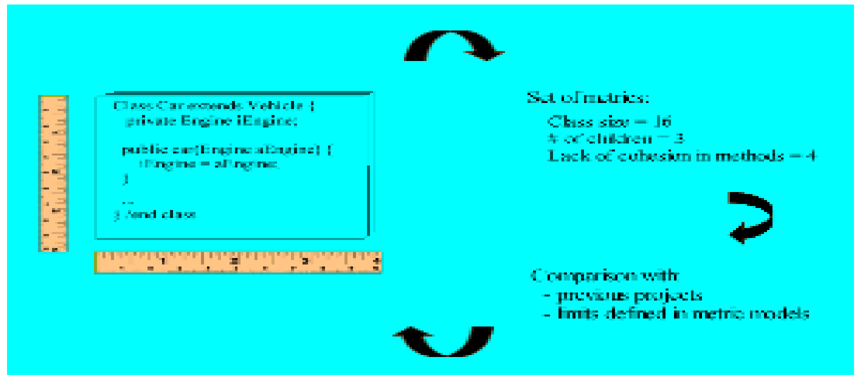
1. 数学概念与模型
2. 实际案例与分析
3. 计算机典型应用

3.计算机典型应用

① 软件缺陷预测

② 其他应用…

软件缺陷预测：背景



(Copied from *staff.cs.utu.fi/opin**not**/kurssit/SemSE/05/SES-Malminen.ppt*)

No	WMC	LCOM	…	SLOC	Fault
1	35	96	…	1255	2
2	73	100		2828	0
…	…	…	…	…	…

No	WMC	LCOM	…	SLOC	Fault
1	15	9	…	135	?
2	45	10		282	?
…	…				

软件缺陷预测：背景

什么是软件度量？

血常规检查

代号	名称	结果	参考值
WBC	白细胞计数值	5.5	4–10 10^9/L
RBC	红细胞计数值	5.53↑	4–5.5 10^12/L
…	…	…	…
LYM%	淋巴细胞比率	0.401 ↑	0.2–0.4
MXD%	中值细胞比率	0.031↓	0.035–0.14
…	…	…	…
RDW-C	红细胞分布宽度	0.135	0.11–0.16 fL
RDW SI	红细胞分布宽度	40.2	37–54
…	…	…	…

软件缺陷预测：背景

什么是软件度量？

代码度量: 从源代码中抽取出的“特征”

- 规模： LOC, Stmts, ...
- 内聚性: LCOM系列, TCC, LCC, CAMC, ...
- 耦合性: CBO, RFC, ...
- 继承相关: DIT, NOC, NOP, NMI, ...
- 复杂性: WMC, CDE, CIE, ...

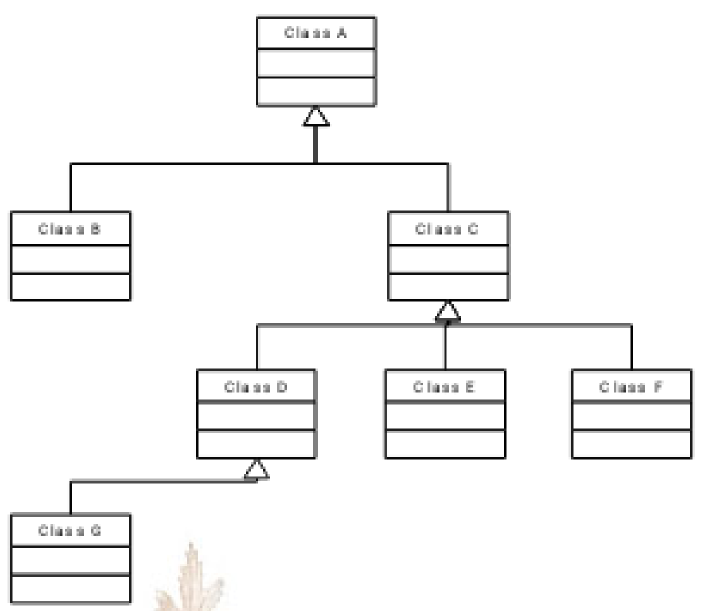
软件缺陷预测：背景

什么是软件度量？

- DIT(A)=0
- DIT(B,C)=1
- DIT(D,E,F)=2
- DIT(G)=3

Another Example

- NOC(A)=2
- NOC(C)=3
- NOC(D)=1
- NOC(B,E,F,G)=0



软件缺陷预测：问题

问题描述

给定一个程序，假定在每个模块上已经收集了许多度量，那么如何根据这些度量值来预测有缺陷的模块？

No	WMC	LCOM	…	SLOC	Fault
1	35	96	…	1255	?
2	73	100		2828	?
…	…				

?

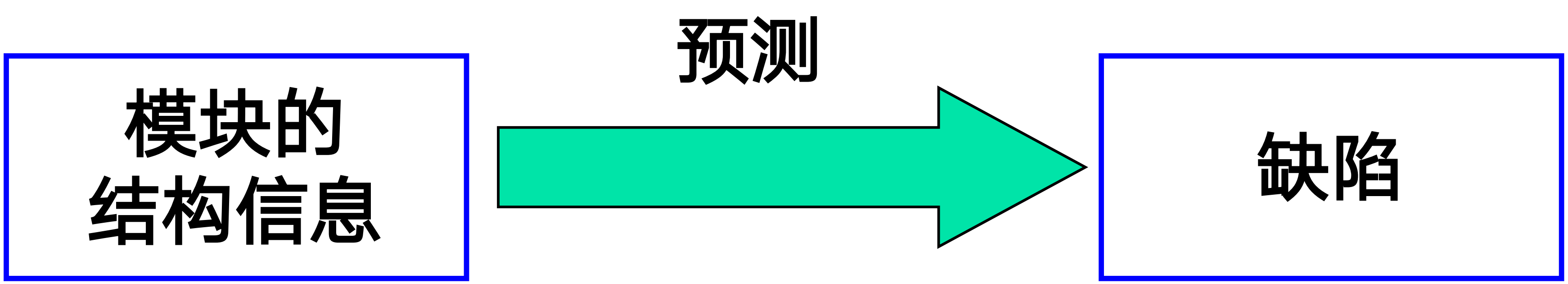
软件缺陷预测：问题

问题描述

① 模块中包含多少个缺陷？ → 预测数量

② 系统中哪些模块包含缺陷？ → 预测类别

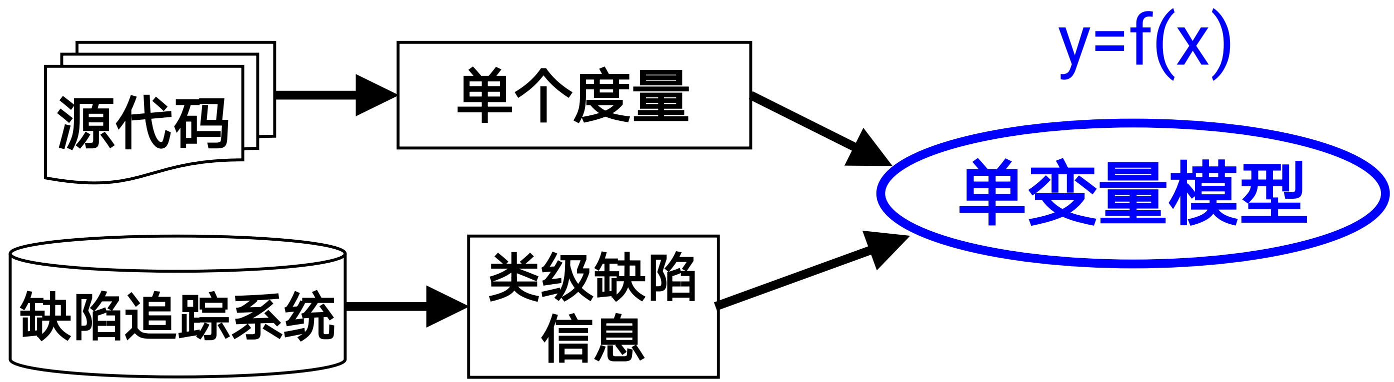
③ 系统中哪些模块最有可能包含缺陷？ → 预测序



软件缺陷预测：方法

预测方法

第一步：
单变量分析



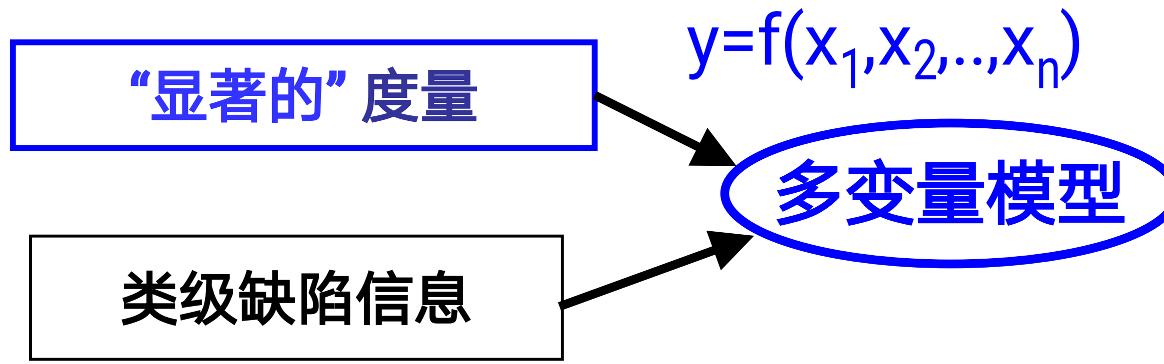
分析单个度量与缺陷的统计相关性(alpha = 0.05)。

f(x)可为线性回归模型、logistic回归模型或者其他模型

软件缺陷预测：方法

预测方法

第二步：
多变量分析

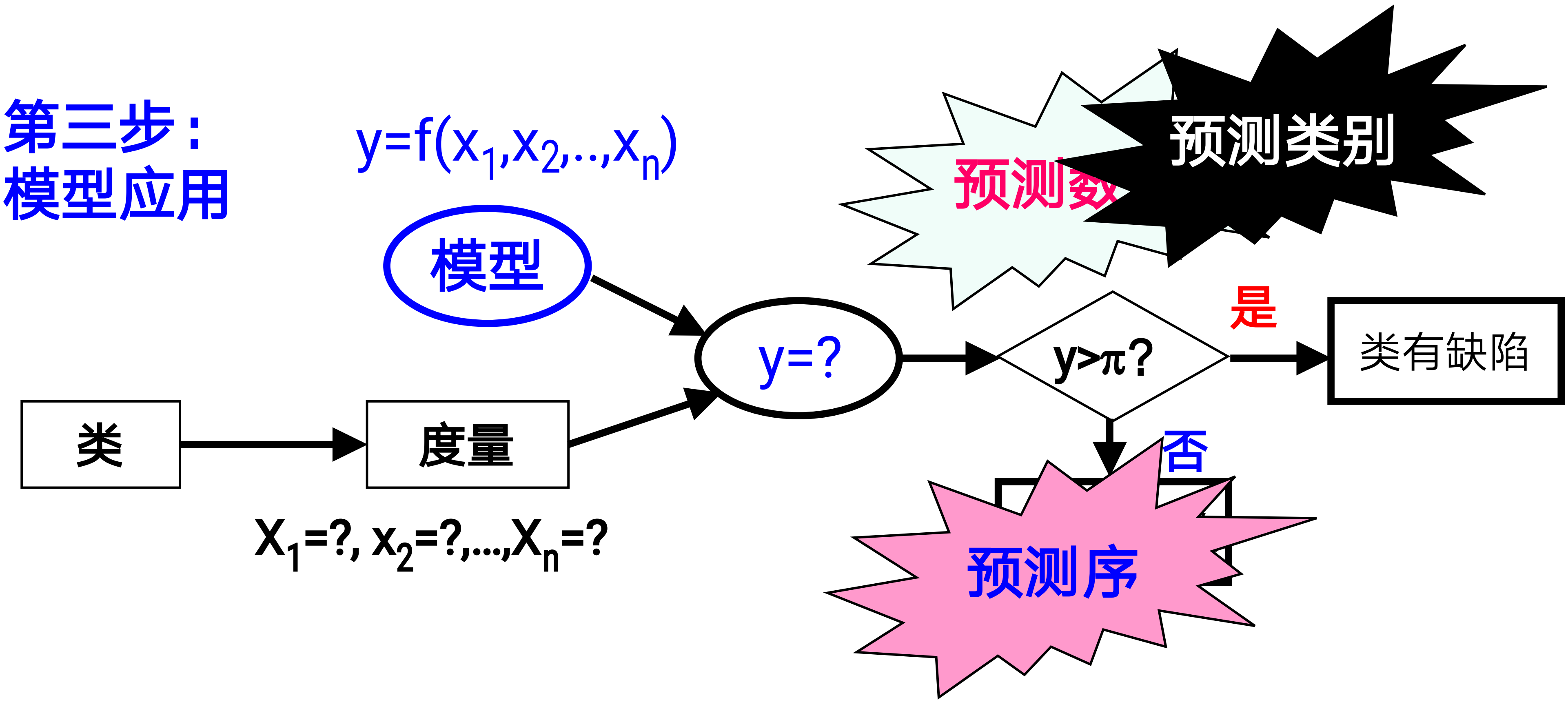


仅选择第1步中统计相关的度量建立多变量模型

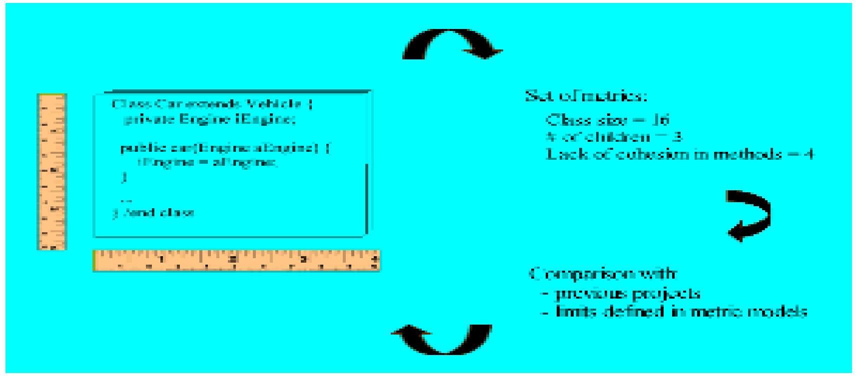
软件缺陷预测：方法

预测方法

第三步：
模型应用



软件缺陷预测：方法



(Copied from *staff.cs.utu.fi/opin**not**/kurssit/SemSE/05/SES-Malminen.ppt*)

No	WMC	LCOM	…	SLOC	Fault
1	35	96	…	1255	2
2	73	100		2828	0
…	…	…	…	…	…

No	WMC	LCOM	…	SLOC	Fault
1	15	9	…	135	?
2	45	10		282	?
…	…				

软件缺陷预测：关键点

预处理

0: 数据预处理

1: 数据分布检查

2: Outlier识别

模型构建

3: 单变量分析

4: 多变量分析

模型评价

5: 模型验证

6: 性能评价

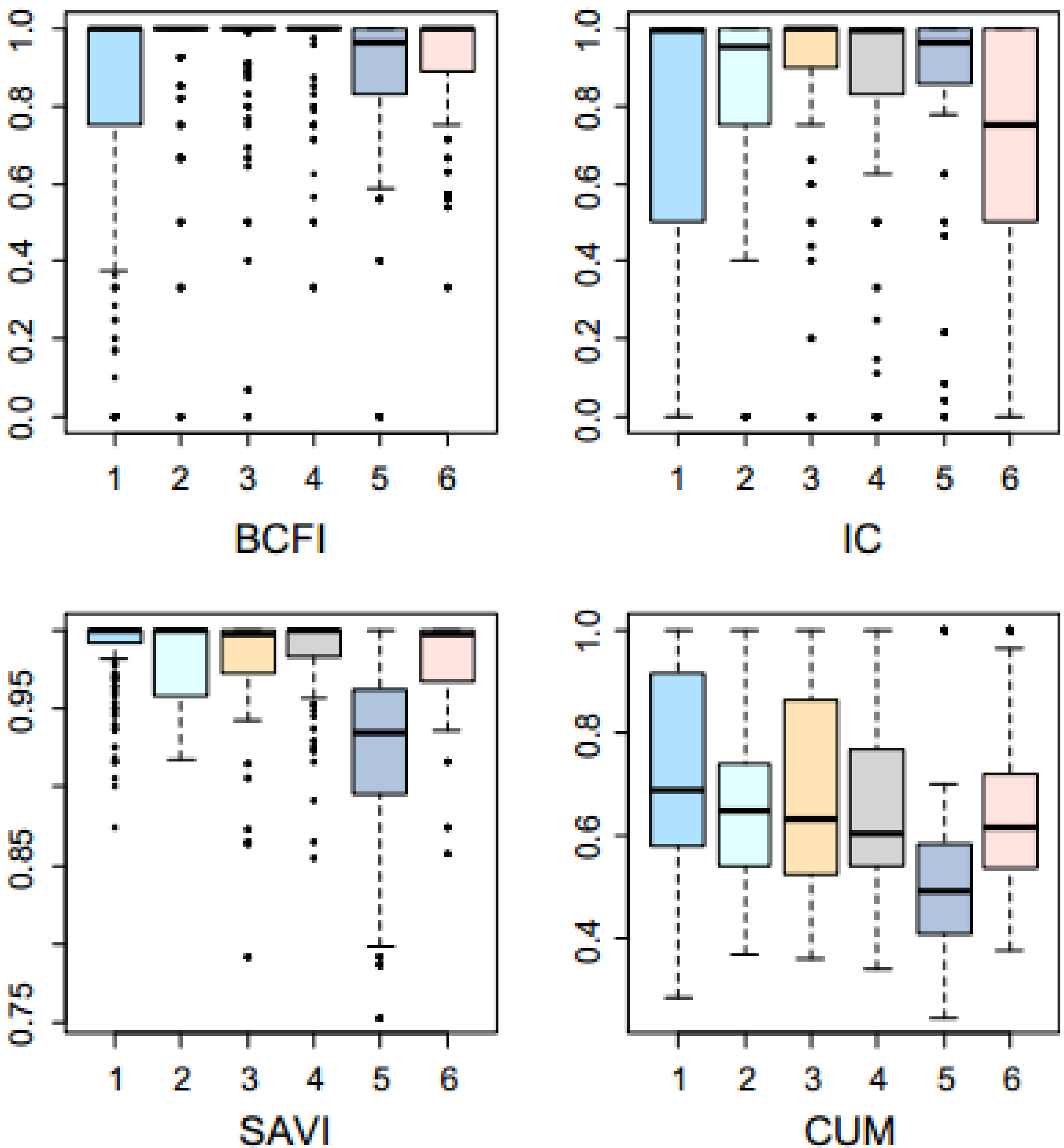
位置：25% + 50%+75%
离散：标准差
偏度 + 峰度
箱线图

数据分布检查举例

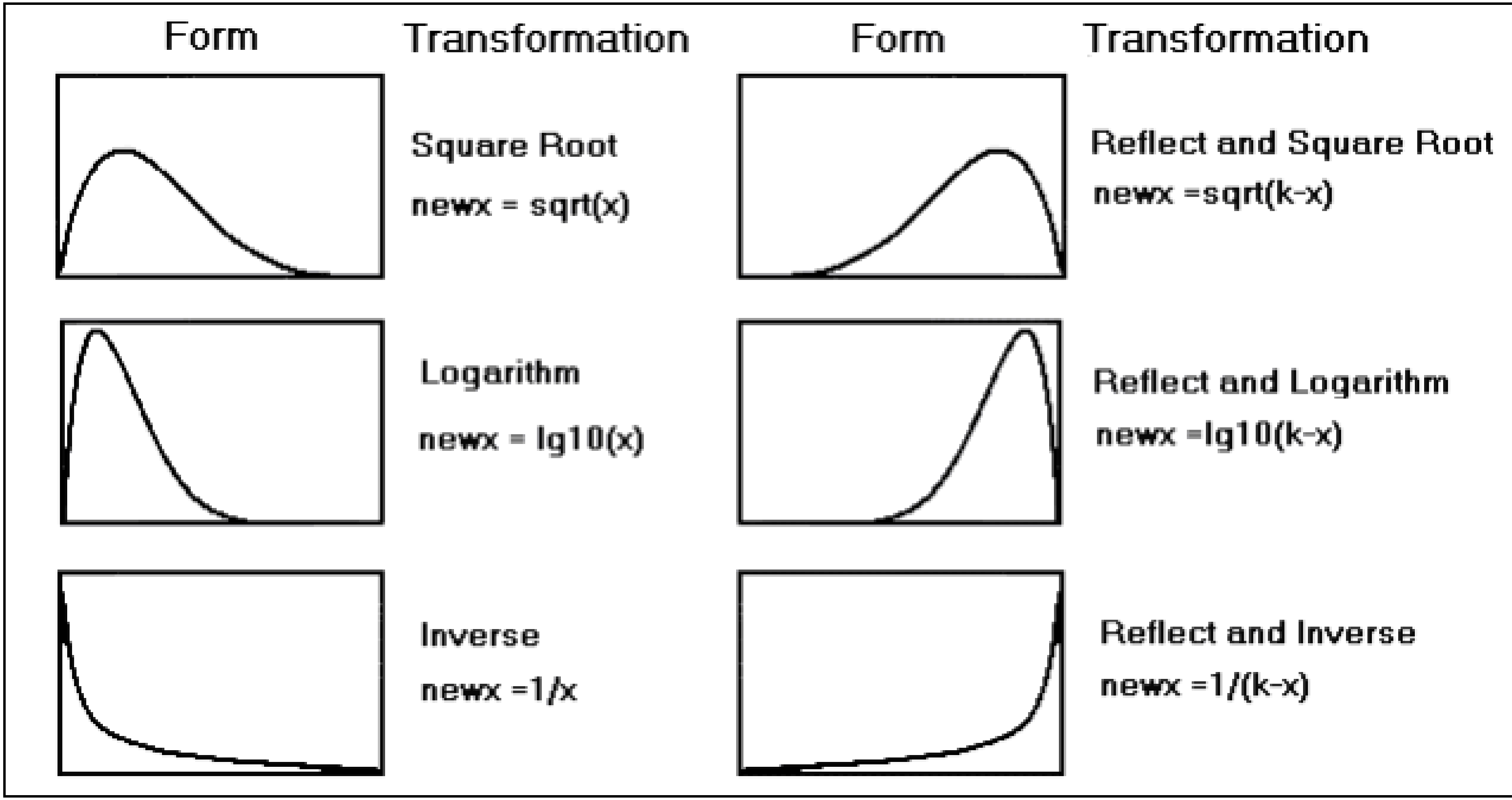
Metric	N	Max.	75%	Median	25%	Min.	Mean	Std. dev.	Skewness	Kurtosis
LCOM1	4830	171850	87	23	6	0	174.247	2615.348	59.225	3851.002
LCOM2	4830	166390	60	13	1	0	151.022	2508.804	60.757	3997.301
LCOM3	4830	492	7	4	2	1	6.250	11.641	19.300	674.415
LCOM4	4830	282	4	2	1	1	3.300	6.796	24.413	825.727
Co	4830	1	0.333	0.089	–0.017	–2	0.026	0.609	–1.195	2.946
Co'	4830	1	0.5	0.306	0.167	0	0.338	0.306	0.922	–0.149
LCOM5	3735	2	0.933	0.833	0.667	0	0.764	0.294	–0.603	2.360
Coh	3735	1	0.458	0.267	0.150	0	0.338	0.249	1.085	0.600
TCC	4417	1	1	0.5	0.167	0	0.503	0.410	0.071	–1.642
LCC	4417	1	1	0.672	0.2	0	0.562	0.425	–0.195	–1.688
ICH	4938	2976	17	4	0	0	16.115	73.573	22.495	722.809

Copied from: Yuming Zhou, et al. Examining the potentially confounding effect of class size on Associations between object-oriented metrics. IEEE Transactions on Software Engineering, 2009, 35(5): 607-623.

数据分布检查举例

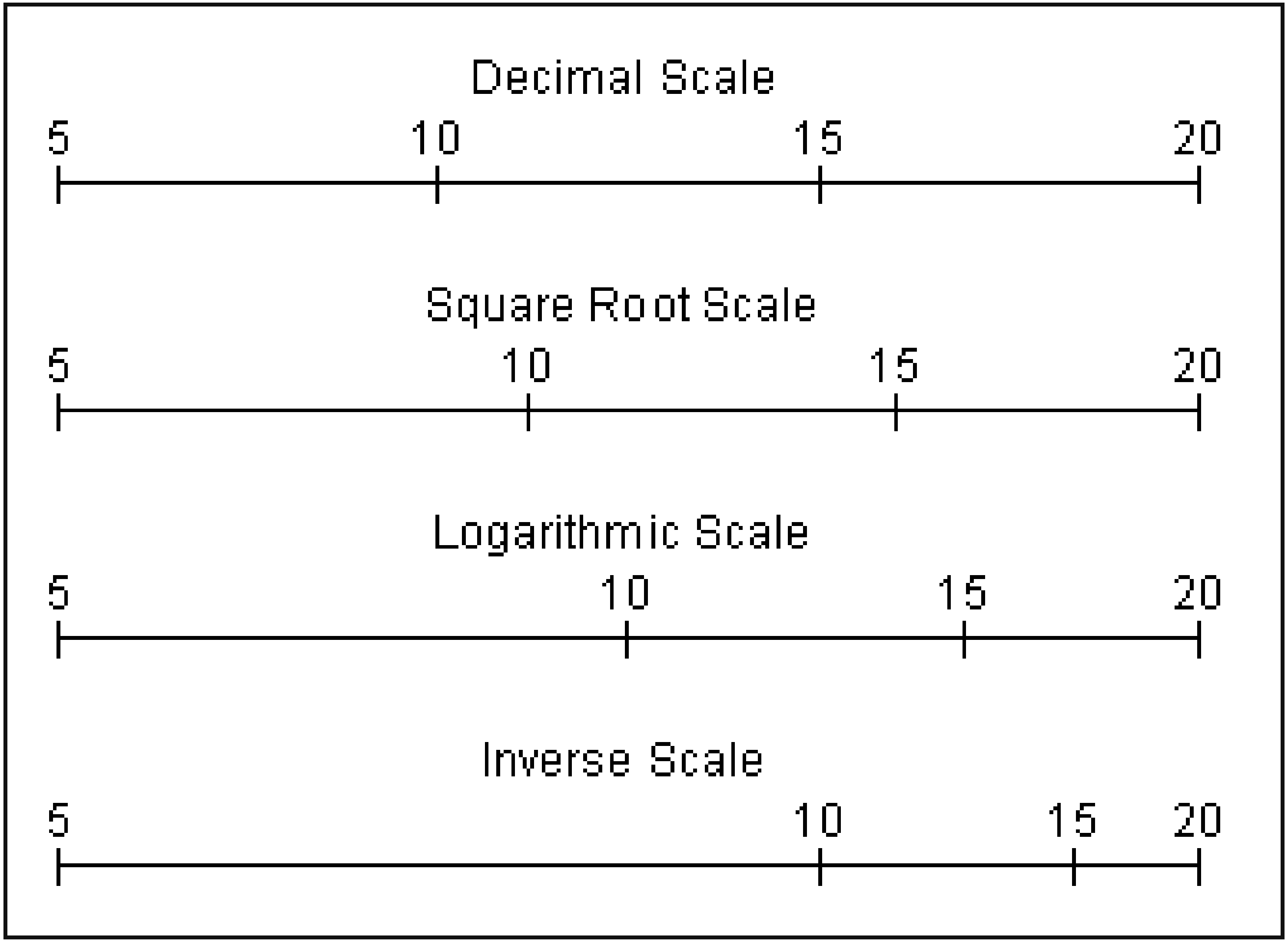


非正态分布：需要变换



The recommendation of which transform to use is often summarized in a pictorial chart like the above. In practice, it is difficult to determine which distribution is most like your variable. It is often more efficient to compute all transformations and examine the statistical properties of each.

非正态分布：需要变换



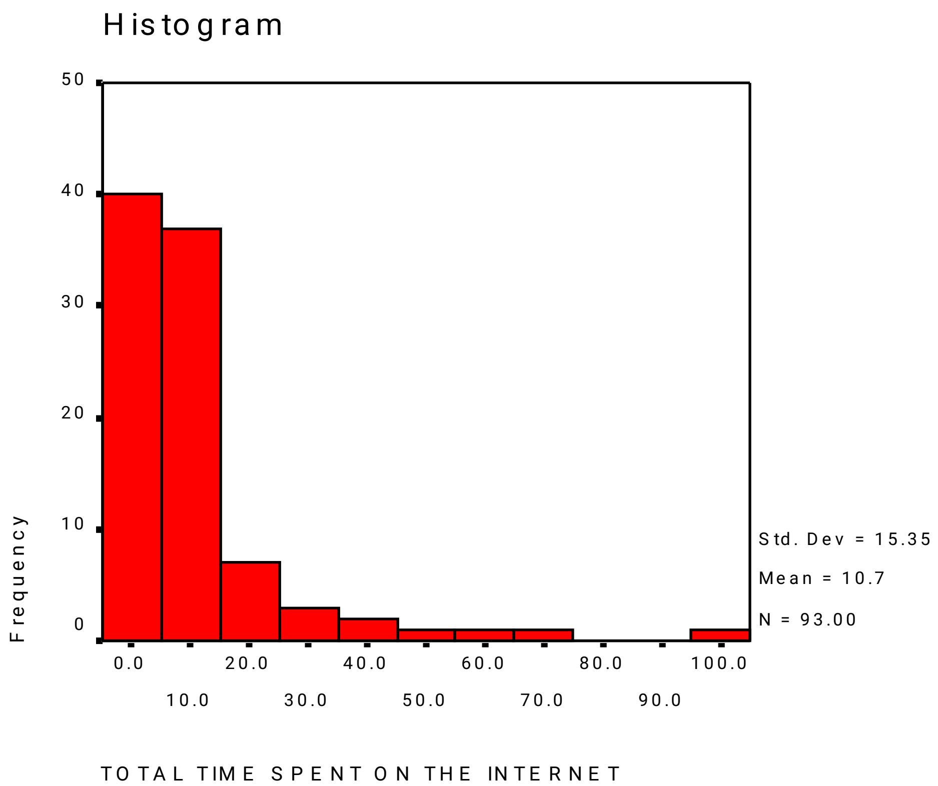
平方根变换

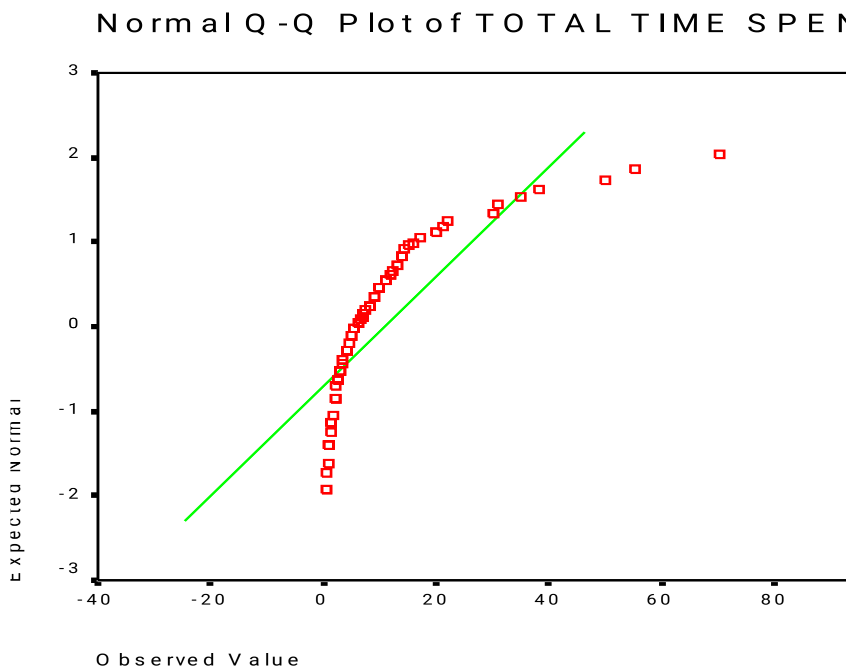
对数变换

倒数变换

Transformations:
Transformations for normality

Both the histogram and the normality plot for *Total Time Spent on the Internet* (netime) indicate that the variable is not normally distributed.





Transformations:
Determine whether reflection is required

Descriptives

			Statistic	Std. Error
TOTAL TIME SPENT ON THE INTERNET	Mean		10.73	1.59
	95% Confidence Interval for Mean	Lower Bound	7.57	
		Upper Bound	13.89	
	5% Trimmed Mean		8.29	
	Median		5.50	
	Variance		235.655	
	Std. Deviation		15.35	
	Minimum		0	
	Maximum		102	
	Range		102	
	Interquartile Range		10.20	
	Skewness		3.532	.250
	Kurtosis		[illegible]14	.495

Skewness, in the table of Descriptive Statistics, indicates whether or not reflection (reversing the values) is required in the transformation.

If Skewness is positive, as it is in this problem, reflection is not required. If Skewness is negative, reflection is required.

Transformations:
Compute the adjustment to the argument

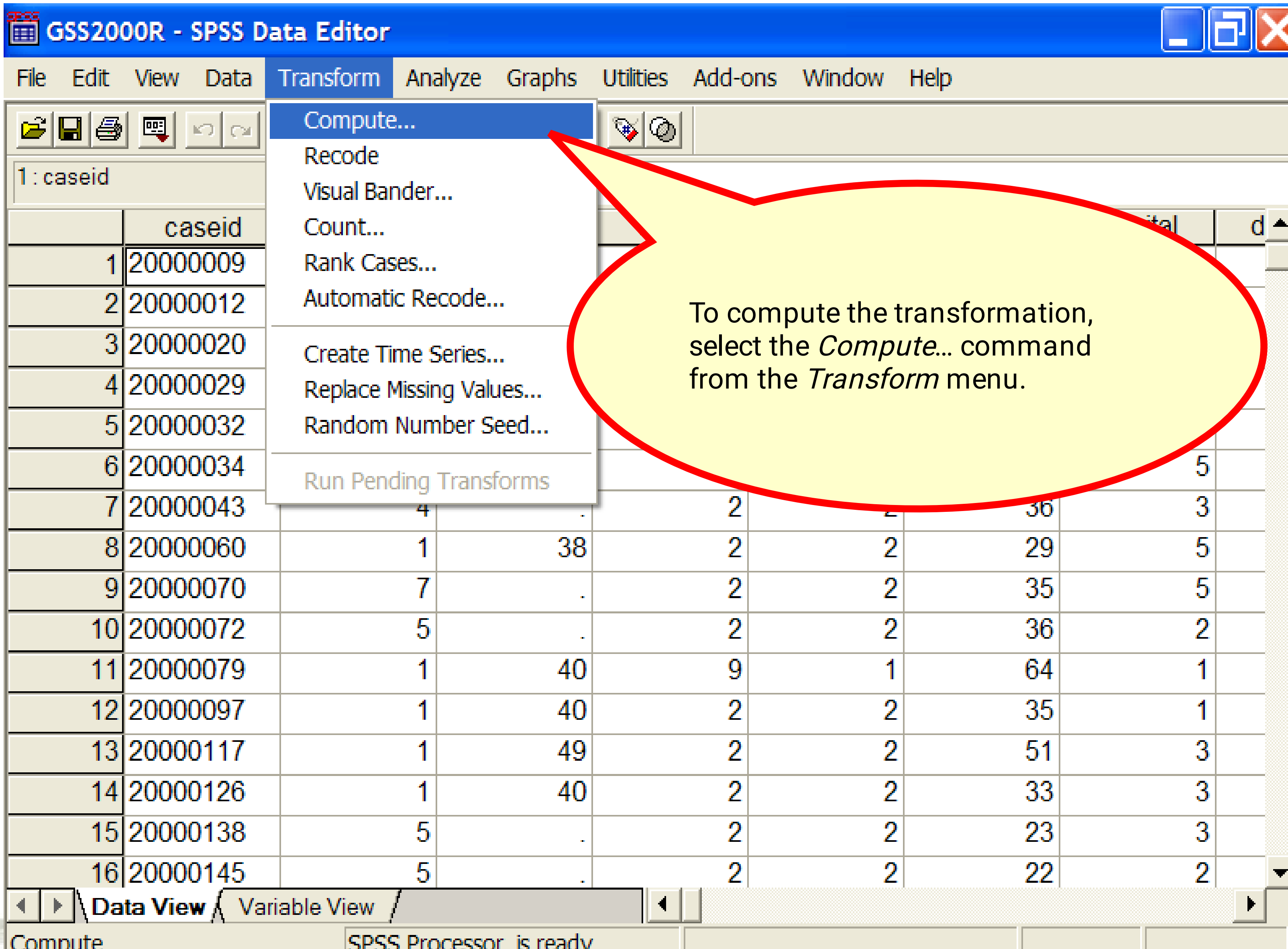
Descriptives

			Statistic	Std. Error
TOTAL TIME SPENT ON THE INTERNET	Mean		10.73	1.59
	95% Confidence Interval for Mean	Lower Bound	7.57	
		Upper Bound	13.89	
	5% Trimmed Mean		8.29	
	Median		5.50	
	Variance		235.655	
	Std. Deviation		15.35	
	Minimum		0	
	Maximum		102	
	Range		102	
	Interquartile Range		10.20	
	Skewness		3.532	.250
	Kurtosis		15.614	.495

In this problem, the minimum value is 0, so 1 will be added to each value in the formula, i.e. the argument to the SPSS functions and formula for the inverse will be:

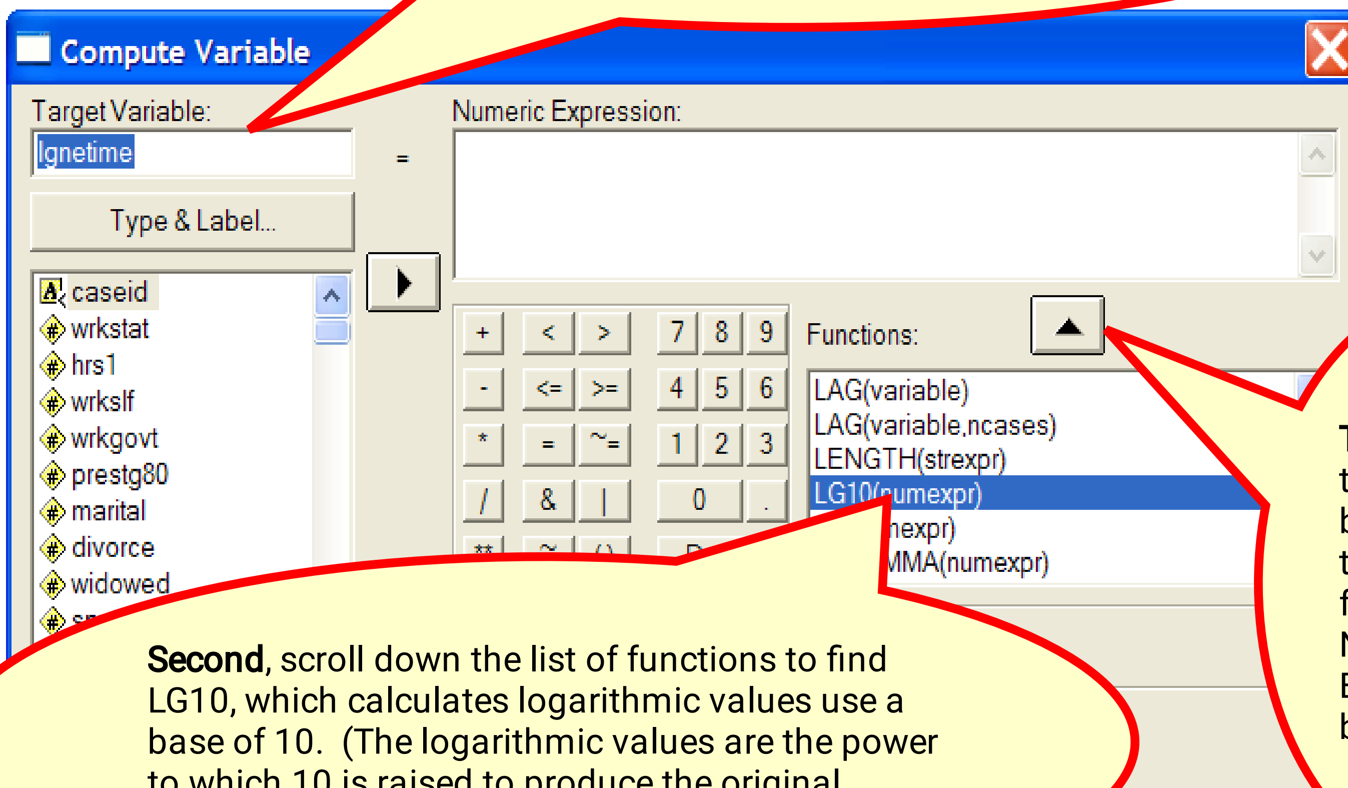
netime + 1.

Transformations:
Computing the logarithmic transformation



Transformations:
Specifying the transform variable name and function

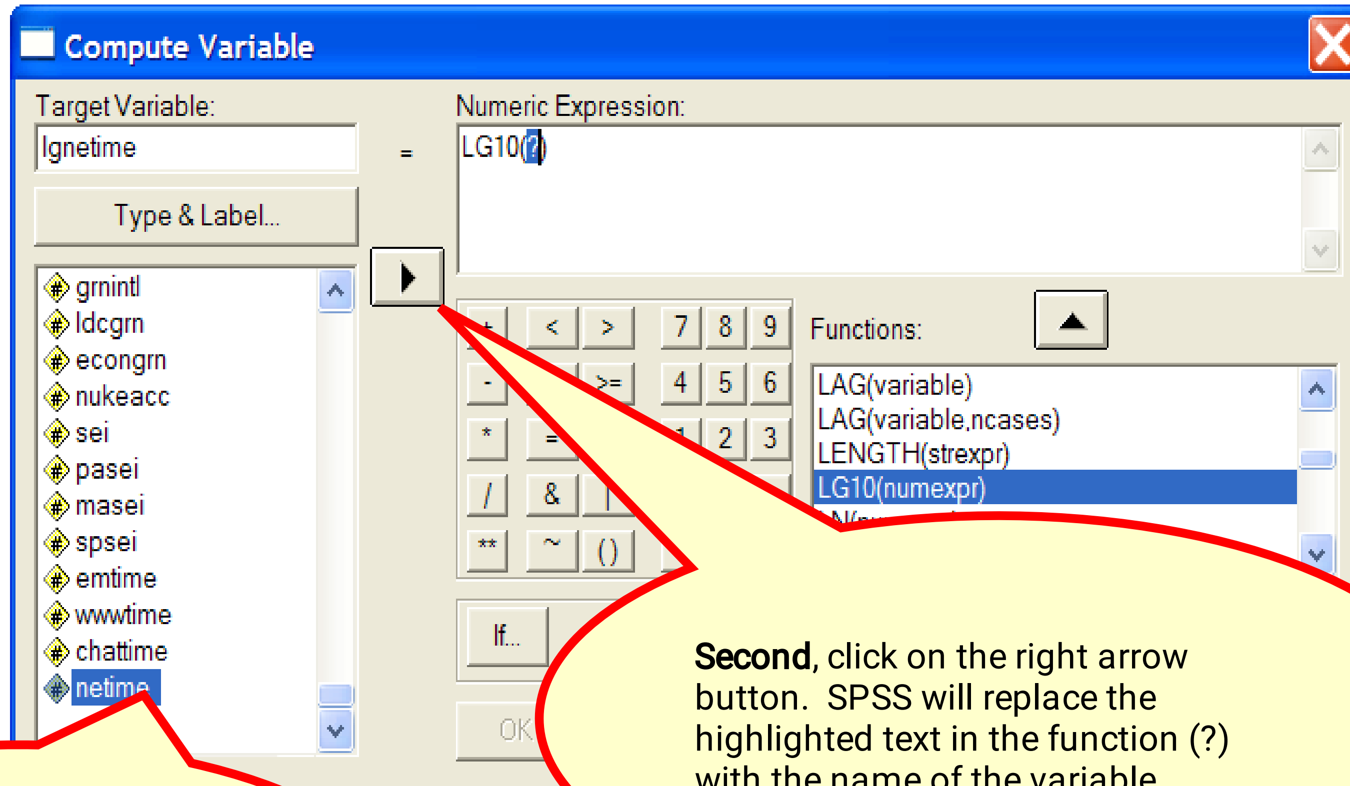
First, in the *Target Variable* text box, type a name for the log transformation variable, e.g. “lgnetime“.



Second, scroll down the list of functions to find LG10, which calculates logarithmic values use a base of 10. (The logarithmic values are the power to which 10 is raised to produce the original number.)

Third, click on the up arrow button to move the highlighted function to the Numeric Expression text box.

Transformations:
Adding the variable name to the function

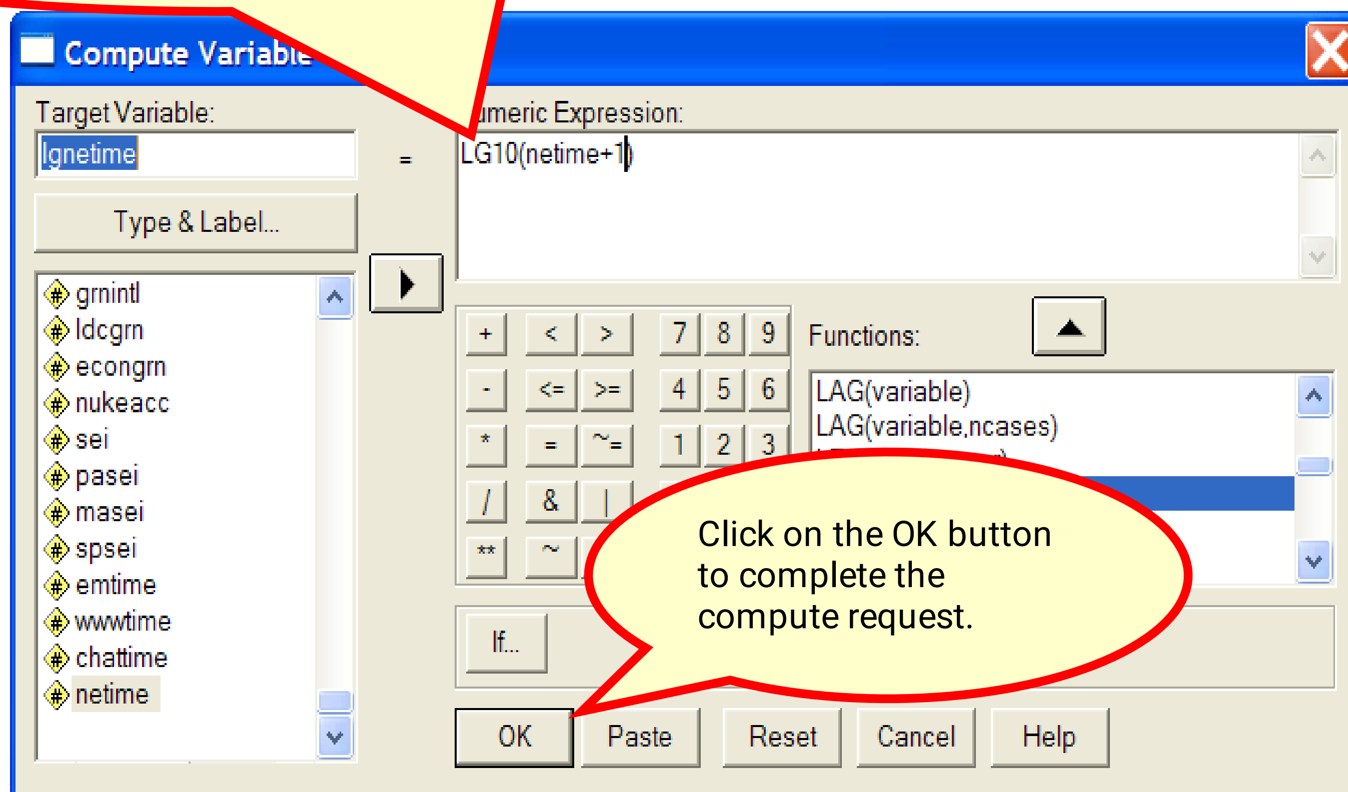


Second, click on the right arrow button. SPSS will replace the highlighted text in the function (?) with the name of the variable.

First, scroll down the list of variables to locate the variable we want to transform. Click on its name so that it is highlighted.

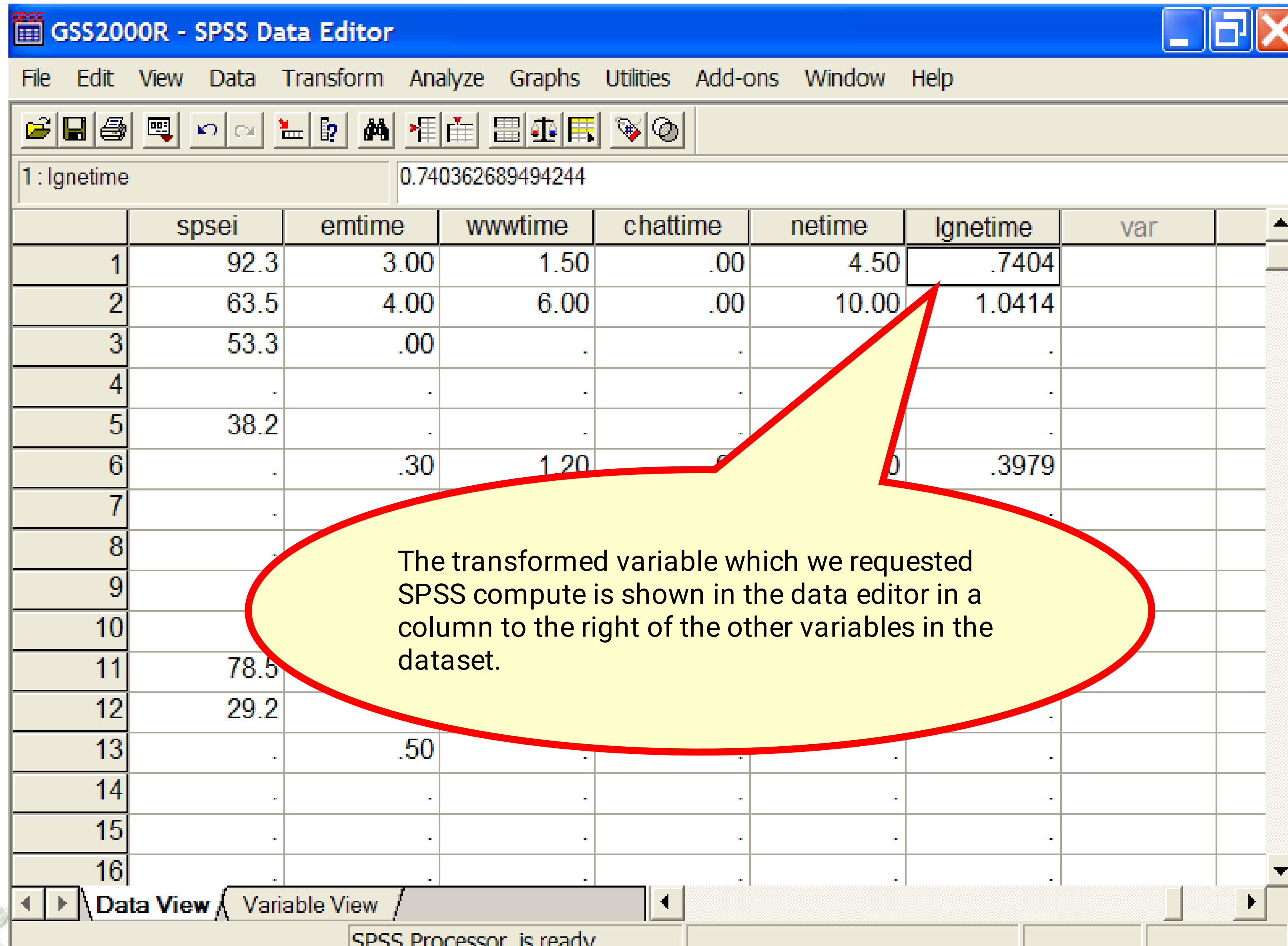
Transformations: Adding the constant to the function

Following the rules stated for determining the constant that needs to be included in the function either to prevent mathematical errors, or to do reflection, we include the constant in the function argument. In this case, we add 1 to the netime variable.

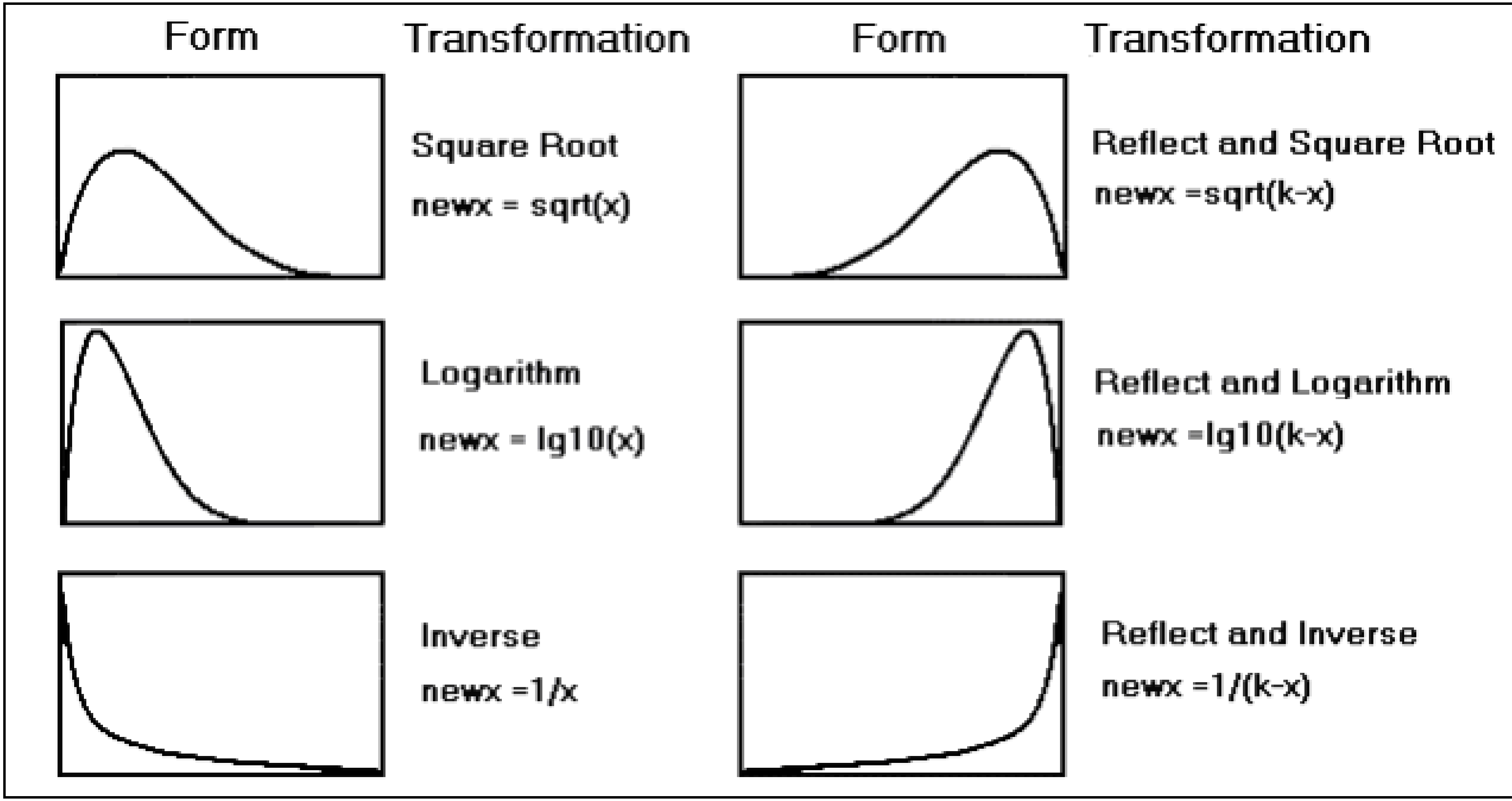


Click on the OK button to complete the compute request.

Transformations:
The transformed variable



非正态分布→正态分布：变换方法



软件缺陷预测：关键点

预处理
- 0: 数据预处理
- 1: 数据分布检查
- 2: Outlier识别

模型构建
- 3: 单变量分析
- 4: 多变量分析

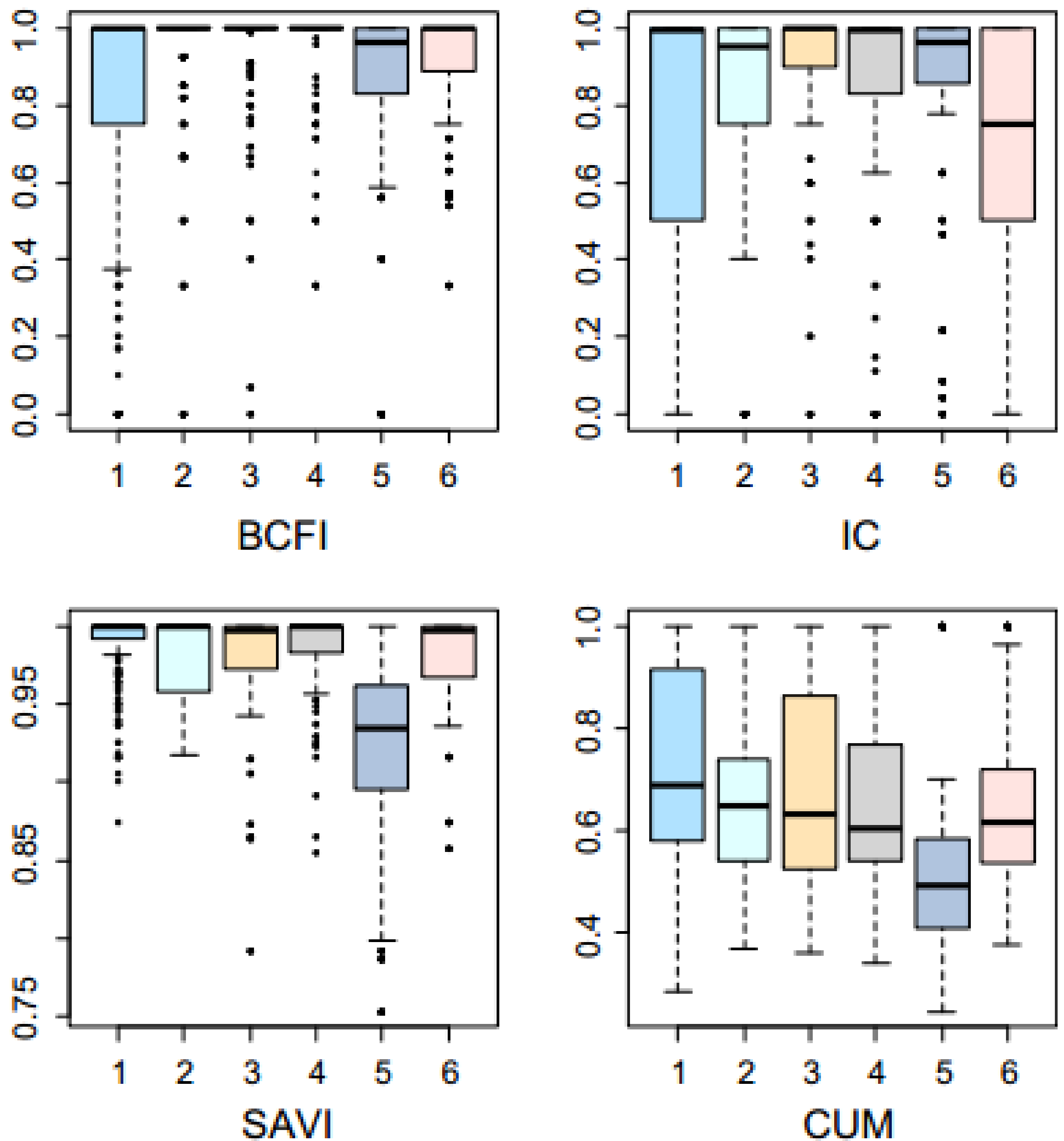
模型评价
- 5: 模型验证
- 6: 性能评价

单变量: standardized score

多变量: Mahalanobis D^2

有影响的: Cook D

单变量的outlier识别



单变量的outlier识别

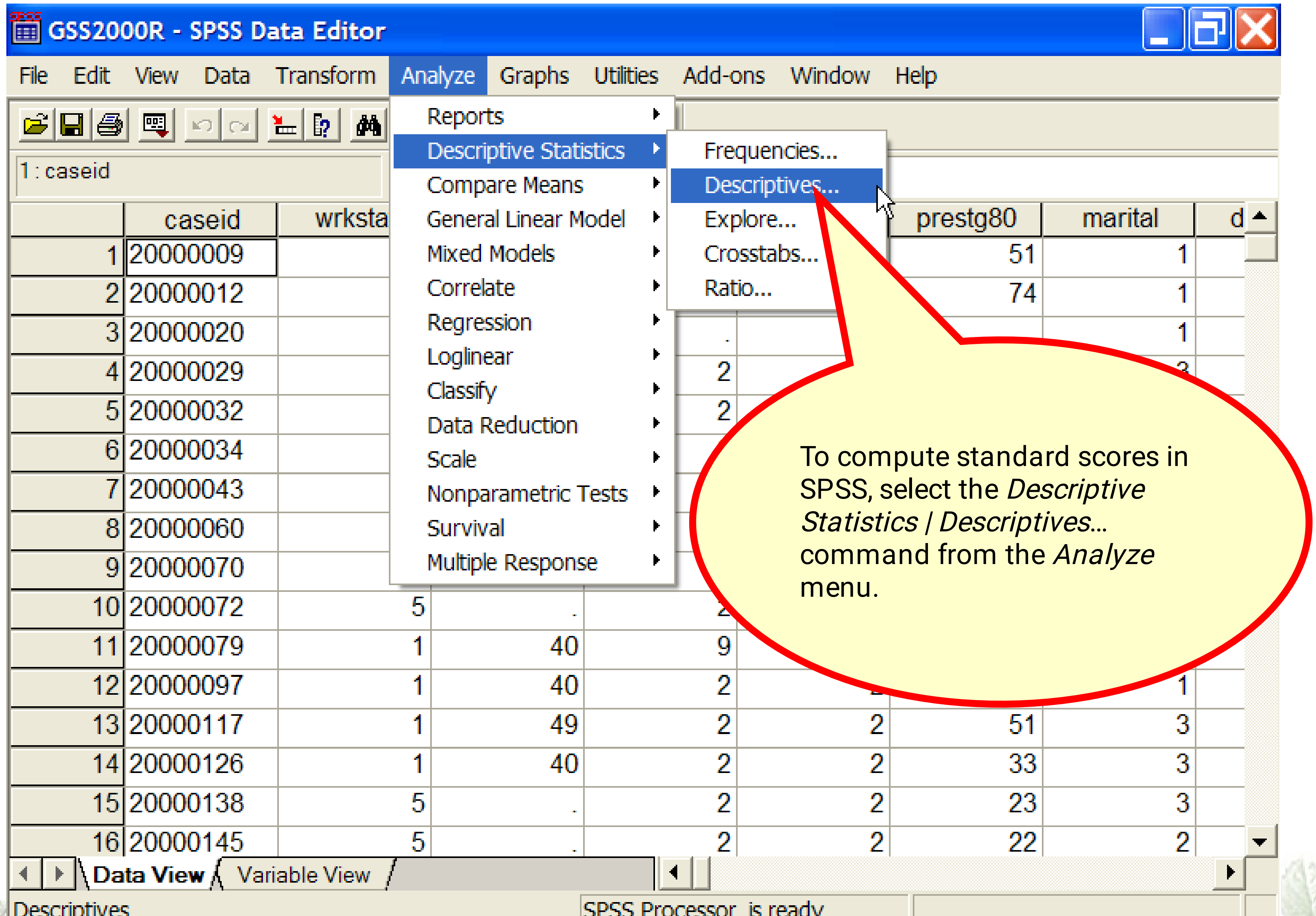
- One way to identify univariate outliers is to convert all of the scores for a variable to standard scores

- If the sample size is small (80 or fewer cases), a case is an outlier if its standard score is ±2.5 or beyond

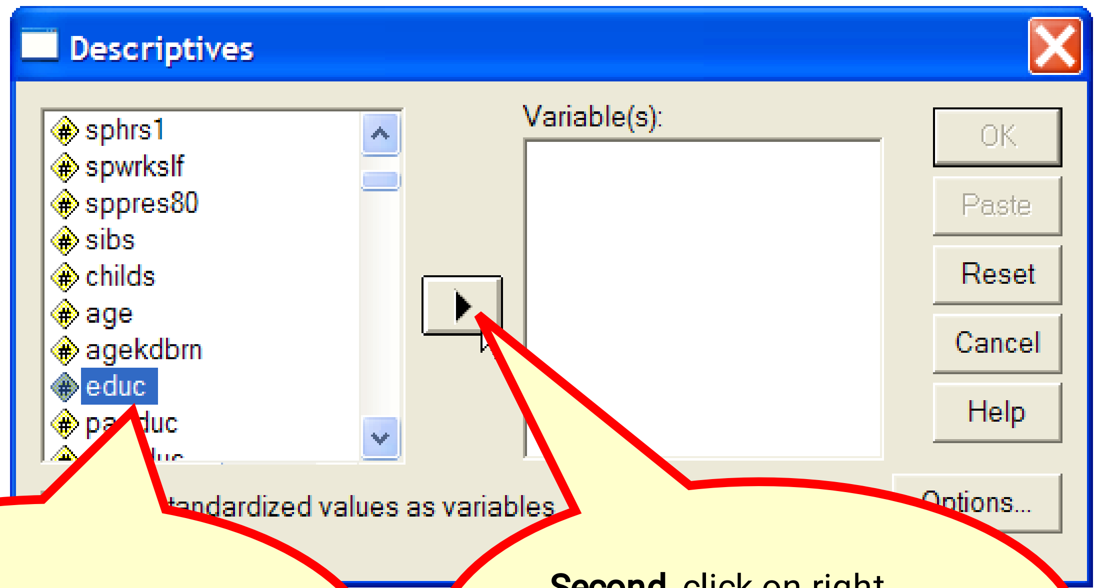
- If the sample size is larger than 80 cases, a case is an outlier if its standard score is ±3.0 or beyond

$$z_i = \frac{X_i - \bar{X}}{s}$$

Descriptive statistics compute standard scores



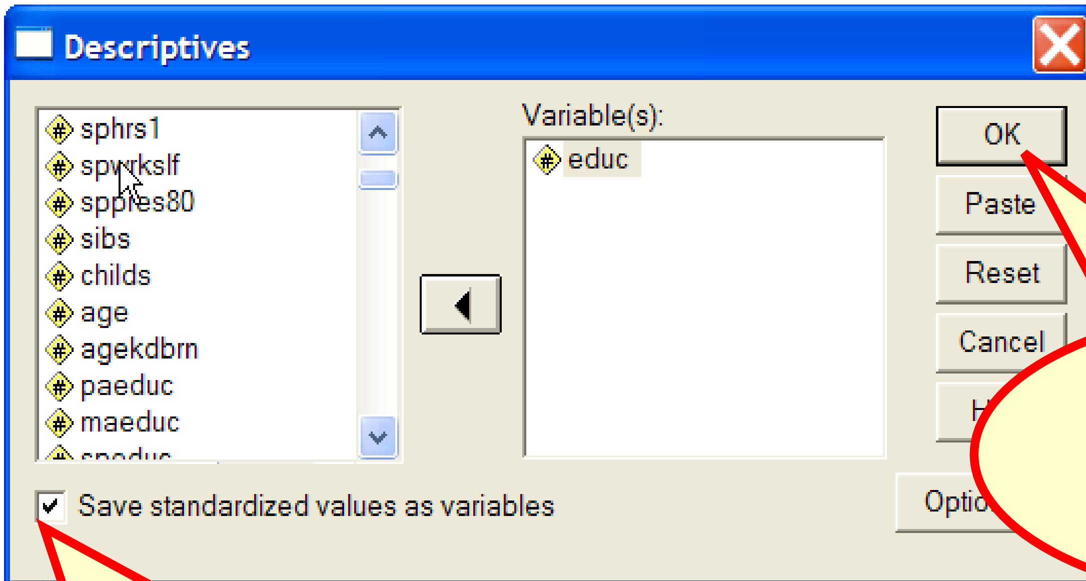
Select the variable(s) for the analysis



First, click on the variable to be included in the analysis to highlight it.

Second, click on right arrow button to move the highlighted variable to the list of variables.

Mark the option for computing standard scores

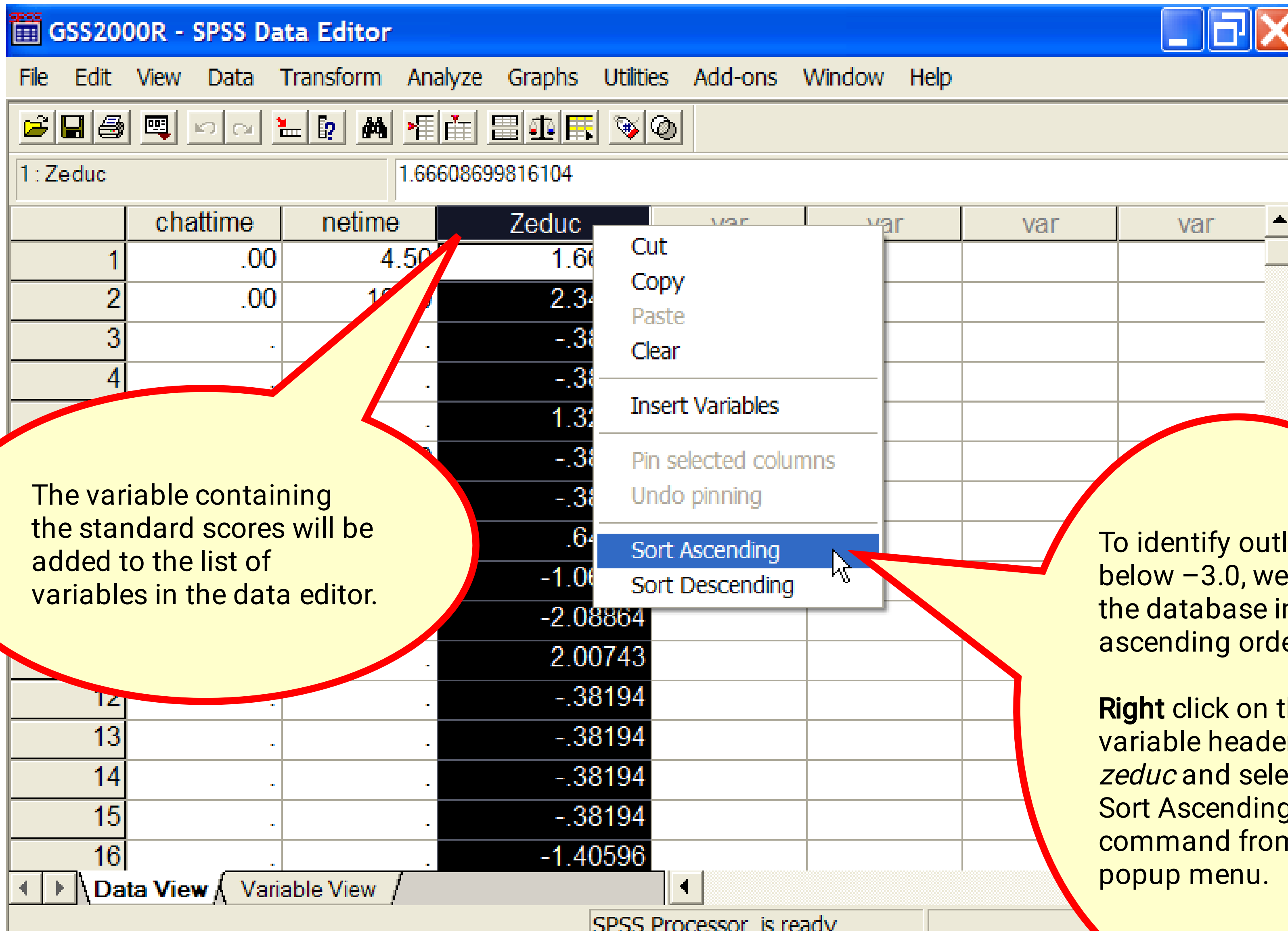


Second, click on the OK button to complete the analysis request.

First, click on the checkbox to save standard score values as a new variable in the dataset.

The new variable will have the letter z prepended to its name, e.g. the standard score variable for “educ” will be “zeduc”.

The z-score variable in the data editor

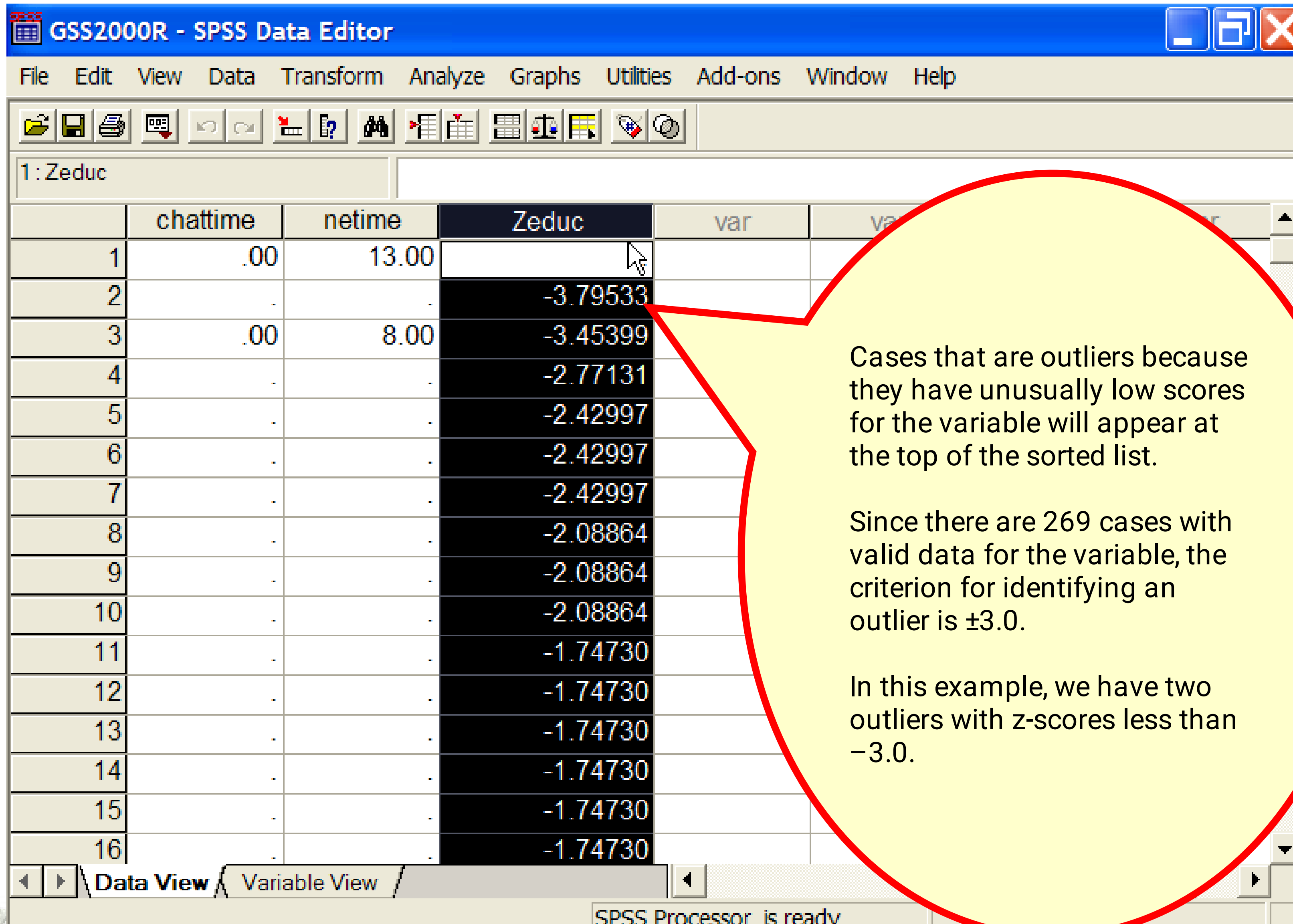


The variable containing the standard scores will be added to the list of variables in the data editor.

To identify outliers below −3.0, we sort the database in ascending order.

Right click on the variable header *zeduc* and select the Sort Ascending command from the popup menu.

Outliers with unusually low scores



Cases that are outliers because they have unusually low scores for the variable will appear at the top of the sorted list.

Since there are 269 cases with valid data for the variable, the criterion for identifying an outlier is ±3.0.

In this example, we have two outliers with z-scores less than −3.0.

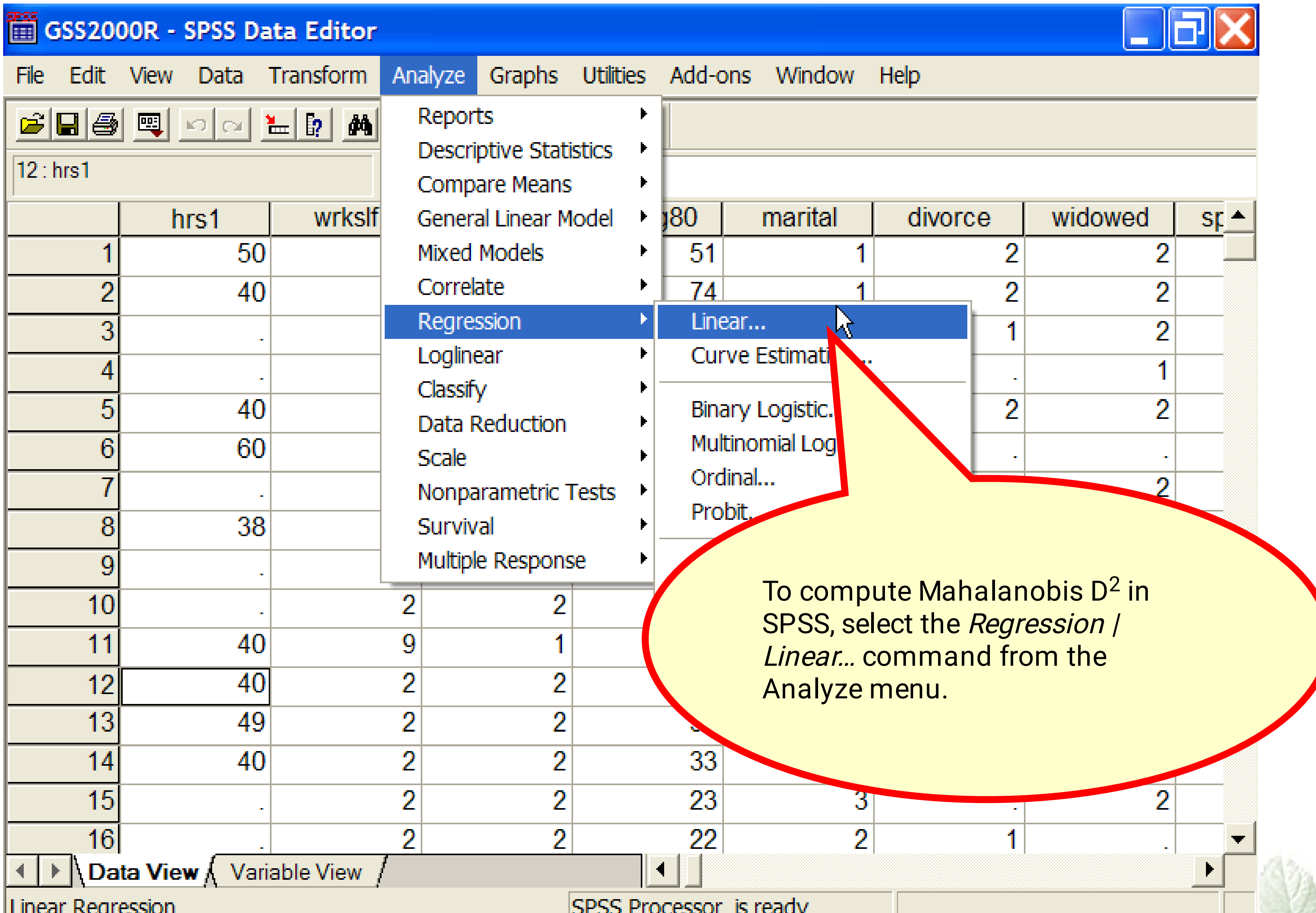
多变量的outlier识别

- Mahalanobis D^2 is a multidimensional version of a z-score. It measures the distance of a case from the centroid (multidimensional mean) of a distribution, given the covariance (multidimensional variance) of the distribution

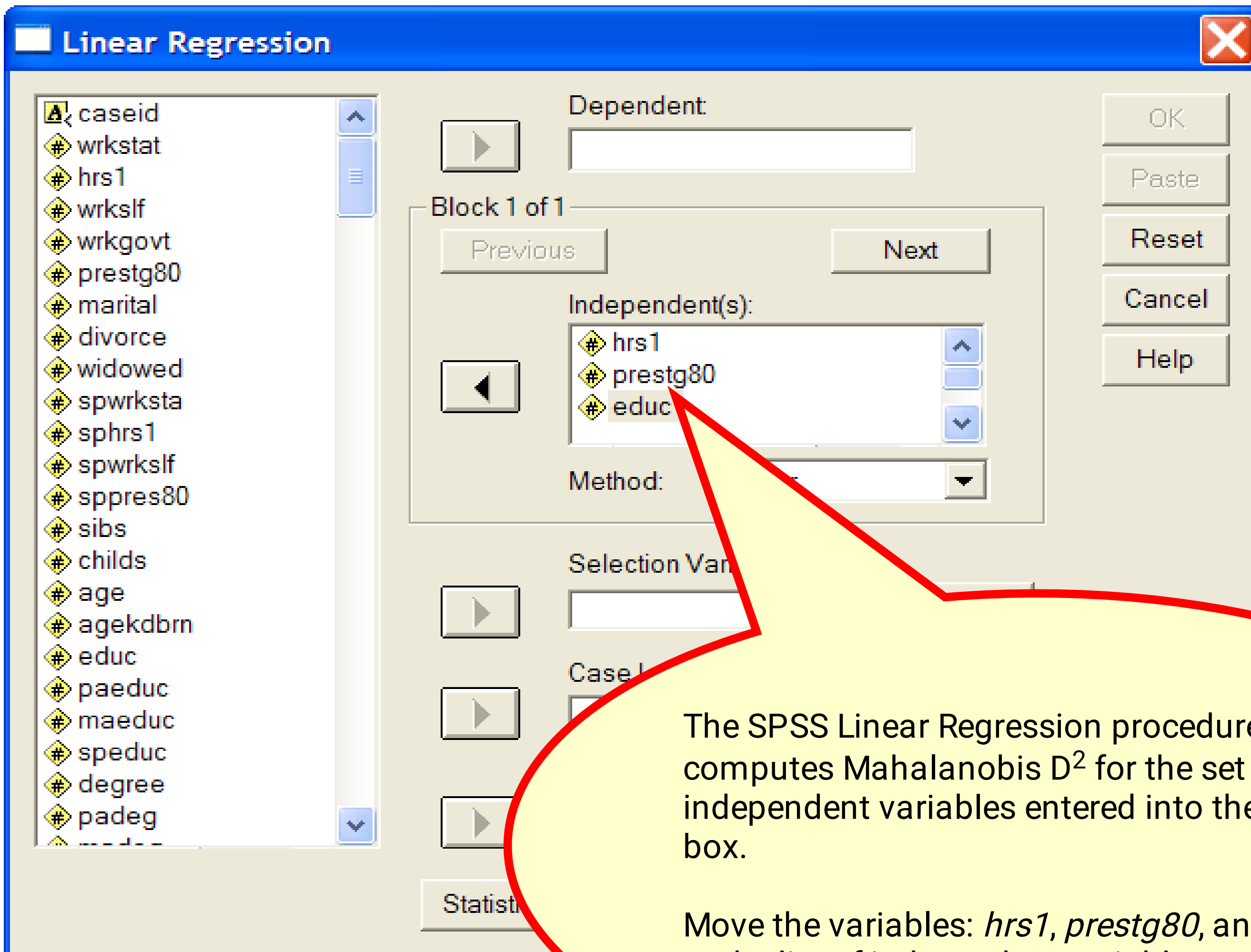
- A case is a multivariate outlier if the probability associated with its D^2 is 0.001 or less. D^2 follows a chi-square distribution with degrees of freedom equal to the number of variables included in the calculation

$$D_i^2 = \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)' \mathbf{S}^{-1} \left(\mathbf{X}_i - \bar{\mathbf{X}}\right)$$

Mahalanobis D^2 is computed by Regression



Adding the independent variables

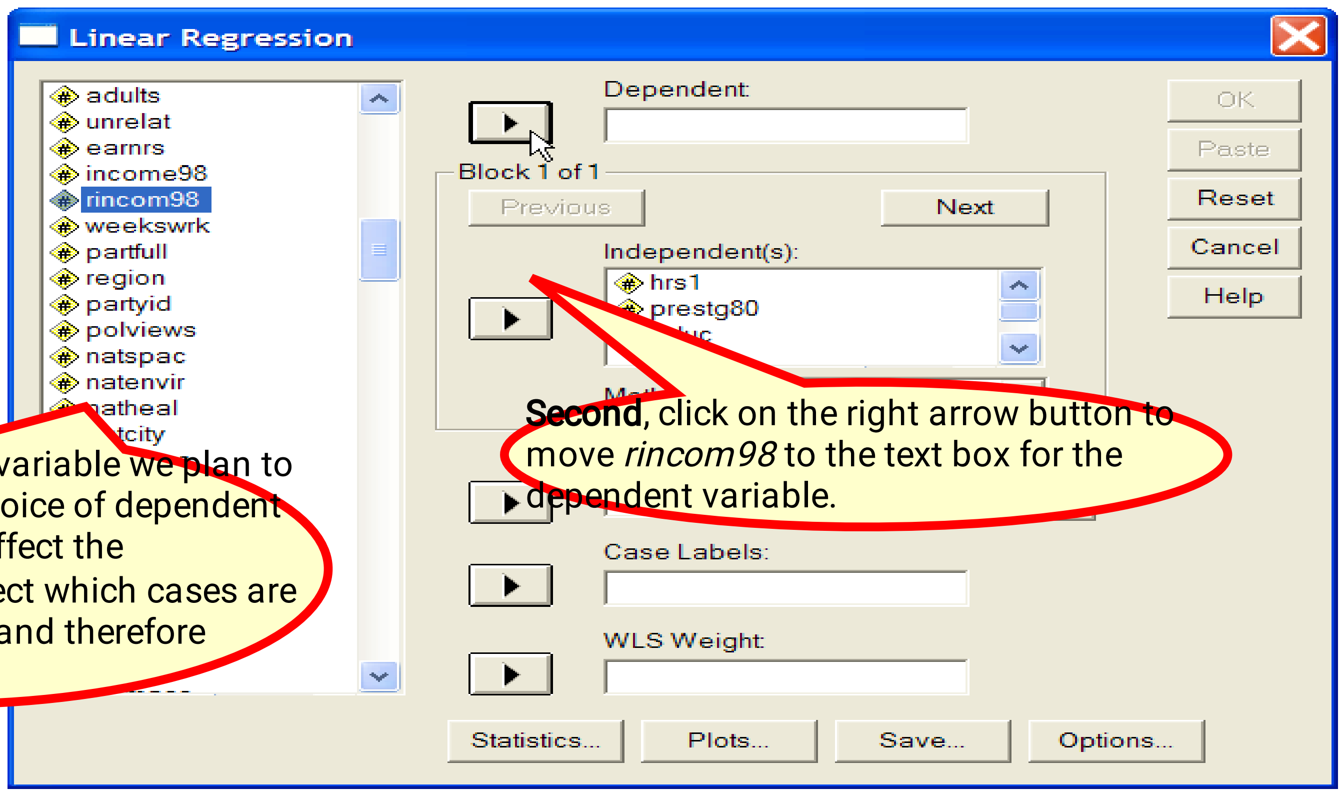


The SPSS Linear Regression procedure computes Mahalanobis D^2 for the set of independent variables entered into the dialog box.

Move the variables: *hrs1*, *prestg80*, and *educ* to the list of independent variables.

Adding the dependent variable

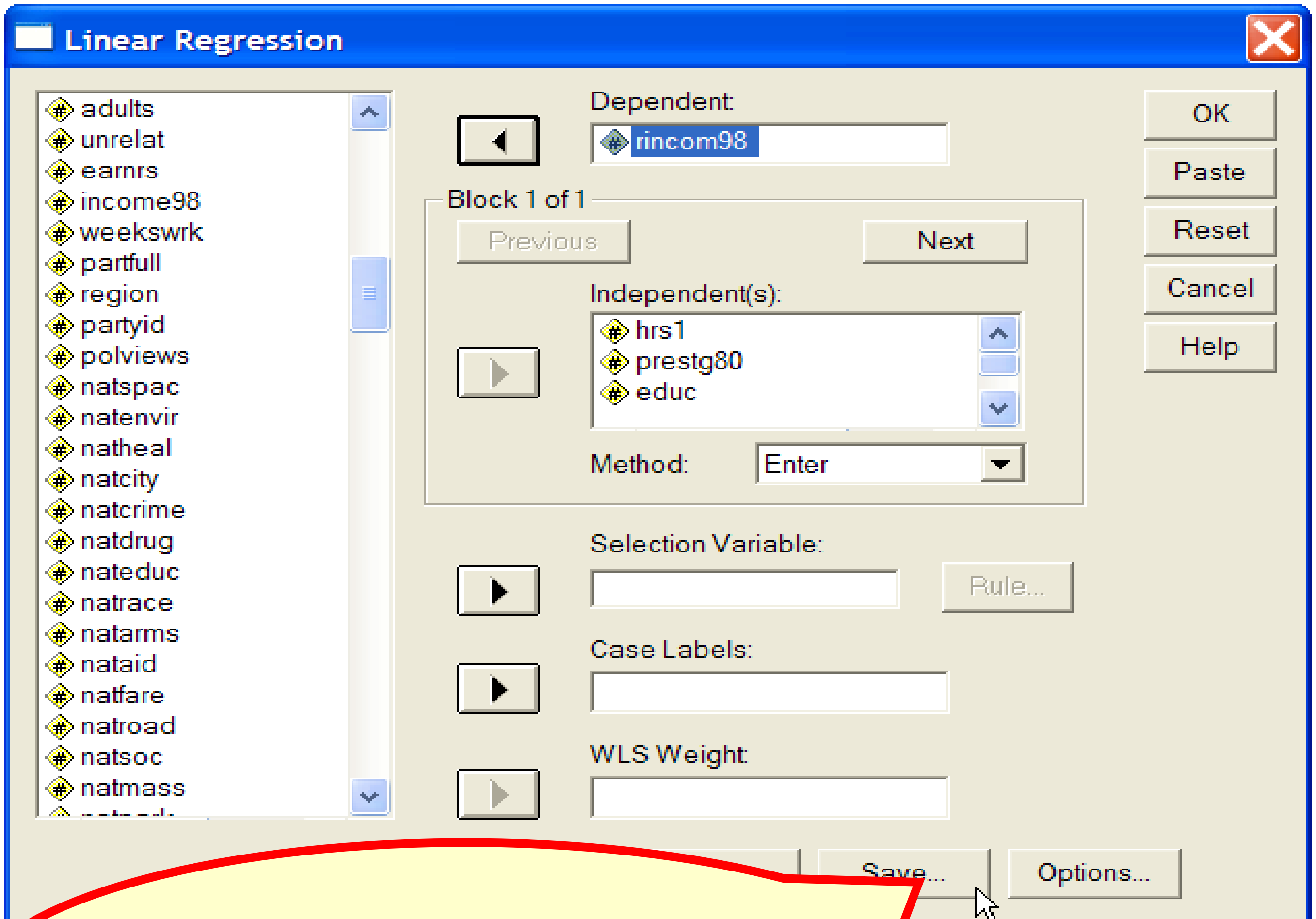
Though the test of multivariate outliers is performed on the independent variables, the Linear Regression procedure requires that a dependent variable be specified.



Second, click on the right arrow button to move *rincom98* to the text box for the dependent variable.

First, select the dependent variable we plan to use in the analysis. The choice of dependent variables will not directly affect the calculation of D^2, it will affect which cases are omitted as missing cases, and therefore indirectly affect the results.

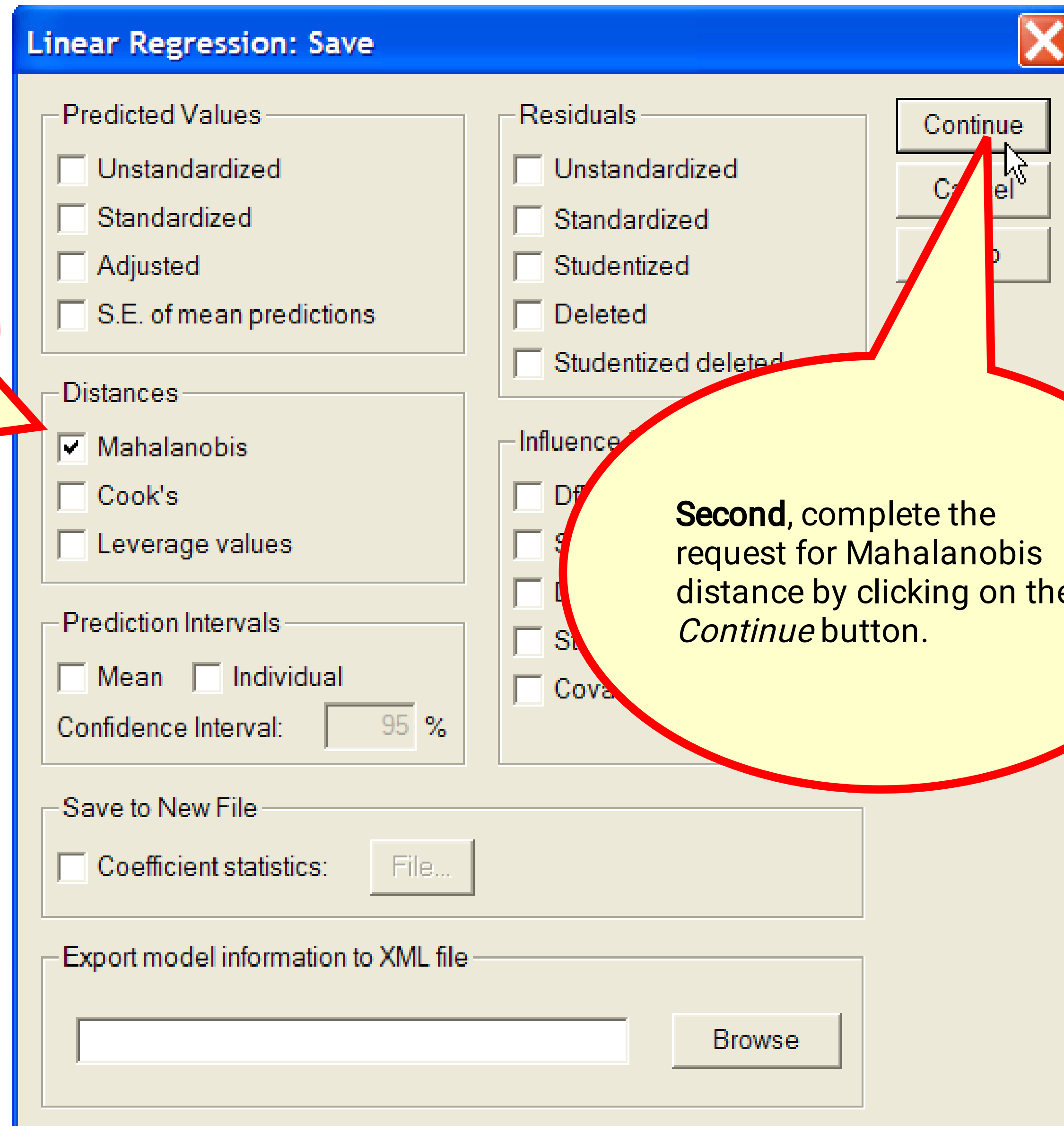
Adding Mahalanobis D^2 to the dataset



To request that SPSS add the value of Mahalanobis D^2 to the data set, click on the *Save* button to open the save dialog box.

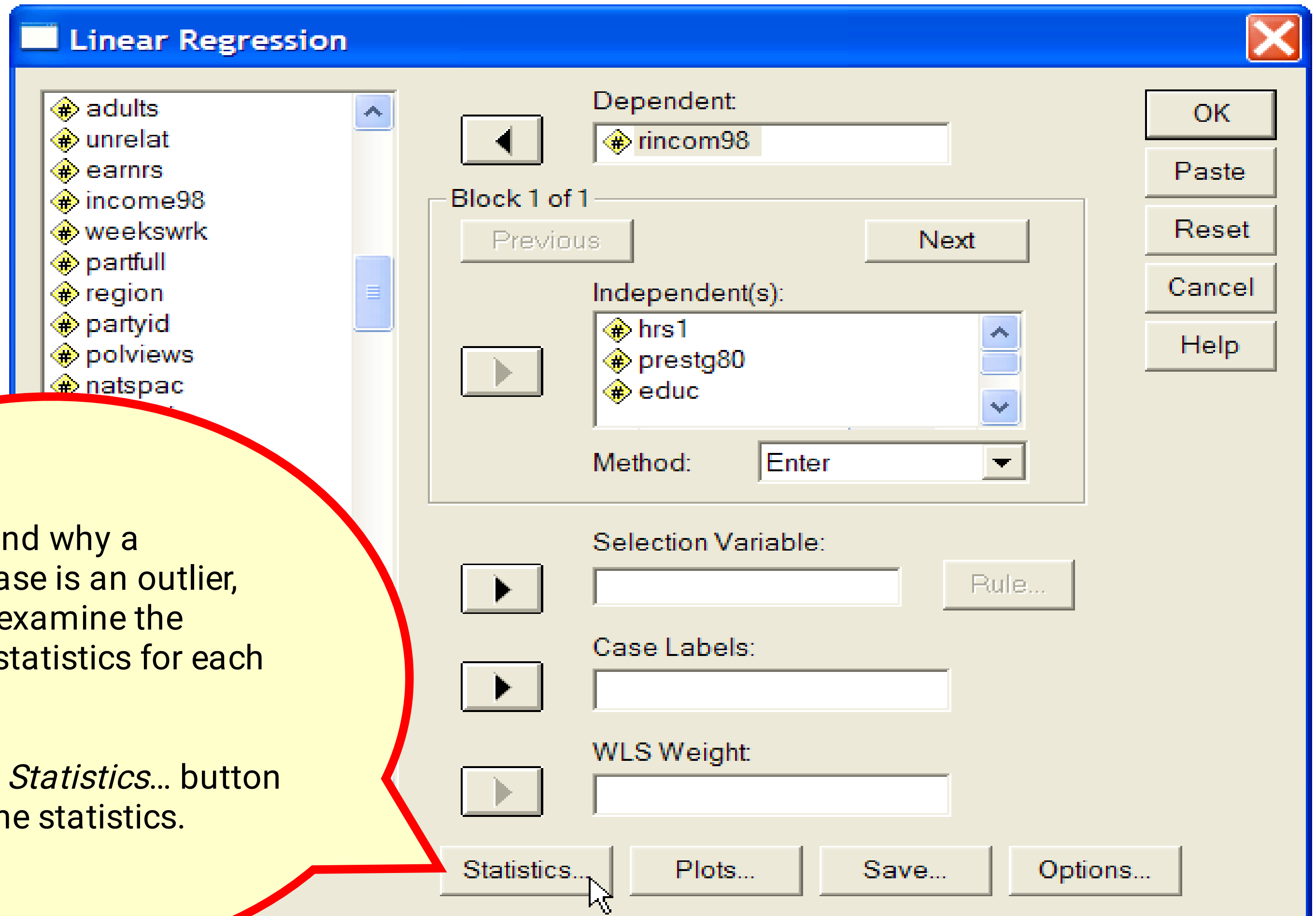
Specify saving Mahalanobis D^2 distance

First, mark the checkbox for *Mahalanobis* in the Distances panel. All other checkboxes can be unchecked.



Second, complete the request for Mahalanobis distance by clicking on the *Continue* button.

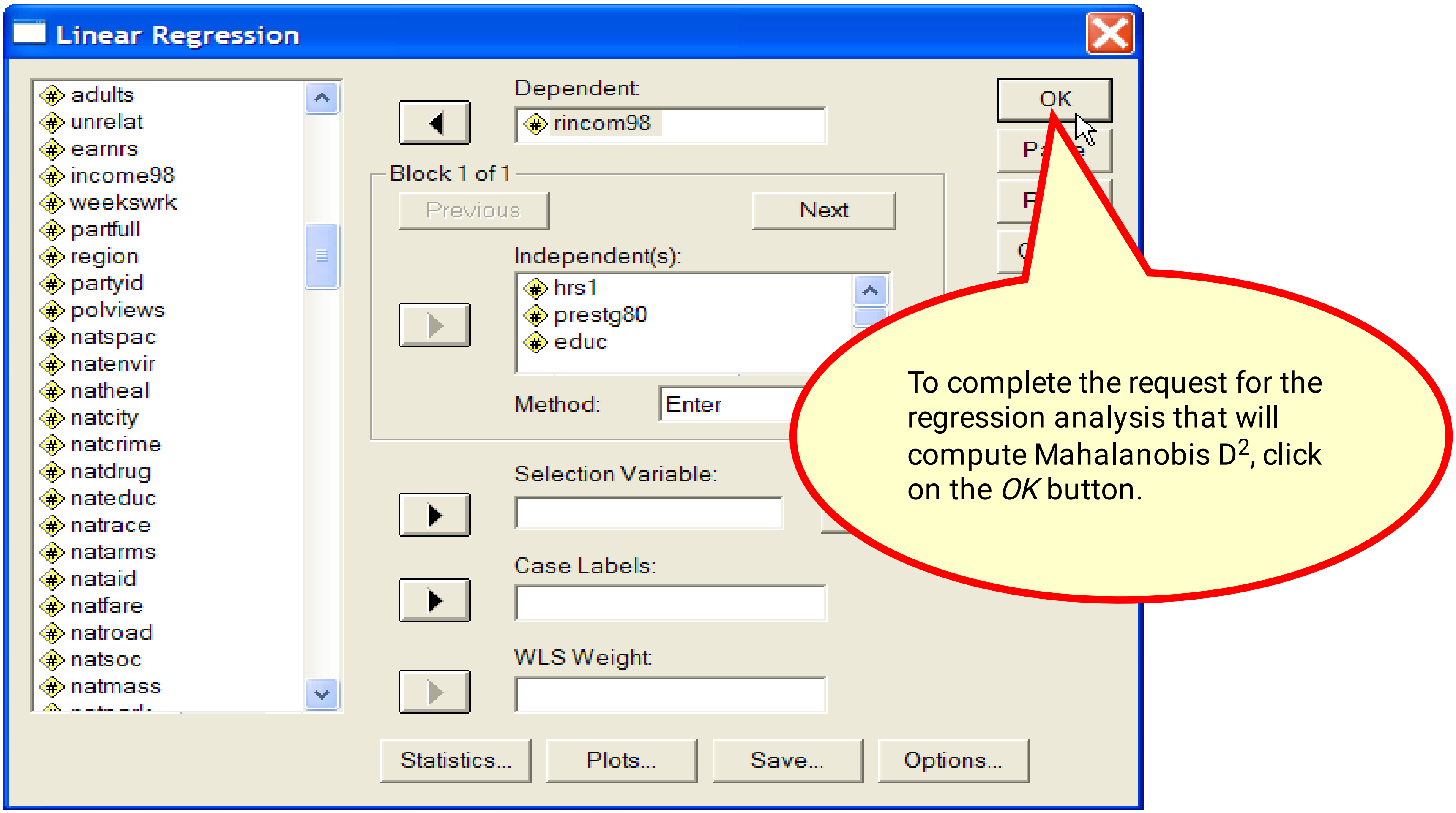
Specify the statistics output needed



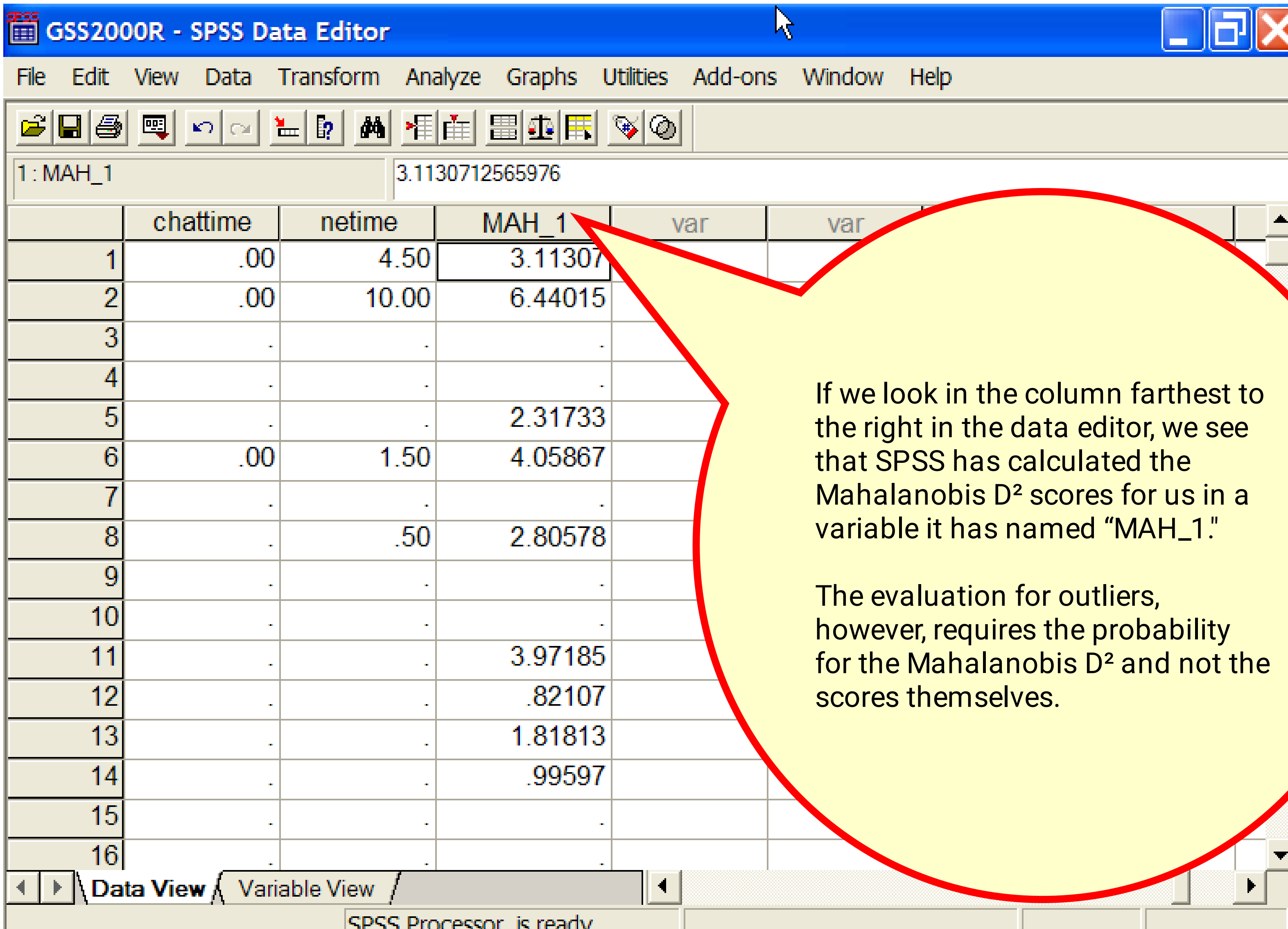
To understand why a particular case is an outlier, we want to examine the descriptive statistics for each variable.

Click on the *Statistics...* button to request the statistics.

Complete the request for Mahalanobis D^2



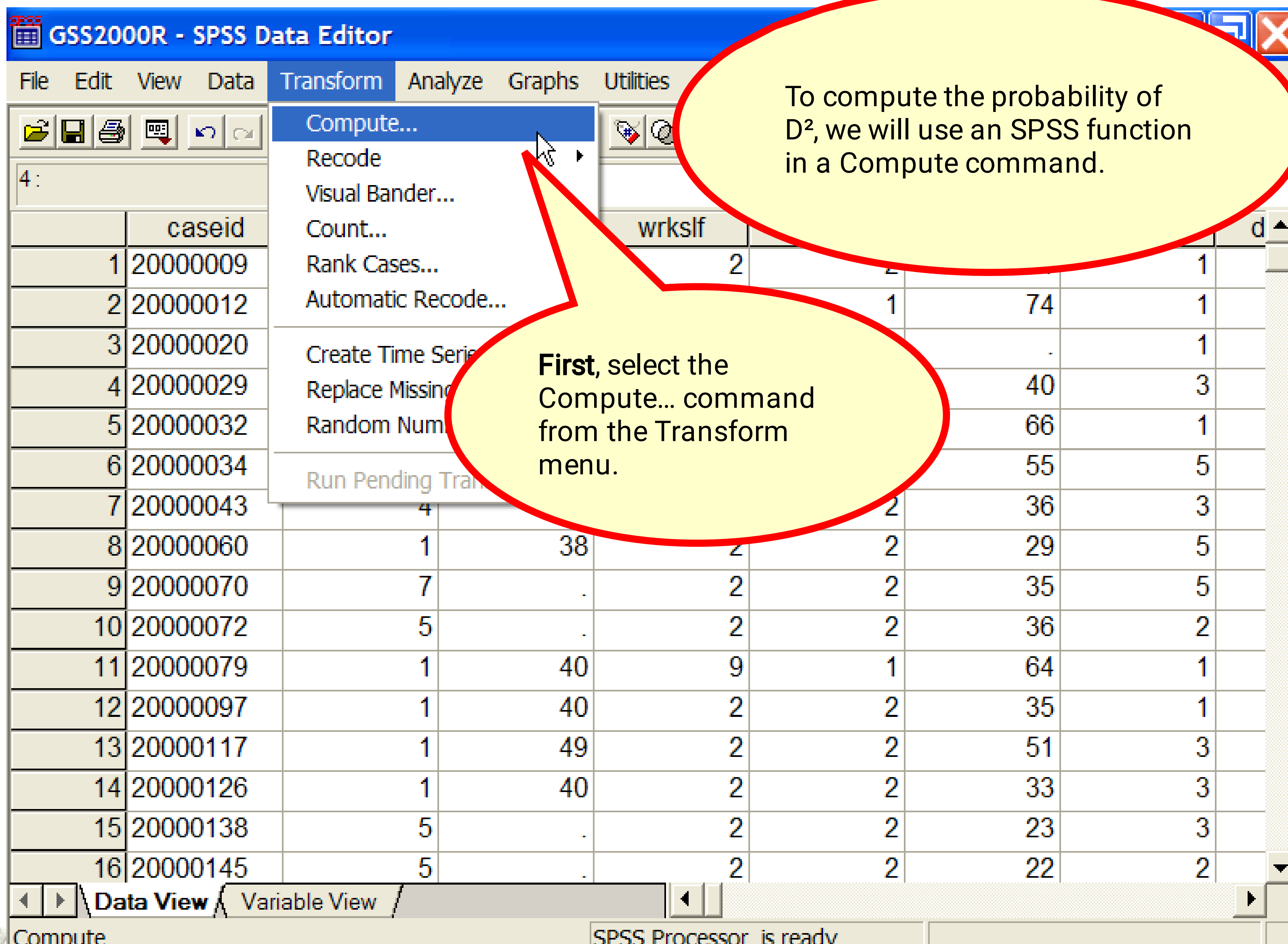
Mahalanobis D^2 scores in the data editor



If we look in the column farthest to the right in the data editor, we see that SPSS has calculated the Mahalanobis D² scores for us in a variable it has named "MAH_1."

The evaluation for outliers, however, requires the probability for the Mahalanobis D² and not the scores themselves.

Computing the probability of D^2

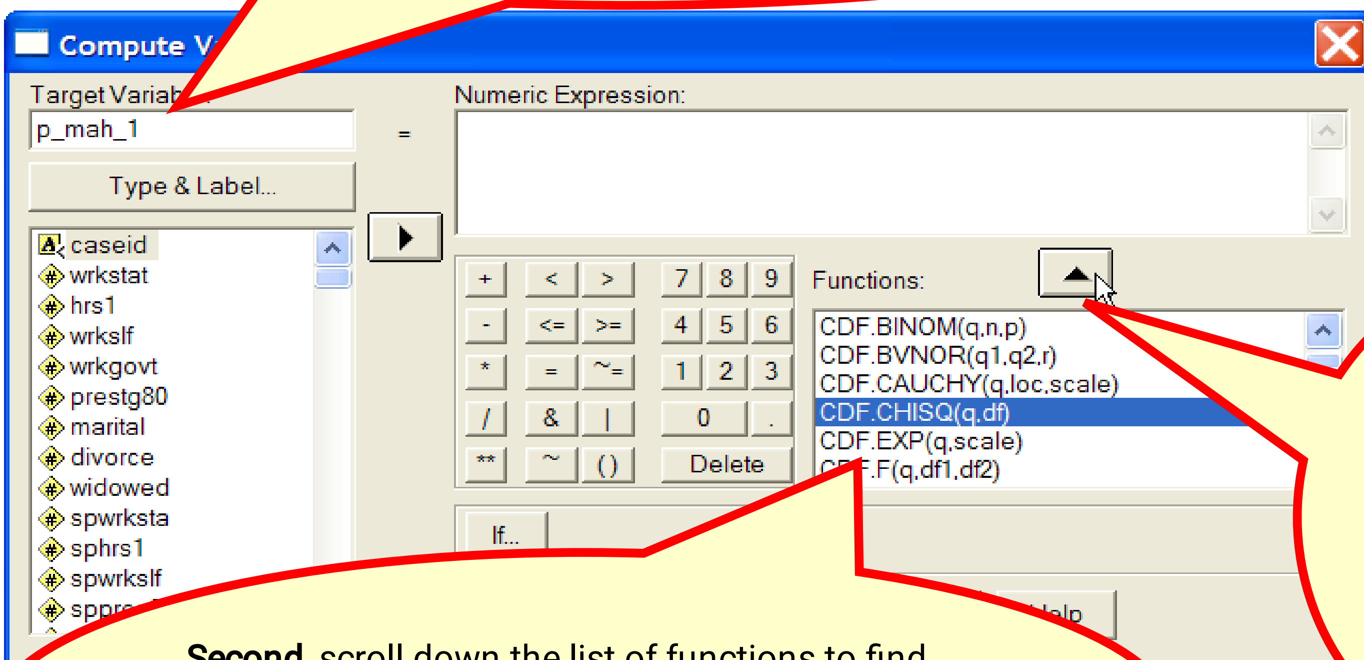


To compute the probability of D^2, we will use an SPSS function in a Compute command.

First, select the Compute... command from the Transform menu.

Specifying the variable name and function

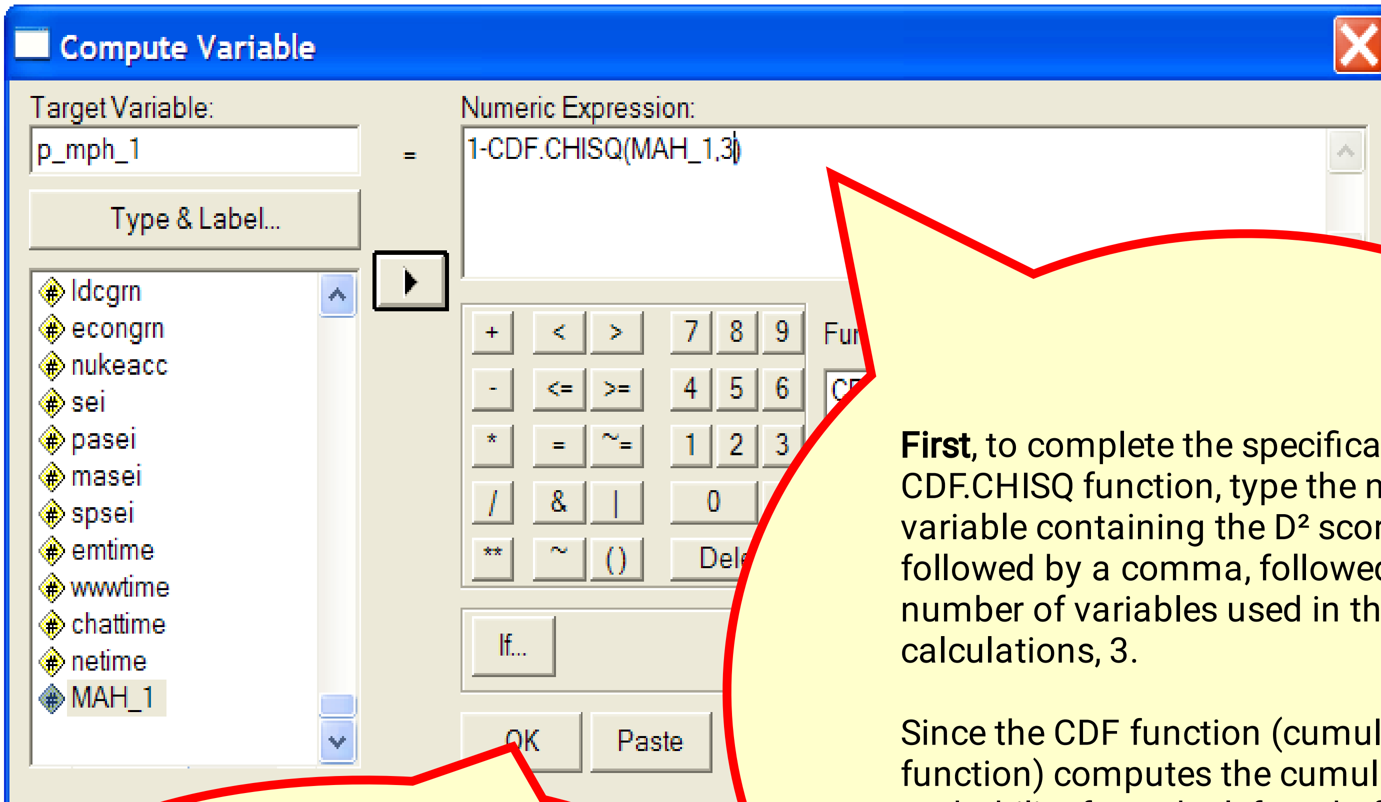
First, in the target variable text box, type the name " p_mah_1" as an acronym for the probability of the mah_1, the Mahalanobis D² score.



Second, scroll down the list of functions to find CDF.CHISQ, which calculates the probability of a variable which follows as chi-square distribution, like Mahalanobis D².

Third, click on the up arrow button to move the highlighted function to the Numeric Expression text box.

Completing the specifications for the function

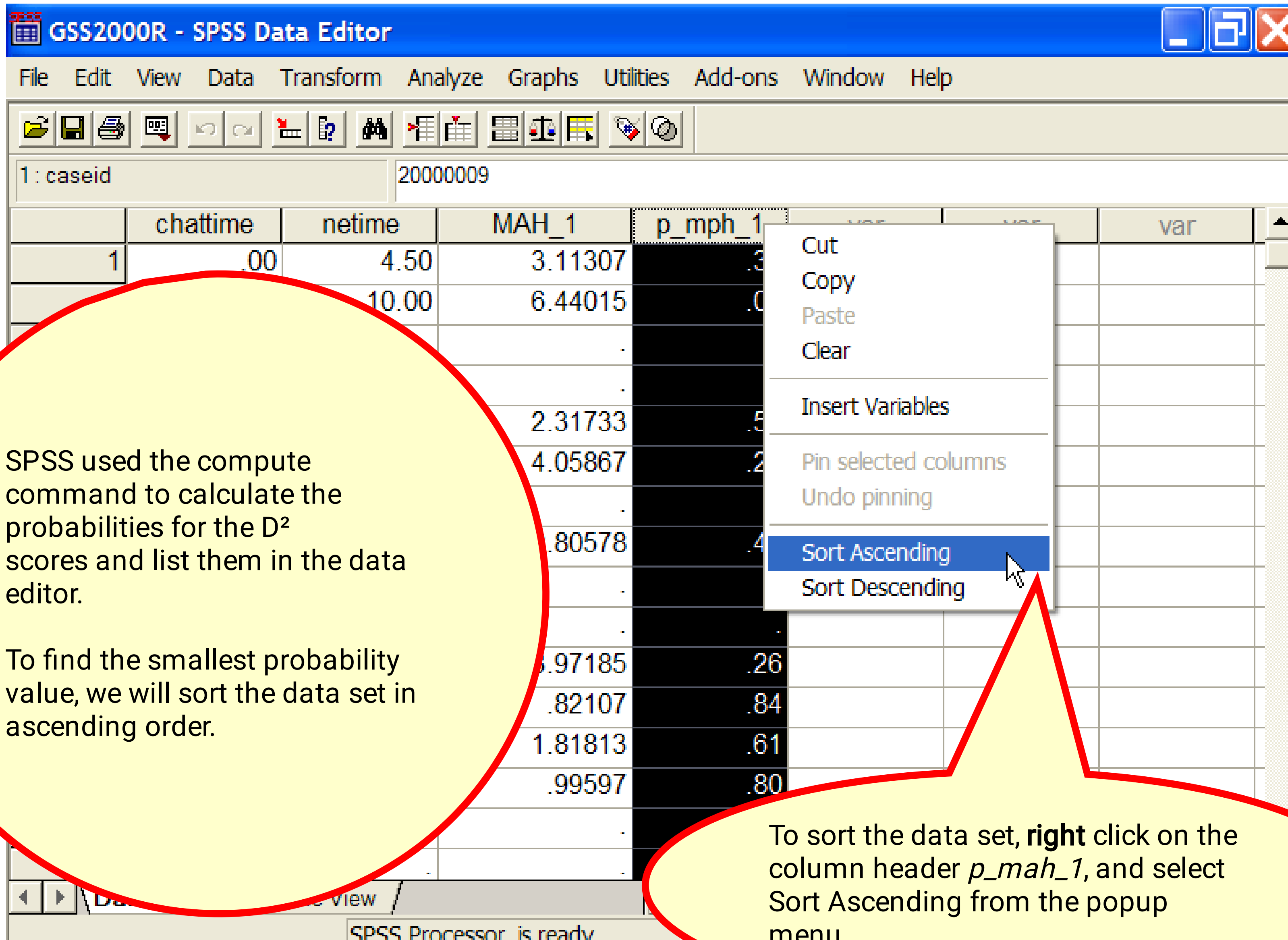


First, to complete the specifications for the CDF.CHISQ function, type the name of the variable containing the D² scores, mah_1, followed by a comma, followed by the number of variables used in the calculations, 3.

Since the CDF function (cumulative density function) computes the cumulative probability from the left end of the distribution up through a given value, we subtract it from 1 to obtain the probability in the upper tail of the distribution.

Second, click on the OK command to signal completion of the computer variable dialog.

Probabilities for D² in the data editor

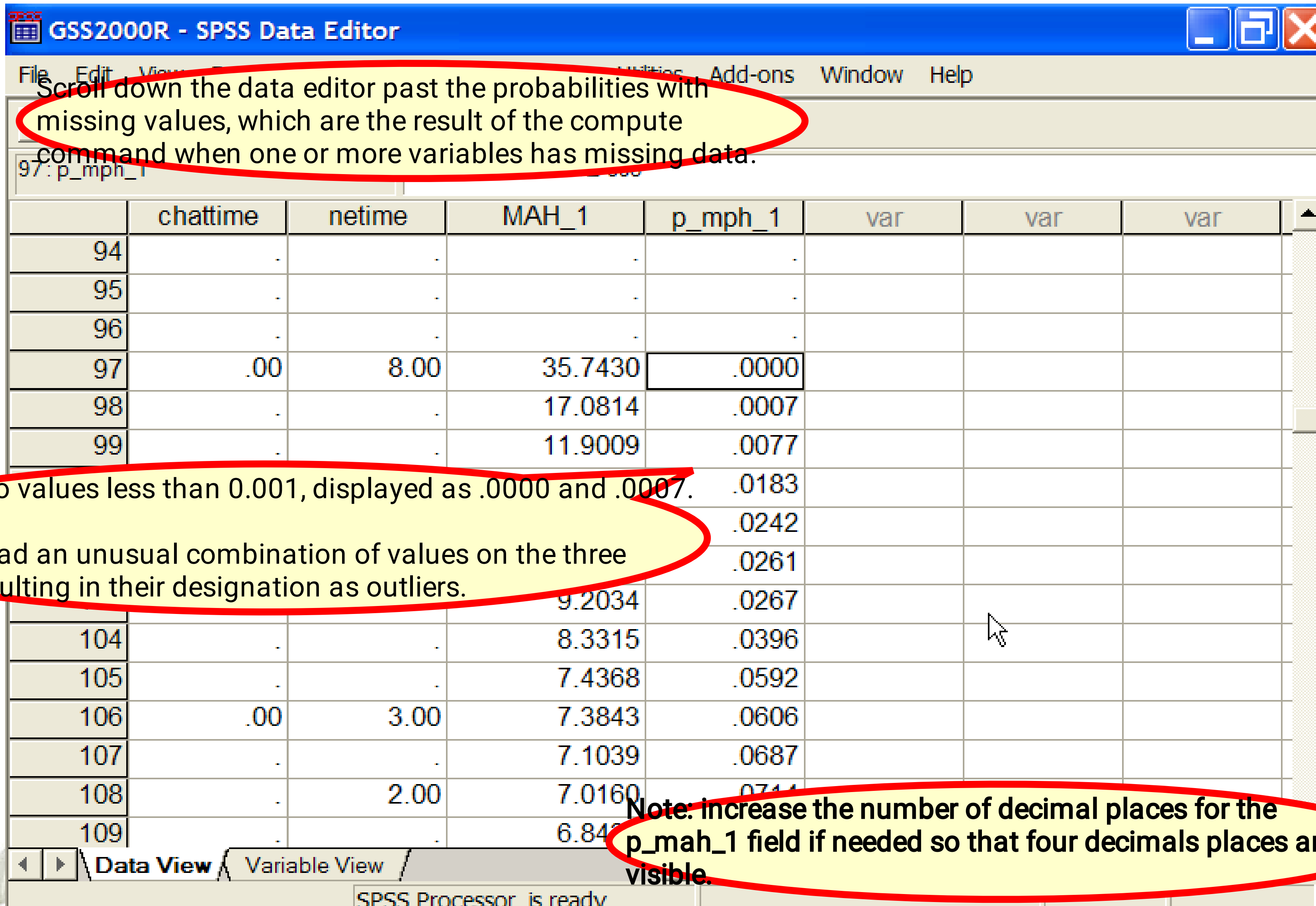


SPSS used the compute command to calculate the probabilities for the D² scores and list them in the data editor.

To find the smallest probability value, we will sort the data set in ascending order.

To sort the data set, **right** click on the column header *p_mah_1*, and select Sort Ascending from the popup menu.

Identifying outliers



Scroll down the data editor past the probabilities with missing values, which are the result of the compute command when one or more variables has missing data.

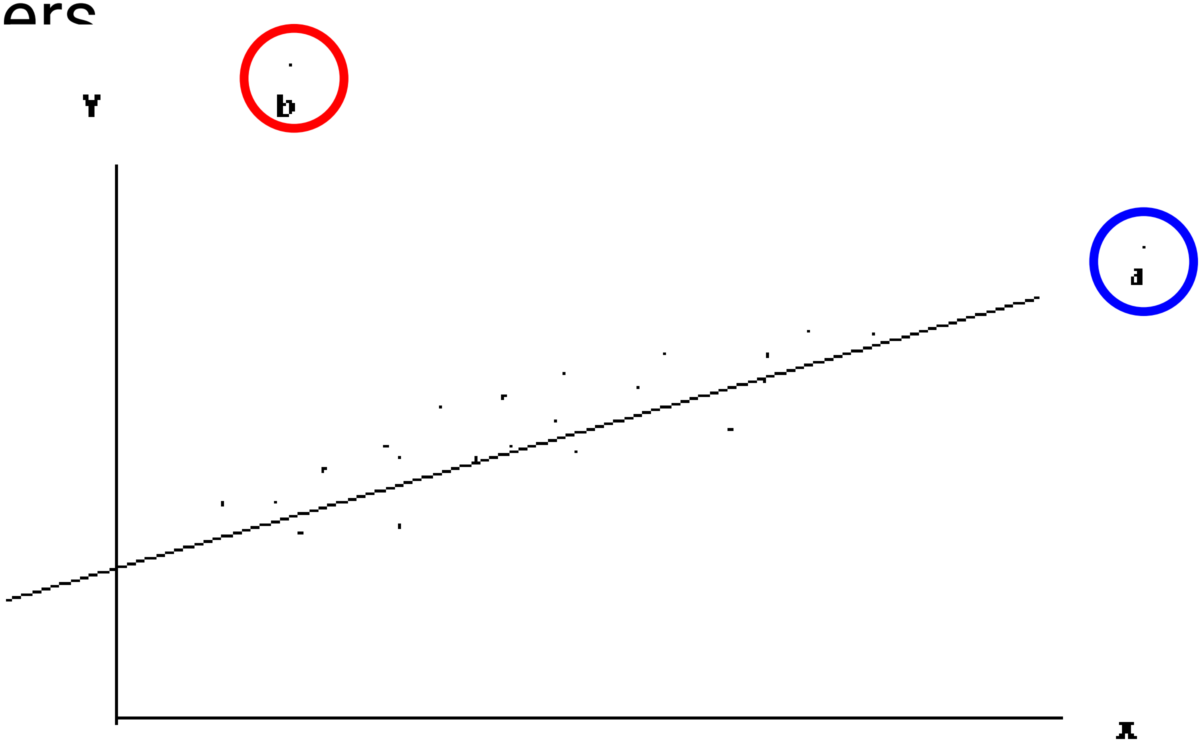
There are two values less than 0.001, displayed as .0000 and .0007.

Two cases had an unusual combination of values on the three variables resulting in their designation as outliers.

Note: increase the number of decimal places for the p_mah_1 field if needed so that four decimals places are visible.

有影响的outlier识别

- Suppose we had a different data set with two outliers



Outlier a does not distort and outlier b does.

有影响的outlier识别

Cook's distance measures the effect of deleting a given observation.

$$D_i = \frac{\sum_{j=1}^{n} (\hat{Y}_j - \hat{Y}_{j(i)})^2}{p \ \mathrm{MSE}}.$$

$\hat{Y}_j$ the prediction from the full regression model for observation j

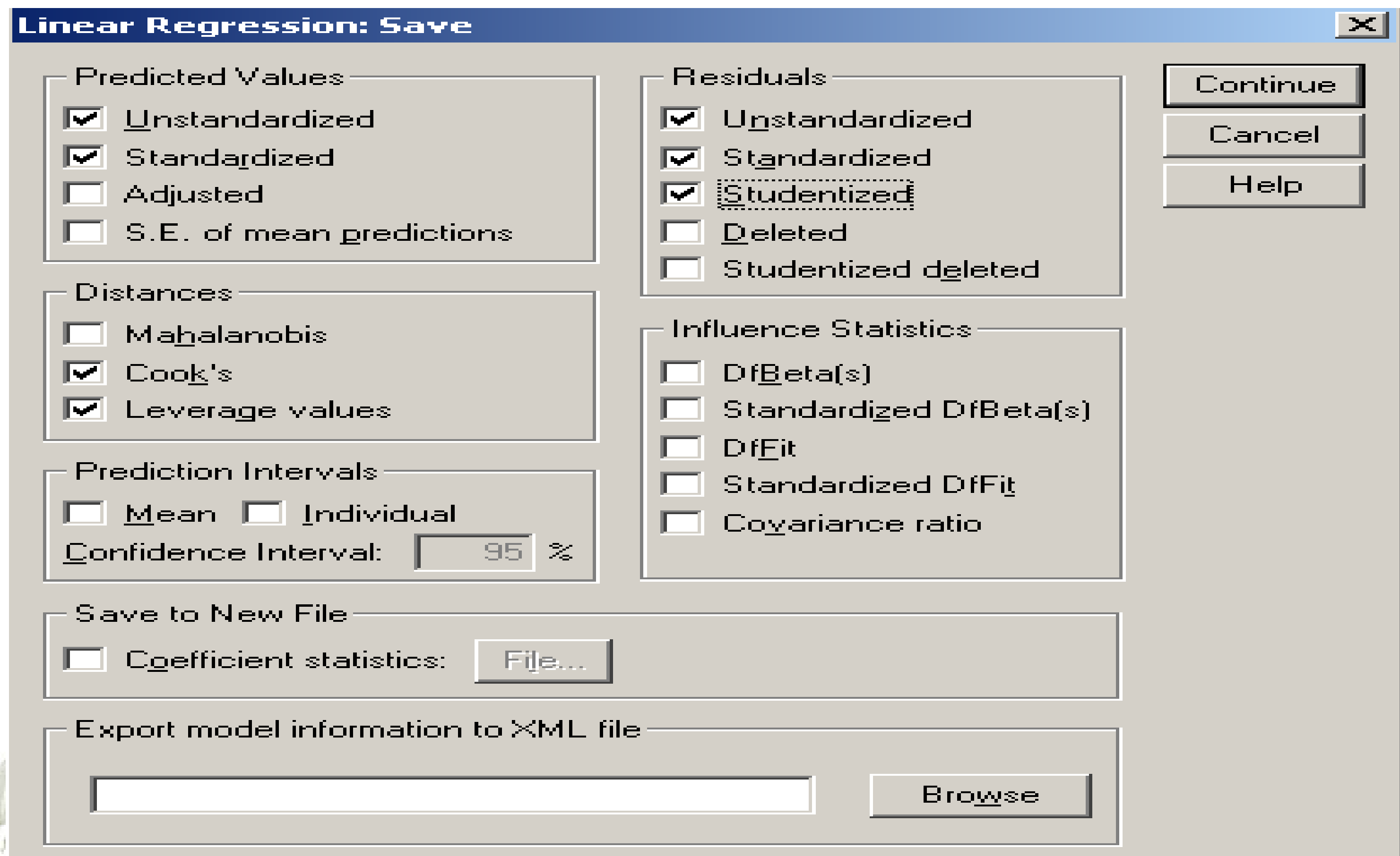
$\hat{Y}_{j(i)}$ the prediction for observation *j* from a refitted regression model in which observation *i* has been omitted

MSE is the mean square error of the regression model

p is the number of fitted parameters in the model

有影响的outlier识别

- List cook, if cook > 4/n
- Belsley suggests 4/(n-k-1) as a cutoff
- In practice, 1 is used as a cutoff



软件缺陷预测：关键点

预处理
- 0: 数据预处理
- 1: 数据分布检查
- 2: Outlier识别

模型构建
- 3: 单变量分析
- 4: 多变量分析

回归方程的选择
回归假设的检查
对不满足回归假设的处理

模型评价
- 5: 模型验证
- 6: 性能评价

多变量线性回归

Purpose of multiple regression

- The purpose of multiple regression is to analyze the relationship between independent variables and a dependent variable

- If there is a relationship, using the information in the independent variables will improve our accuracy in predicting values for the dependent variable

多变量线性回归

Types of multiple regression

- **Standard multiple regression**

  evaluate the relationships between a set of independent variables and a dependent variable

- **Hierarchical/sequential regression**

  examine the relationships between a set of independent variables and a dependent variable, after controlling for the effects of some other independent variables on the dependent variable

- **Stepwise regression**

  identify the subset of independent variables that has the strongest relationship to a dependent variable

多变量线性回归

Standard multiple regression

- All of the independent variables are entered into the regression equation at the same time
- Multiple R and R^2 measure the strength of the relationship
- An F test is used to determine if the relationship can be generalized to the population represented by the sample
- A t-test is used to evaluate the individual relationship between each independent variable and the dependent variable

多变量线性回归

Hierarchical multiple regression

the independent variables are entered in two stages

- In the first stage, the independent variables that we want to control for are entered into the regression. In the second stage, the independent variables whose relationship we want to examine after the controls are entered
- A statistical test of the change in R^2 from the first stage is used to evaluate the importance of the variables entered in the second stage

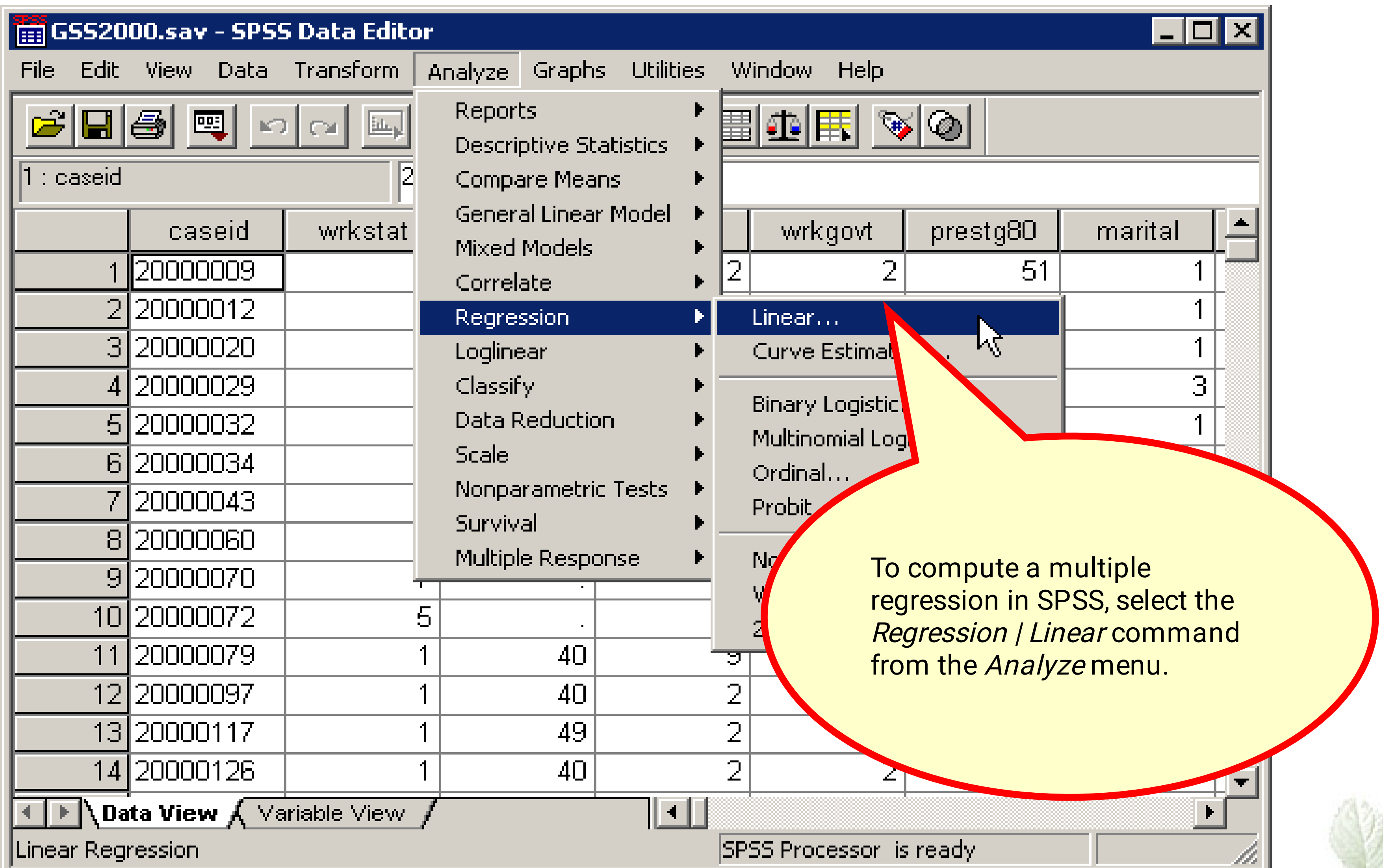
多变量线性回归

Stepwise multiple regression

- Variables are added to the regression equation one at a time, using the statistical criterion of maximizing the R^2 of the included variables
- When none of the possible addition can make a statistically significant improvement in R^2, the analysis stops

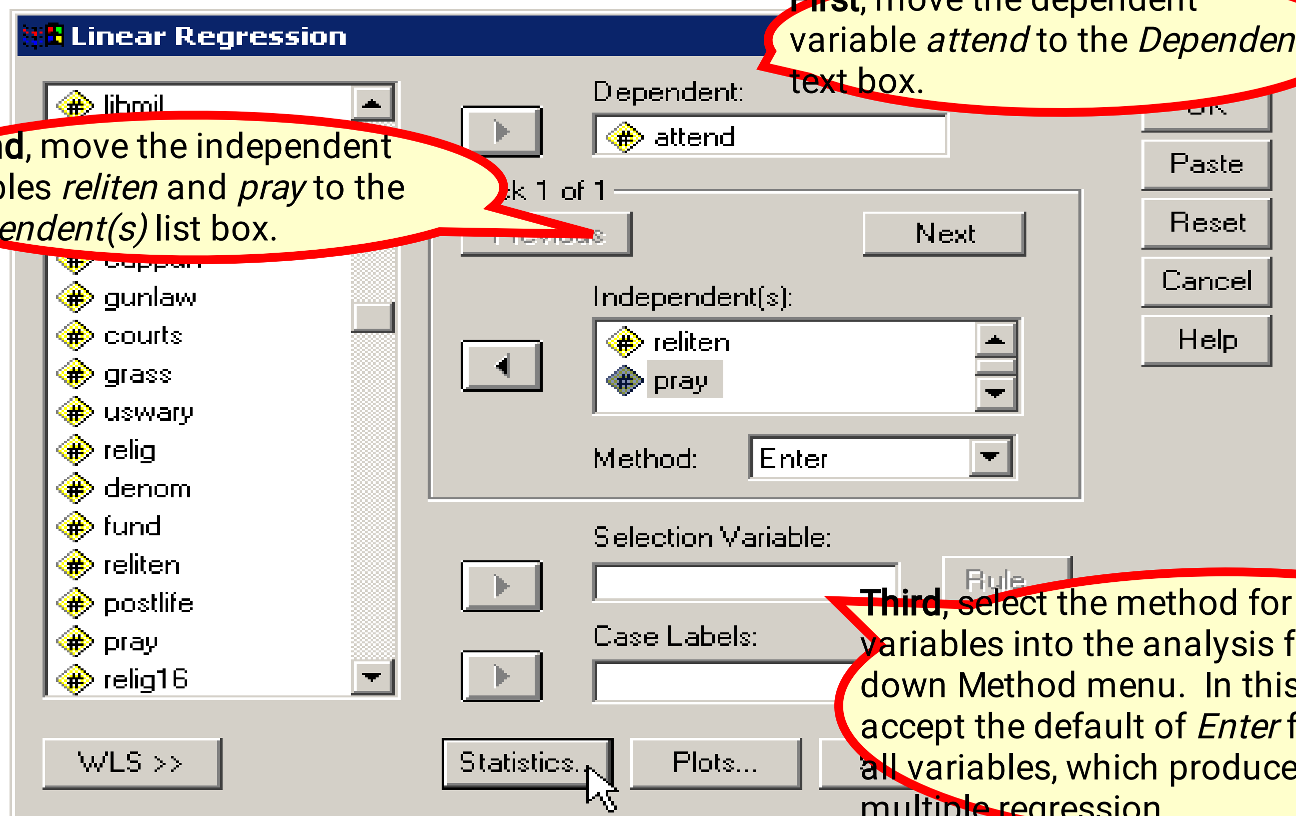
多变量线性回归

standard multiple regression: request



多变量线性回归

standard multiple regression: specify



First, move the dependent variable *attend* to the *Dependent* text box.

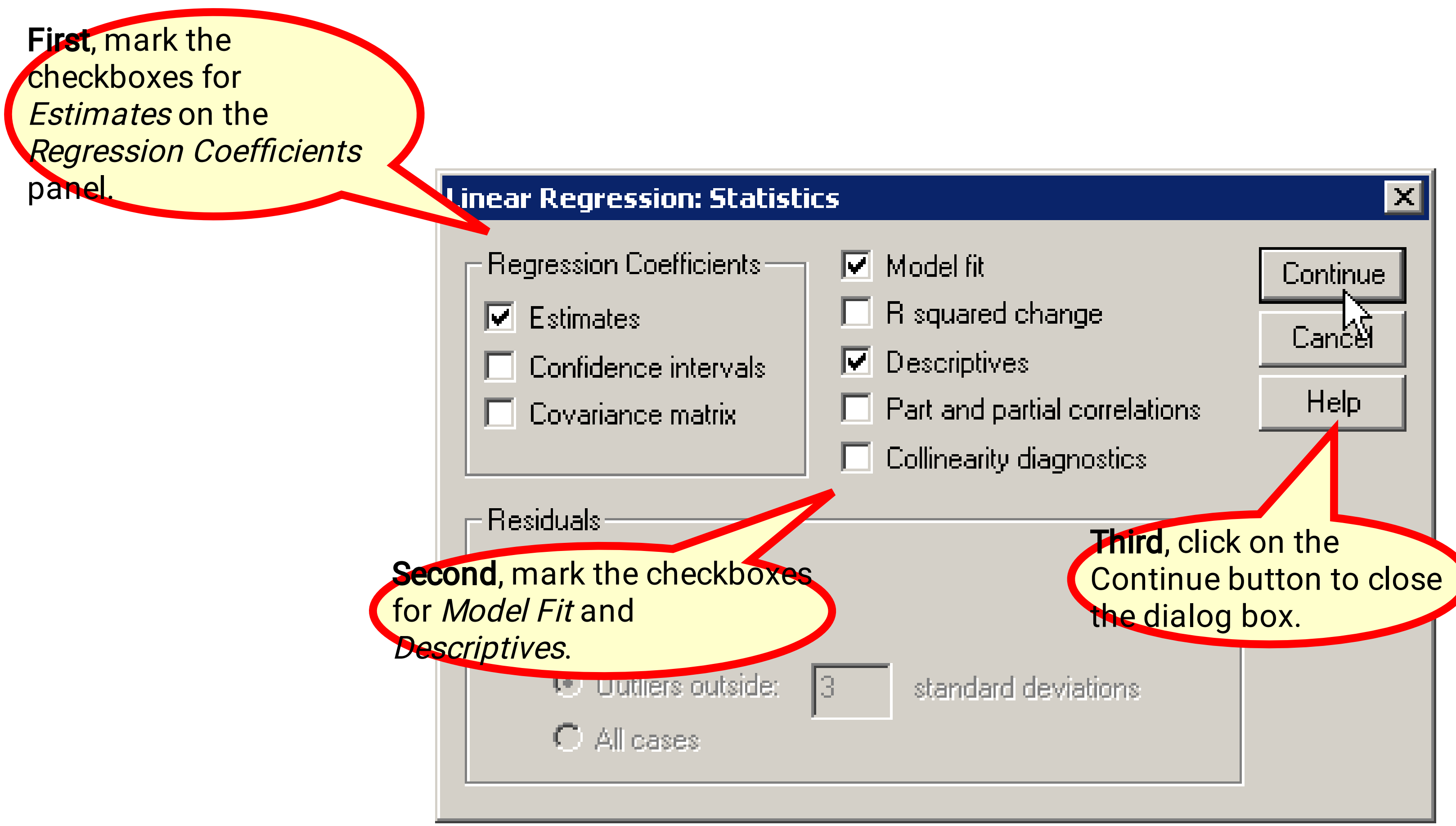
Second, move the independent variables *reliten* and *pray* to the *Independent(s)* list box.

Third, select the method for entering the variables into the analysis from the drop down Method menu. In this example, we accept the default of *Enter* for direct entry of all variables, which produces a standard multiple regression.

Fourth, click on the *Statistics...* button to specify the statistics options that we want.

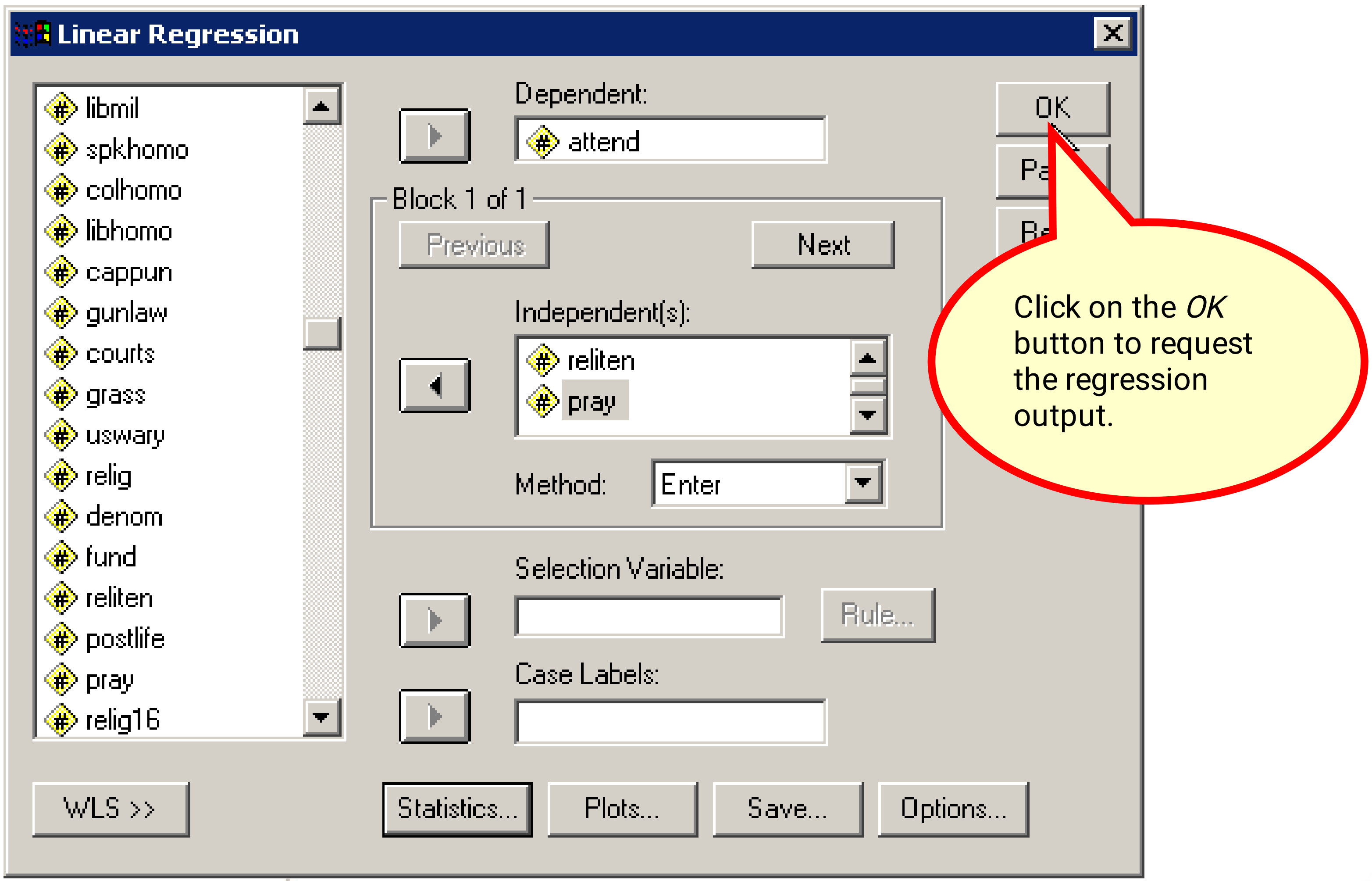
多变量线性回归

standard multiple regression: specify



多变量线性回归

standard multiple regression: request



多变量线性回归

standard multiple regression: output

The probability of the F statistic (49.824) for the overall regression relationship is <0.001, less than or equal to the level of significance of 0.05. We reject the null hypothesis that there is no relationship between the set of independent variables and the dependent variable ($R^2 = 0$). We support the research hypothesis that there is a statistically significant relationship between the set of independent variables and the dependent variable.

ANOVA [b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	374.757	2	187.379	49.824	.000[a]
	Residual	413.685	110	3.761		
	Total	788.442	112			

a. Predictors: (Constant), HOW OFTEN DOES R PRAY, STRENGTH OF AFFILIATION

b. Dependent Variable: HOW OFTEN R ATTENDS RELIGIOUS SERVICES

多变量线性回归

standard multiple regression: output

The Multiple R for the relationship between the set of independent variables and the dependent variable is 0.689, which would be characterized as strong using the rule of thumb than a correlation less than or equal to 0.20 is characterized as very weak; greater than 0.20 and less than or equal to 0.40 is weak; greater than 0.40 and less than or equal to 0.60 is moderate; greater than 0.60 and less than or equal to 0.80 is strong; and greater than 0.80 is very strong.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.689[a]	.475	.466	1.939

a. Predictors: (Constant), HOW OFTEN DOES R PRAY, STRENGTH OF AFFILIATION

多变量线性回归

standard multiple regression: output

For the independent variable strength of affiliation, the probability of the t statistic (-5.857) for the b coefficient is <0.001 which is less than or equal to the level of significance of 0.05. We reject the null hypothesis that the slope associated with strength of affiliation is equal to zero (b = 0) and conclude that there is a statistically significant relationship between strength of affiliation and frequency of attendance at religious services.

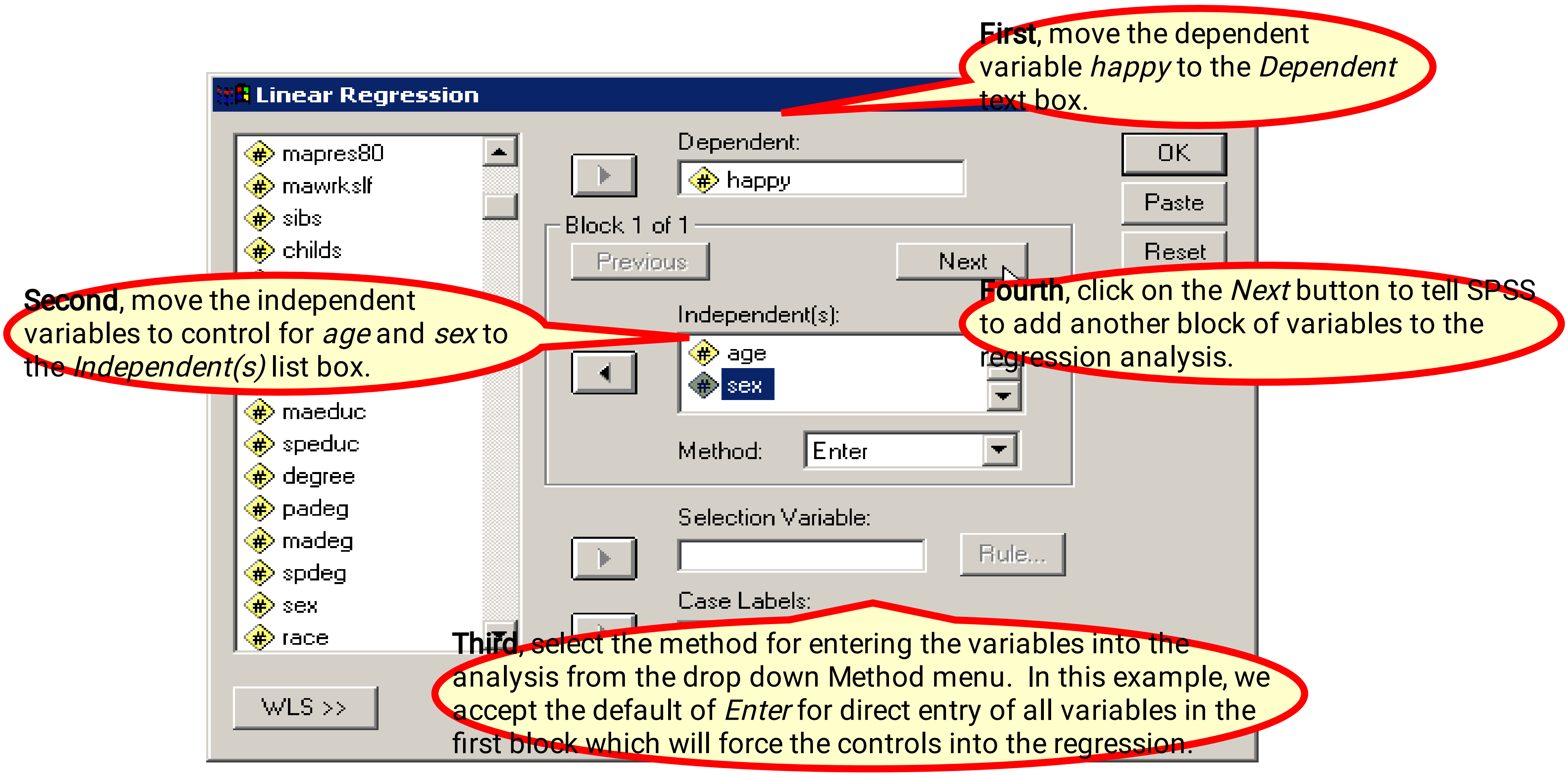
Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	7.167	.442		16.206	.000
	STRENGTH OF AFFILIATION	-1.138	.194	-.465	-5.857	.000
	HOW OFTEN DOES R PRAY	-.554	.134	-.329	-4.145	.000

a. Dependent Variable: HOW OFTEN R ATTENDS RELIGIOUS SERVICES

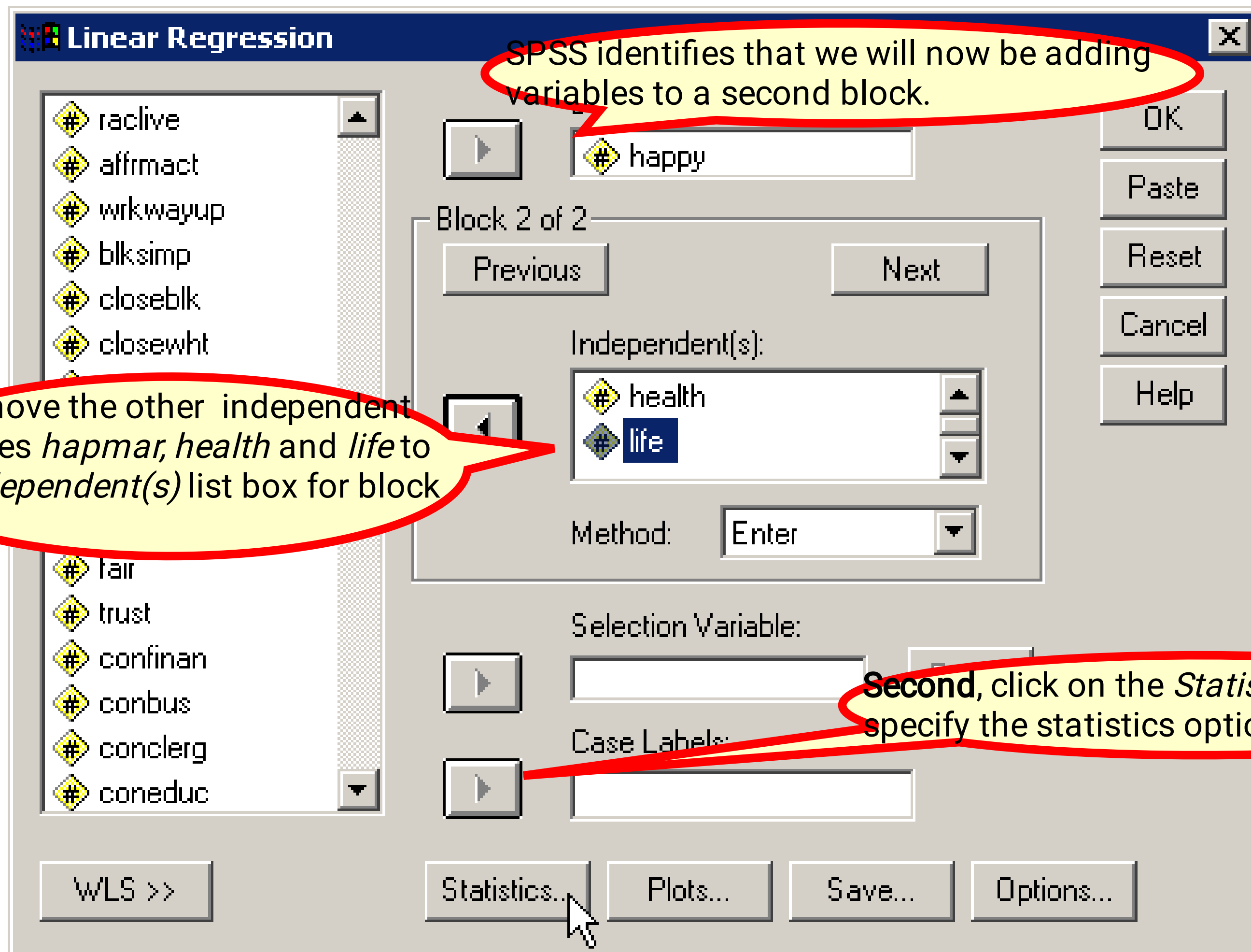
多变量线性回归

Hierarchical multiple regression: specify



多变量线性回归

Hierarchical multiple regression: specify



SPSS identifies that we will now be adding variables to a second block.

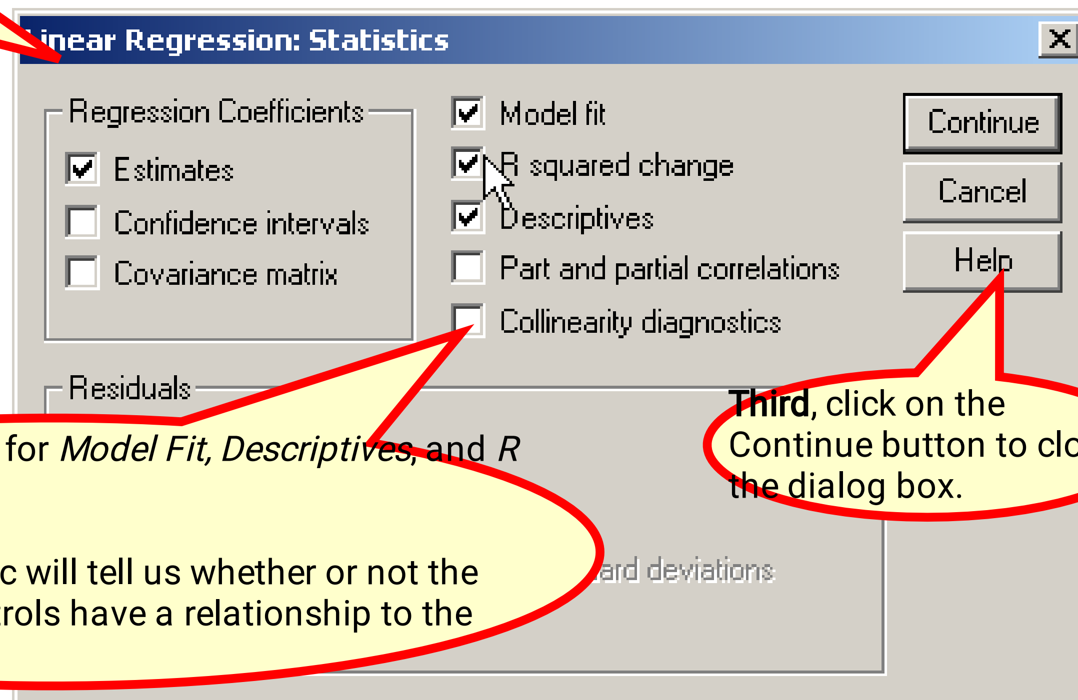
First, move the other independent variables *hapmar, health* and *life* to the *Independent(s)* list box for block 2.

Second, click on the *Statistics...* button to specify the statistics options that we want.

多变量线性回归

Hierarchical multiple regression: specify

First, mark the checkboxes for *Estimates* on the *Regression Coefficients* panel.



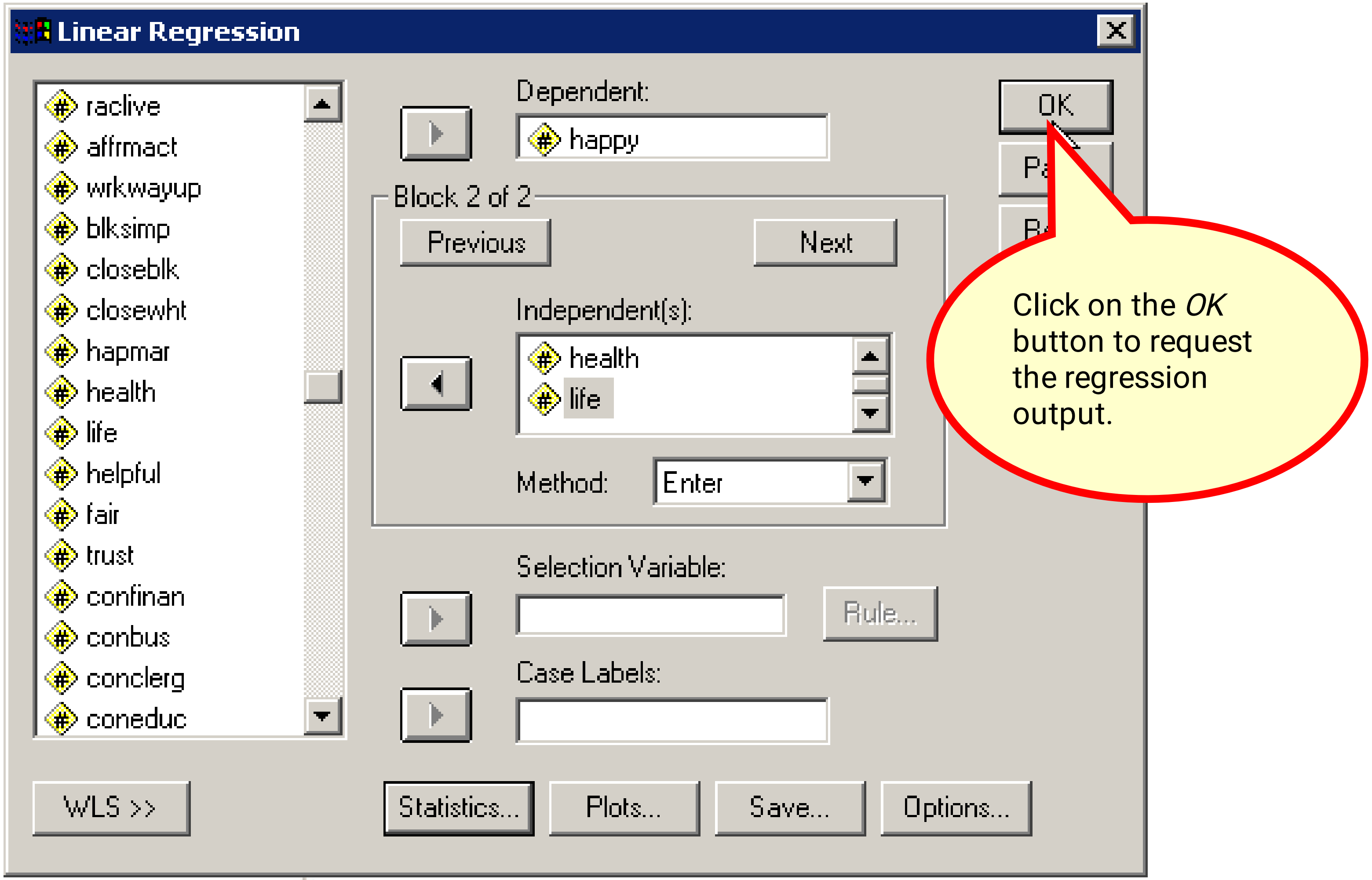
Second, mark the checkboxes for *Model Fit, Descriptives,* and *R squared change.*

The *R squared change* statistic will tell us whether or not the variables added after the controls have a relationship to the dependent variable.

Third, click on the Continue button to close the dialog box.

多变量线性回归

Hierarchical multiple regression: request



多变量线性回归

Hierarchical multiple regression: output

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	.006	2	.003	.007	.993[a]
	Residual	34.894	87	.401		
	Total	34.900	89			
2	Regression	12.601	5	2.520	9.493	.000[b]
	Residual	22.299	84	.265		
	Total	34.900	89			

a. Predictors: (Constant), RESPONDENTS SEX, AGE OF RESPONDENT

b. Predictors: (Constant), RESPONDENTS SEX, AGE OF RESPONDENT, IS LIFE EXCITING OR DULL, HAPPINESS OF MARRIAGE, CONDITION OF HEALTH

c. Dependent Variable: GENERAL HAPPINESS

The probability of the F statistic (9.493) for the overall regression relationship for all indpendent variables is <0.001, less than or equal to the level of significance of 0.05. We reject the null hypothesis that there is no relationship between the set of all independent variables and the dependent variable ($R^2 = 0$). We support the research hypothesis that there is a statistically significant relationship between the set of all independent variables and the dependent variable.

多变量线性回归

Hierarchical multiple regression: output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. F Change
1	.013[a]	.000	-.023	.633	.000	.007	2	87	.993
2	.601[b]	.361	.323	.515	.361	15.814	3	84	.000

a. Predictors: (Constant), RESPONDENTS SEX, AGE OF RESPONDENT

b. Predictors: (Constant), RESPONDENTS SEX, AGE OF RESPONDENT, IS LIFE EXCITING OR DULL, HAPPINESS OF MARRIAGE, CONDITION OF HEALTH

The R Square Change statistic for the increase in R^2 associated with the added variables (happiness of marriage, condition of health, and attitude toward life) is 0.361. Using a proportional reduction in error interpretation for R^2, information provided by the added variables reduces our error in predicting general happiness by 36.1%.

多变量线性回归

Hierarchical multiple regression: output

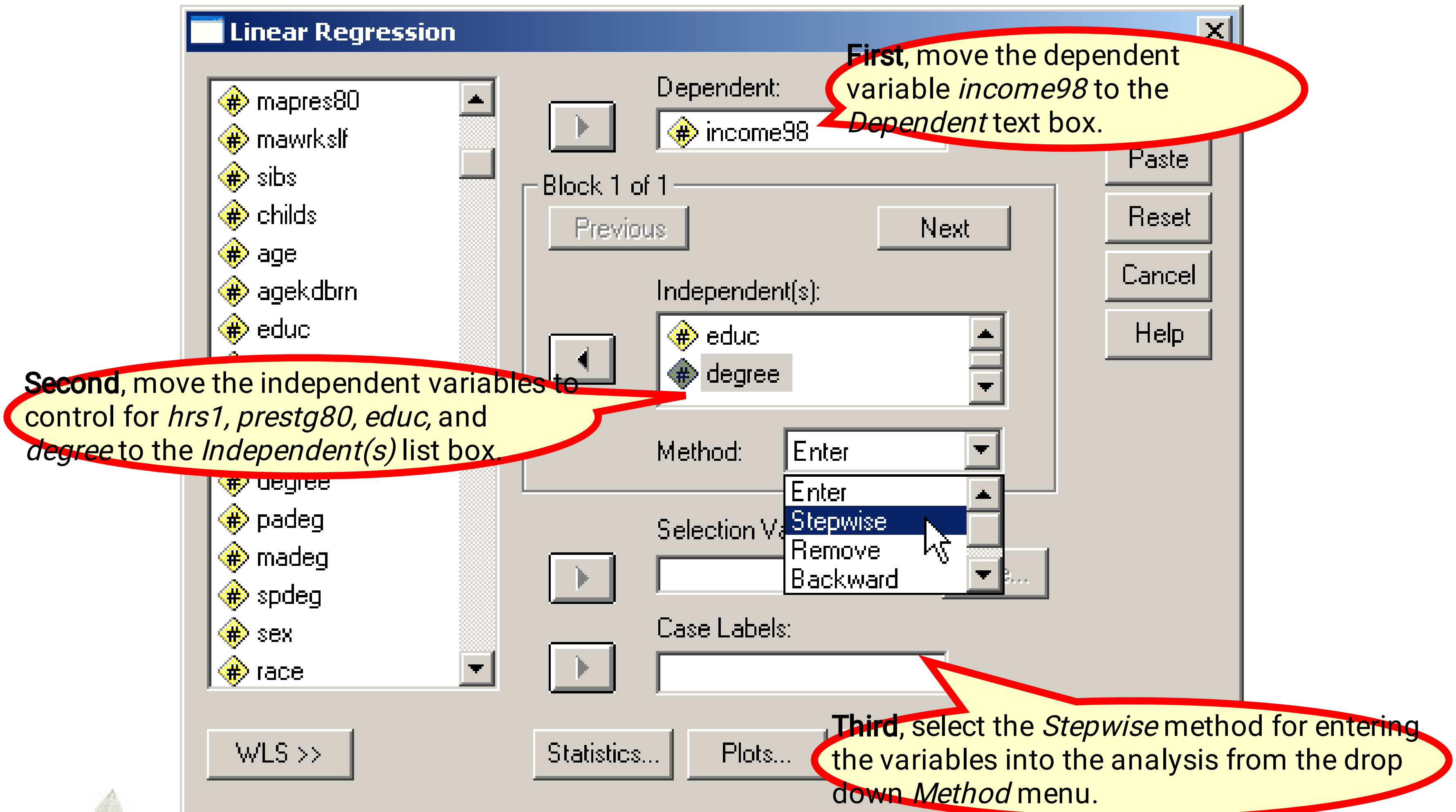
Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	1.594	.341		4.677	.000
	AGE OF RESPONDENT	.000	.005	.012	.107	.915
	RESPONDENTS SEX	.011	.140	.008	.078	.938
2	(Constant)	.432	.341		1.268	.208
	AGE OF RESPONDENT	-.001	.004	-.035	-.385	.701
	RESPONDENTS SEX	-.013	.115	-.010	-.113	.911
	HAPPINESS OF MARRIAGE	.599	.104	.517	5.741	.000
	CONDITION OF HEALTH	.101	.072	.131	1.408	.163
	IS LIFE EXCITING OR DULL	.170	.108	.142	1.570	.120

a. Dependent Variable: GENERAL HAPPINESS

多变量线性回归

Stepwise multiple regression: specify



多变量线性回归

Stepwise multiple regression: output

The best subset of predictors for total family income included the independent variables: highest academic degree and occupational prestige score.

Variables Entered/Removed [a]

Model	Variables Entered	Variables Removed	Method
1	RS HIGHEST DEGREE	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	RS OCCUPATIONAL PRESTIGE SCORE (1980)	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: TOTAL FAMILY INCOME

多变量线性回归

Stepwise multiple regression: output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.492[a]	.242	.237	3.607
2	.532[b]	.283	.273	3.522

a. Predictors: (Constant), RS HIGHEST DEGREE

b. Predictors: (Constant), RS HIGHEST DEGREE, RS OCCUPATIONAL PRESTIGE SCORE (1980)

The Multiple R for the relationship between the subset of independent variables that best predict the dependent variable is 0.532, which would be characterized as moderate using the rule of thumb than a correlation less than or equal to 0.20 is characterized as very weak; greater than 0.20 and less than or equal to 0.40 is weak; greater than 0.40 and less than or equal to 0.60 is moderate; greater than 0.60 and less than or equal to 0.80 is strong; and greater than 0.80 is very strong.

多变量线性回归

Multiple regression and assumptions

- each of the independent/dependent variables are **normally distributed**
- the relationships between the independent and dependent variables are **linear**
- the relationship between metric and dichotomous variables is **homoscedastic**
- the **errors are independent** and there is **no serial correlation**

多变量线性回归

Multiple regression and outliers

- Outliers can distort the regression results. When an outlier is included in the analysis, it pulls the regression line towards itself. This can result in a solution that is more accurate for the outlier, but less accurate for all of the other cases in the data set

- We will check for univariate outliers on the dependent variable and multivariate outliers on the independent variables

多变量线性回归

Relationship between assumptions and outliers

- The problems of satisfying assumptions and detecting outliers are intertwined. For example, if a case has a value on the dependent variable that is an outlier, it will affect the skew, and hence, the normality of the distribution

- Removing an outlier may improve the distribution of a variable

- Transforming a variable may reduce the likelihood that the value for a case will be characterized as an outlier

多变量线性回归

Order of analysis is important

- The order in which we check assumptions and detect outliers will affect our results because we may get a different subset of cases in the final analysis

- In order to maximize the number of cases available to the analysis, we will evaluate assumptions first. We will substitute any transformations of variable that enable us to satisfy the assumptions

- We will use any transformed variables that are required in our analysis to detect outliers

多变量线性回归

Strategy for solving problems

① Run type of regression specified using full data set

② Test the dependent variable for normality and decide which transformations should be used

③ Test the independent variables for normality, linearity, homoscedasticity and decide which transformations should be used

④ Substitute transformations and run regression entering all independent variables, saving studentized residuals and Mahalanobis distance scores. Compute probabilities for D^2

⑤ Remove the outliers (studentized residual greater than 3 or Mahalanobis D^2 with $p \le 0.001$), and run regression with the method and variables specified in the problem

⑥ Compare R^2 for analysis using transformed variables and omitting outliers (step 5) to R^2 obtained for model using all data and original variables (step 1)

多变量线性回归

Transforming dependent variables

- If dependent variable is not normally distributed:
  - Try log, square root, and inverse transformation. Use first transformed variable that satisfies normality criteria
  - If no transformation satisfies normality criteria, use untransformed variable and add caution for violation of assumption

- If a transformation satisfies normality, use the transformed variable in the tests of the independent variables

多变量线性回归

Transforming independent variables - 1

- If independent variable is normally distributed and linearly related to dependent variable, use as is

- If independent variable is normally distributed but not linearly related to dependent variable:
  - Try log, square root, square, and inverse transformation. Use first transformed variable that satisfies linearity criteria and does not violate normality criteria
  - If no transformation satisfies linearity criteria and does not violate normality criteria, use untransformed variable and add caution for violation of assumption

多变量线性回归

Transforming independent variables - 2

- If independent variable is linearly related to dependent variable but not normally distributed:
  - Try log, square root, and inverse transformation. Use first transformed variable that satisfies normality criteria and does not reduce correlation
  - Try log, square root, and inverse transformation. Use first transformed variable that satisfies normality criteria and has significant correlation
  - If no transformation satisfies normality criteria with a significant correlation, use untransformed variable and add caution for violation of assumption

Transforming independent variables - 3

- If independent variable is not linearly related to dependent variable and not normally distributed:
  - Try log, square root, square, and inverse transformation. Use first transformed variable that satisfies normality criteria and has significant correlation.
  - If no transformation satisfies normality criteria with a significant correlation, used untransformed variable and add caution for violation of assumption

Problem 1

In the dataset GSS2000.sav, is the following statement true, false, or an incorrect application of a statistic? Assume that there is no problem with missing data. Use a level of significance of 0.05 for the regression analysis. Use a level of significance of 0.01 for evaluating assumptions. Use 0.0160 as the criteria for identifying influential cases. Validate the results of your regression analysis by splitting the sample in two, using 788035 as the random number seed.

The variables "age" [age], "sex" [sex], and "respondent's socioeconomic index" [sei] have a strong relationship to the variable "how many in family earned money" [earnrs].

Survey respondents who were older had fewer family members earning money. The variables sex and respondent's socioeconomic index did not have a relationship to how many in family earned money.

1. True
2. True with caution
3. False
4. Inappropriate application of a statistic

多变量线性回归

Problem 1

When we test for influential cases using Cook's distance, we need to compute a critical value for comparison using the formula:

$$4 / (n - k - 1)$$

where n is the number of cases and k is the number of independent variables. The correct value (0.0160) is provided in the problem.

The problem may give us different levels of significance for the analysis.

In this problem, we are told to use 0.05 as alpha for the regression, but 0.01 for testing assumptions.

In the … GSS2000.sav, is the following … statement true, false, or an in… application of a statistic? Assume that there is no problem … with missing data. Use a level of significance of 0.05 for the regression analysis. Use a level of significance of 0.01 for evaluating assumptions. Use 0.0160 as the criteria for identifying influential cases. Validate the results of your regression analysis by splitting the sample in two, using 788035 as the random number see…

The random number seed (788035) for the split sample validation is provided.

After evaluating assumptions, outliers, and influential cases, we will decide whether we should use the model with transformations and excluding outliers, or the model with the original form of the variables and all cases.

… age … ex" [se… … family earned money …

多变量线性回归

Problem 1

In the dat ... the following statem ... an
in ... ume tha ...
... nce ...
an ... 0.01 f ...
Use 0.01 ... identifying influe ...
results of your regression analysis by splitting th ...
788035 as the random number seed.

When a problem states that there is a relationship between some independent variables and a dependent variable, we do standard multiple regression.

The variables listed first in the problem statement are the independent variables (IVs): "age" [age], " sex" [sex], and "respondent's socioeconomic index" [sei].

The variables "age" [age], "sex" [sex], and "respondent's socioeconomic index" [sei] have a strong relationship to the variable "how ma y in family earned money" [earnrs].

The variable that is the target of the relationship is the dependent variable (DV): " how many in family earned money" [earnrs].

Survey respondents who were older had f ... g
money. The variables sex and respondent ... t
have a relationship to how many in family ...

1. True
2. True with caution
3. False
4. Inappropriate application of a statistic

多变量线性回归

Problem 1

In the dataset GSS2000.sav, is the followin... true, false, or an
incorrect application of a stati... ...lem with
missing data. Use a level
analysis. Use a level of s
Use 0.0160 as the criteria ... e
results of your regression analy... ...o, using
788035 as the random number see

In order for a problem to be true, we will have to find that there is a statistically significant relationship between the set of IVs and the DV, and the strength of the relationship stated in the problem must be correct.

The variables "age" [age], "sex" [sex], and "respondent's socioeconomic index" [sei] have a strong relationship to the variable "how many in family earned money" [earnrs].

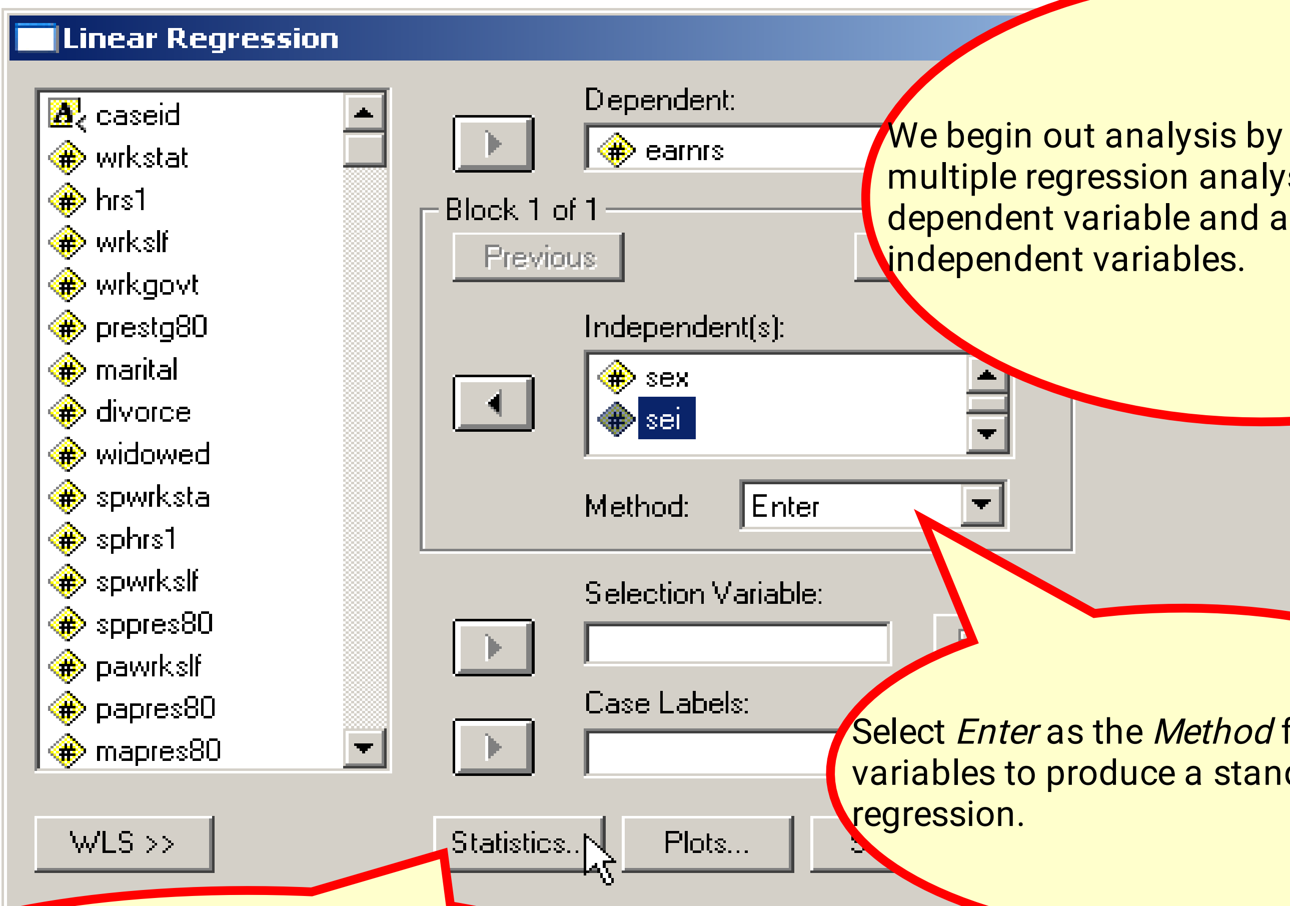
Survey respondents who were older had fewer family members earning money. The variables sex and respondent's socioeconomic index did not have a relationship to how many ... ...mily earned money.

In addition, the relationship or lack of relationship between the individual IV's and the DV must be identified correctly, and must be characterized correctly.

1. True
2. True with caution
3. False
4. Inappropriate application of a statistic

多变量线性回归

The baseline regression



We begin out analysis by running a standard multiple regression analysis with earnrs as the dependent variable and age, sex, and sei as the independent variables.

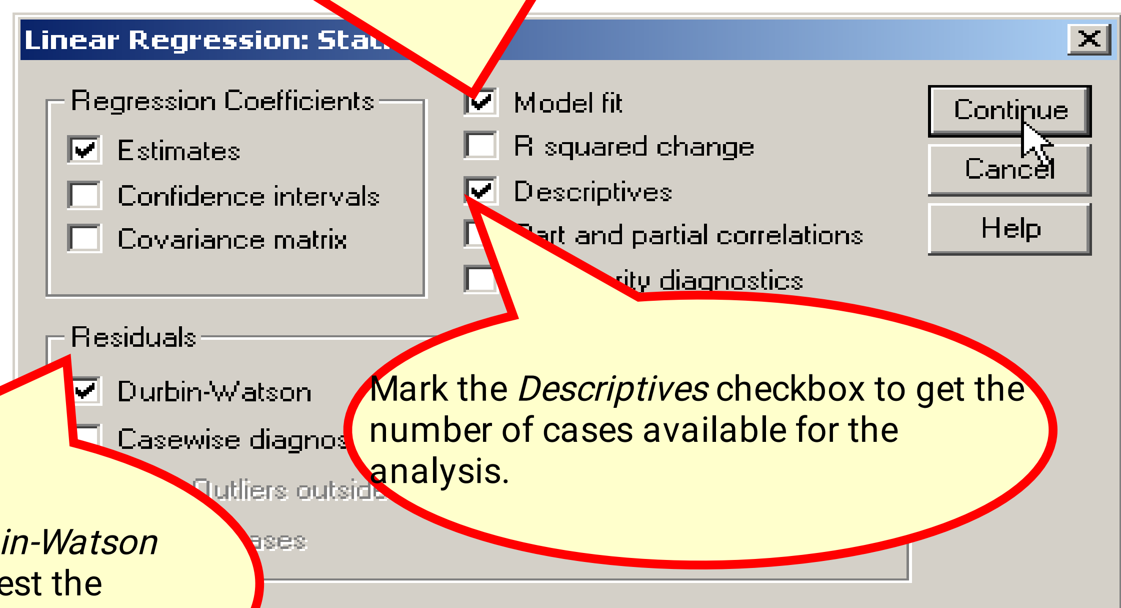
Select *Enter* as the *Method* for including variables to produce a standard multiple regression.

Click on the *Statistics...* button to select statistics we will need for the analysis.

多变量线性回归

The baseline regression

Retain the default checkboxes for *Estimates* and *Model fit* to obtain the baseline R^2, which will be used to determine whether we should use the model with transformations and excluding outliers, or the model with the original form of the variables and all cases.



Mark the *Descriptives* checkbox to get the number of cases available for the analysis.

Mark the checkbox for the *Durbin-Watson* statistic, which will be used to test the assumption of independence of errors.

多变量线性回归

Initial sample size

Descriptive Statistics

	Mean	Std. Deviation	N
EARNRS	1.47	1.008	254
AGE	46.62	16.642	254
SEX	1.57	.496	254
SEI	48.601	19.1110	254

The initial sample size before excluding outliers and influential cases is 254. With 3 independent variables, the ratio of cases to variables is 84.7 to 1, satisfying both the minimum and preferred sample size for multiple regression.

If the sample size did not initially satisfy the minimum requirement, regression analysis is not appropriate.

多变量线性回归

R^2 before transformations or removing outliers

The R^2 of 0.187 is the benchmark that we will use to evaluate the utility of transformations and the elimination of outliers/influential cases.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
	.433[a]	.187	.178	.915	1.849

), SEI, AGE, SEX

EARNRS

Prior to any transformations of variables to satisfy the assumptions of multiple regression or removal of outliers, the proportion of variance in the dependent variable explained by the independent variables (R^2) was 18.7%.

The relationship is statistically significant, though we would not stop if it were not significant because the lack of significance may be a consequence of violation of assumptions or the inclusion of outliers and influential cases.

ANOVA[b]

	df	Mean Square	F	Sig.
8	3	16.069	19.213	.000[a]
99	250	.836		
307	253			

SEI, AGE, SEX

: EARNRS

多变量线性回归

Assumption of independence of errors: the Durbin-Watson statistic

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.433[a]	.187	.178	.915	1.849

a. Predictors: (Constant)

The Durbin-Watson statistic is used to test for the presence of serial correlation among the residuals, i.e., the assumption of independence of errors, which requires that the residuals or errors in prediction do not follow a pattern from case to case.

The value of the Durbin-Watson statistic ranges from 0 to 4. As a general rule of thumb, the residuals are not correlated if the Durbin-Watson statistic is approximately 2, and an acceptable range is 1.50 - 2.50.

The Durbin-Watson statistic for this problem is 1.849 which falls within the acceptable range.

If the Durbin-Watson statistic was not in the acceptable range, we would add a caution to the findings for a violation of regression assumptions.

多变量线性回归

Normality of dependent variable: how many in family earned money

Descriptives

			Statistic	Std. Error
HOW MANY IN FAMILY EARNED MONEY	Mean		1.43	.061
	95% Confidence Interval for Mean	Lower Bound	1.31	
		Upper Bound	1.56	
	5% Trimmed Mean		1.37	
	Median		1.00	
	Variance		1.015	
	Std. Deviation		1.008	
	Minimum		0	
	Maximum		5	
	Range		5	
	Interquartile Range		1.00	
	Skewness		.742	.149
	Kurtosis		1.324	.296

The dependent variable "how many in family earned money" [earnrs] does not satisfy the criteria for a normal distribution.

The skewness (0.742) fell between -1.0 and +1.0, but the kurtosis (1.324) fell outside the range from -1.0 to +1.0.

多变量线性回归

Normality of dependent variable: how many in family earned money

Descriptives

			Statistic	Std. Error
Logarithm of EARNRS [LG10(1+EARNRS)]	Mean		.34676	.011783
	95% Confidence Interval for Mean	Lower Bound	.32356	
		Upper Bound	.36996	
	5% Trimmed Mean		.34693	
	Median		.30103	
	Variance		.037	
	Std. Deviation		.193257	
	Minimum		.000	
	Maximum		.778	
	Range		.778	
	Interquartile Range		.17609	
	Skewness		-.483	
	Kurtosis		-.309	

The logarithmic transformation improves the normality of "how many in family earned money" [earnrs]. In evaluating normality, the skewness (-0.483) and kurtosis (-0.309) were both within the range of acceptable values from -1.0 to +1.0.

The square root transformation also has values of skewness and kurtosis in the acceptable range.

However, by our order of preference for which transformation to use, the logarithm is preferred to the square root or inverse.

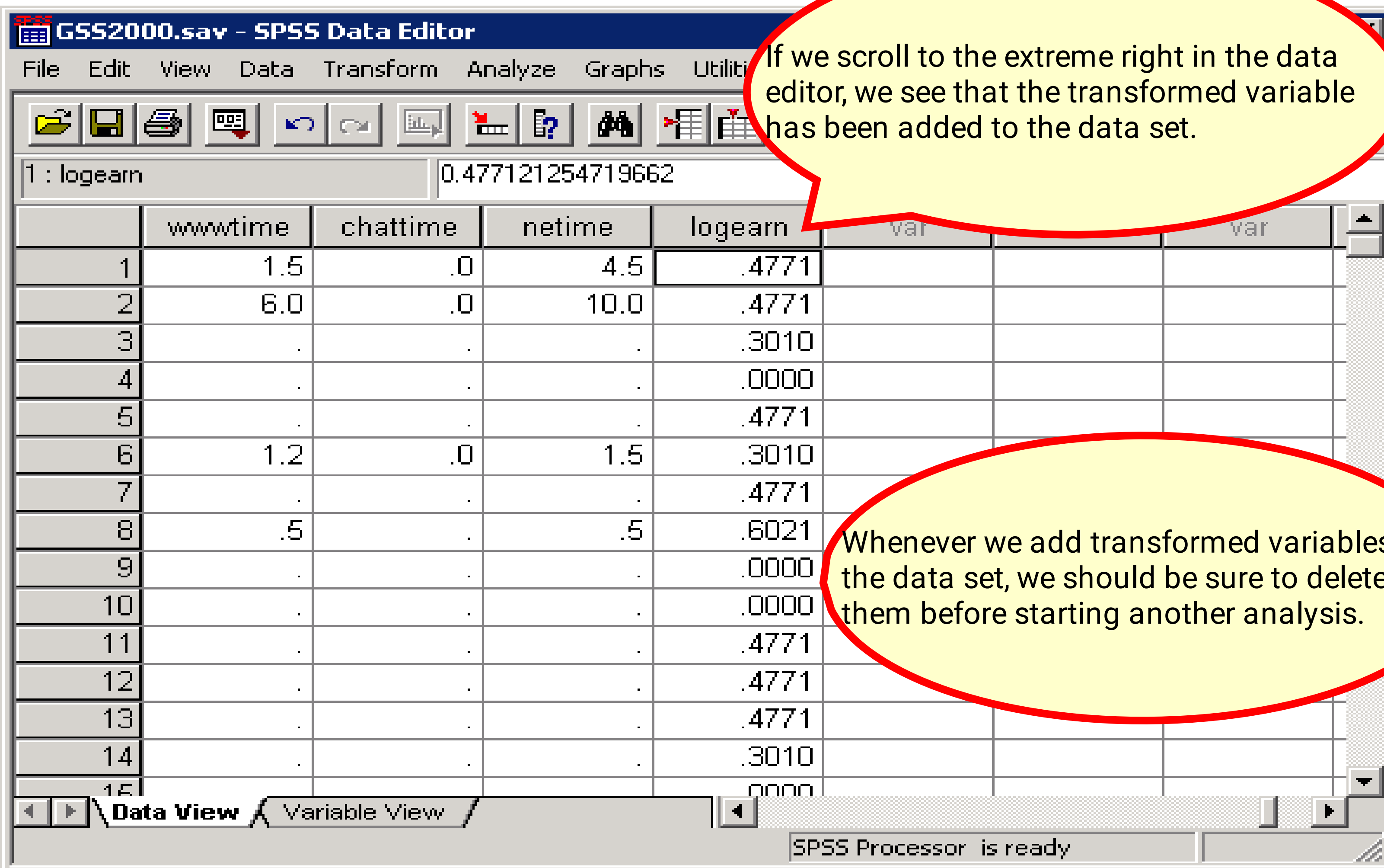
Transformation for how many in family earned money

- The logarithmic transformation improves the normality of "how many in family earned money" [earnrs].

- We will substitute the logarithmic transformation of how many in family earned money as the dependent variable in the regression analysis.

多变量线性回归

The transformed variable in the data editor



If we scroll to the extreme right in the data editor, we see that the transformed variable has been added to the data set.

Whenever we add transformed variables to the data set, we should be sure to delete them before starting another analysis.

多变量线性回归

Normality/linearity of independent variable: age

Descriptives

			Statistic	Std. Error
AGE OF RESPONDENT	Mean		45.99	1.023
	95% Confidence Interval for Mean	Lower Bound	43.98	
		Upper Bound	48.00	
	5% Trimmed Mean		45.31	
	Median		43.50	
	Variance		282.465	
	Std. Deviation		16.807	
	Minimum		19	
	Maximum		89	
	Range		70	
	Interquartile Range		24.00	
	Skewness		.595	.148
	Kurtosis		-.351	.295

In evaluating normality, the skewness (0.595) and kurtosis (-0.351) were both within the range of acceptable values from -1.0 to +1.0.

多变量线性回归

Normality/linearity of independent variable: age

Correlations

		Logarithm of EARNRS [LG10(1+EARNRS)]	AGE RESP DE
Logarithm of EARNRS [LG10(1+EARNRS)]	Pearson Correlation	1	
	Sig. (2-tailed)	.	
	N	269	
AGE OF RESPONDENT	Pearson Correlation	-.493**	
	Sig. (2-tailed)	.000	
	N	269	
Logarithm of AGE [LG10(AGE)]	Pearson Correlation	-.417**	
	Sig. (2-tailed)	.000	
	N	269	
Square of AGE [(AGE)**2]	Pearson Correlation	-.552**	
	Sig. (2-tailed)		
	N		
Square Root of AGE [SQRT(AGE)]			
Inverse of AGE [-1/(AGE)]			

**. Correlation is significant at the 0.01 le

The evidence of linearity in the relationship between the independent variable "age" [age] and the dependent variable "log transformation of how many in family earned money" [logearn] was the statistical significance of the correlation coefficient (r = -0.493). The probability for the correlation coefficient was <0.001, less than or equal to the level of significance of 0.01. We reject the null hypothesis that r = 0 and conclude that there is a linear relationship between the variables.

The independent variable "age" [age] satisfies the criteria for both the assumption of normality and the assumption of linearity with the dependent variable "log transformation of how many in family earned money" [logearn].

多变量线性回归

Normality/linearity of independent variable: respondent's socioeconomic index

Descriptives

			Statistic	Std. Error
RESPONDENT'S SOCIOECONOMIC INDEX	Mean		48.710	1.1994
	95% Confidence Interval for Mean	Lower Bound	46.348	
		Upper Bound	51.072	
	5% Trimmed Mean		47.799	
	Median		39.600	
	Variance		366.821	
	Std. Deviation		19.1526	
	Minimum		19.4	
	Maximum		97.2	
	Range		77.8	
	Interquartile Range		31.100	
	Skewness		.585	.153
	Kurtosis		-.862	.304

The independent variable "respondent's socioeconomic index" [sei] satisfies the criteria for the assumption of normality, but does not satisfy the assumption of linearity with the dependent variable "log transformation of how many in family earned money" [logearn].

In evaluating normality, the skewness (0.585) and kurtosis (-0.862) were both within the range of acceptable values from -1.0 to +1.0.

多变量线性回归

Normality/linearity of independent variable: respondent's socioeconomic index

Correlations

		Logarithm of EARNRS [LG10(1+EARNRS)]	RESPONDENTS SOCIOECONOMIC INDEX
Logarithm of EARNRS [LG10(1+EARNRS)]	Pearson Correlation	1	.05
	Sig. (2-tailed)	.	
	N	269	
RESPONDENT'S SOCIOECONOMIC INDEX	Pearson Correlation	.055	
	Sig. (2-tailed)	.385	
	N	254	
Logarithm of SEI [LG10(SEI)]	Pearson Correlation	.073	
	Sig. (2-tailed)	.243	
	N	254	
Square of SEI [(SEI)**2]	Pearson Correlation	.036	
	Sig. (2-tailed)	.563	
	N	254	
Square Root of SEI [SQRT(SEI)]	Pearson Correlation	.064	
	Sig. (2-tailed)	.309	
	N	254	
Inverse of SEI [-1/(SEI)]	Pearson Correlation	.092	.9
	Sig. (2-tailed)	.142	.00
	N	254	255

**. Correlation is significant at the 0.01 level (2-tailed).

The probability for the correlation coefficient was 0.385, greater than the level of significance of 0.01. We cannot reject the null hypothesis that r = 0, and cannot conclude that there is a linear relationship between the variables.

Since none of the transformations to improve linearity were successful, it is an indication that the problem may be a weak relationship, rather than a curvilinear relationship correctable by using a transformation. A weak relationship is not a violation of the assumption of linearity, and does not require a caution.

多变量线性回归

Homoscedasticity of independent variable: Sex

Descriptives

	RESPONDENTS SEX		Statistic	Std. Error
Logarithm of EARNRS [LG10(1+EARNRS)]	1	Mean	.339924	.0186863
		Variance	.038	
		Std. Deviation	.1959831	
	2	Mean	[illegible]	[illegible]
		Variance	[illegible]	
		Std. Deviation	[illegible]	

Oneway

Test of Homogeneity of Variances

Logarithm of EARNRS [LG10(1+EARNRS)]

Levene Statistic	df1	df2	Sig.
.088	1	267	.767

Based on the Levene Test, the variance in "log transformation of how many in family earned money" [logearn] is homogeneous for the categories of "sex" [sex].

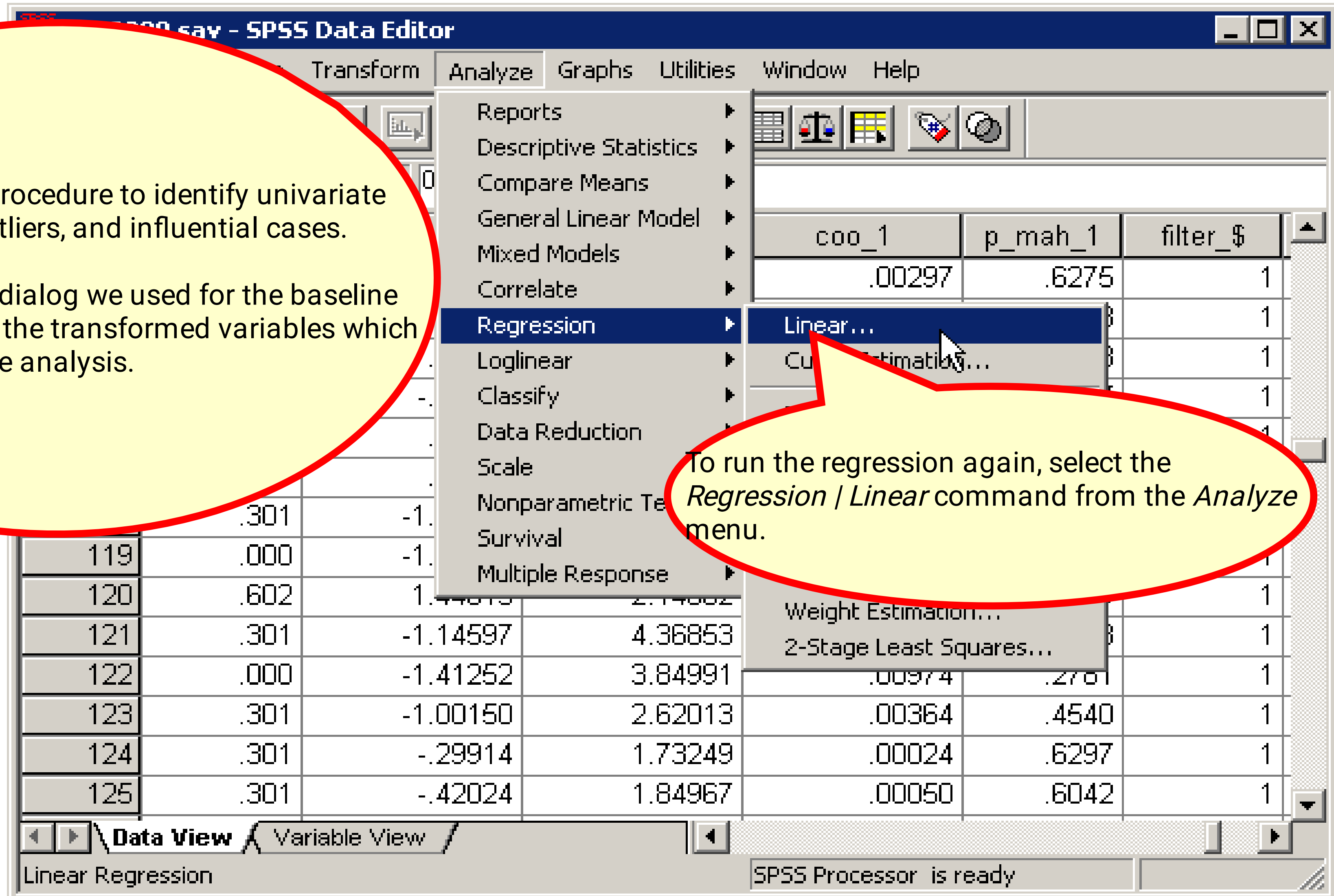
The probability associated with the Levene Statistic (0.767) is greater than the level of significance, so we fail to reject the null hypothesis and conclude that the homoscedasticity assumption is satisfied.

多变量线性回归

The regression to identify outliers and influential cases

We use the regression procedure to identify univariate outliers, multivariate outliers, and influential cases.

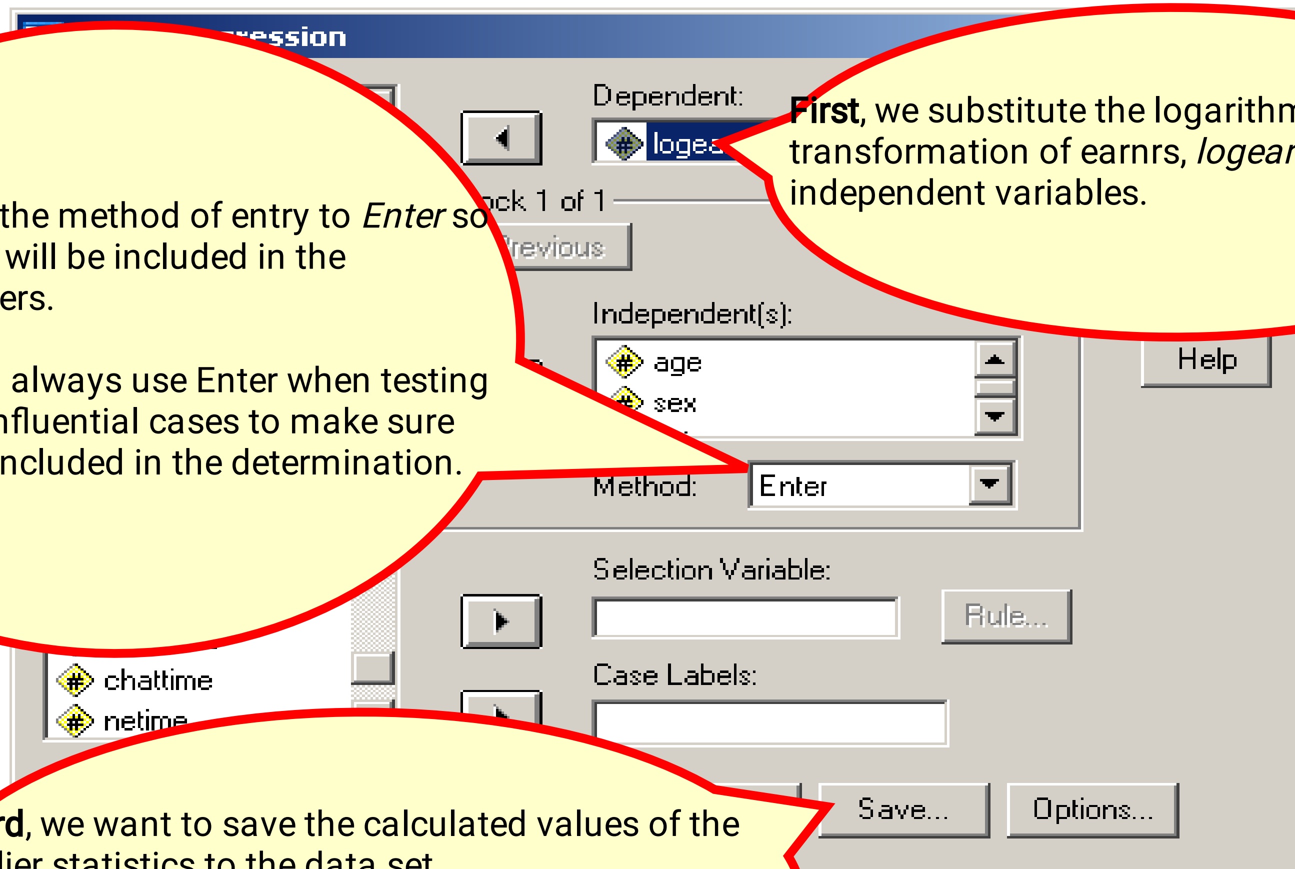
We start with the same dialog we used for the baseline analysis and substitute the transformed variables which we think will improve the analysis.



To run the regression again, select the *Regression | Linear* command from the *Analyze* menu.

多变量线性回归

The regression to identify outliers and influential cases



First, we substitute the logarithmic transformation of earnrs, *logearn*, into the list of independent variables.

Second, we keep the method of entry to *Enter* so that all variables will be included in the detection of outliers.

NOTE: we should always use Enter when testing for outliers and influential cases to make sure all variables are included in the determination.

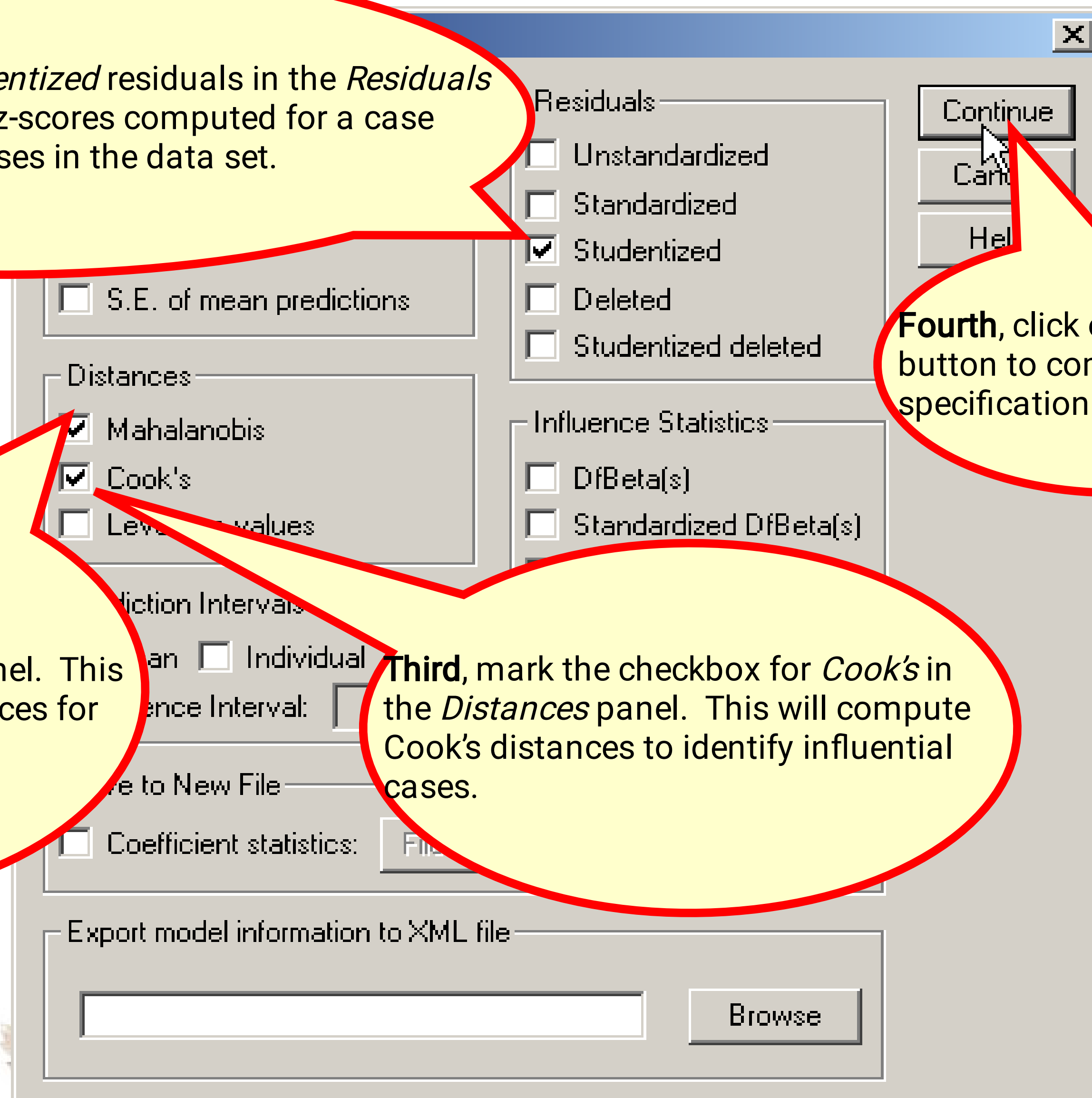
Third, we want to save the calculated values of the outlier statistics to the data set.

Click on the *Save...* button to specify what we want to save.

多变量线性回归

Saving the measures of outliers/influential cases

First, mark the checkbox for *Studentized* residuals in the *Residuals* panel. Studentized residuals are z-scores computed for a case based on the data for all other cases in the data set.



Fourth, click on the *OK* button to complete the specifications.

Second, mark the checkbox for *Mahalanobis* in the *Distances* panel. This will compute Mahalanobis distances for the set of independent variables.

Third, mark the checkbox for *Cook's* in the *Distances* panel. This will compute Cook's distances to identify influential cases.

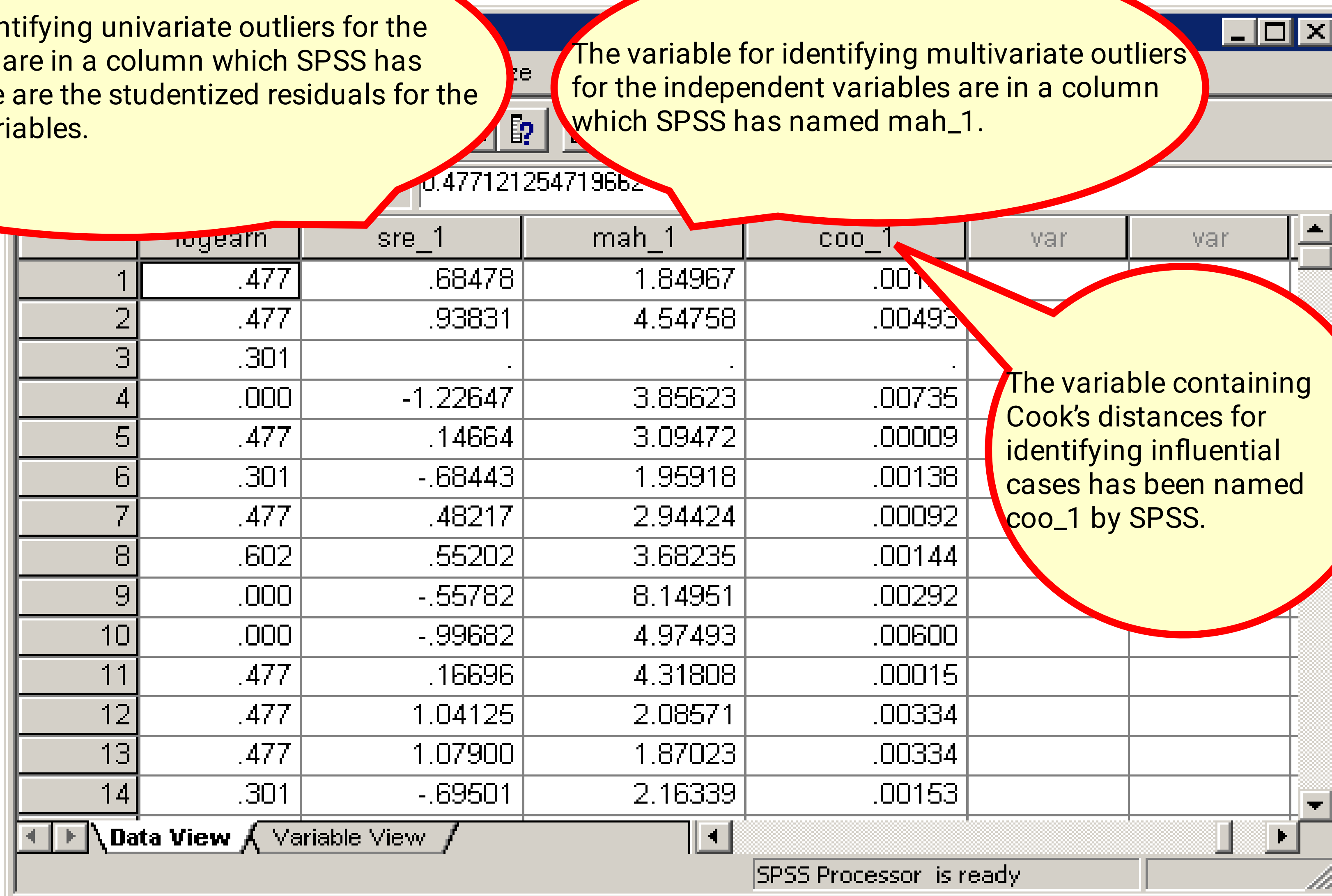
多变量线性回归

The variables for identifying outliers/influential cases

The variable for identifying univariate outliers for the dependent variable are in a column which SPSS has named sre_1. These are the studentized residuals for the log transformed variables.

The variable for identifying multivariate outliers for the independent variables are in a column which SPSS has named mah_1.

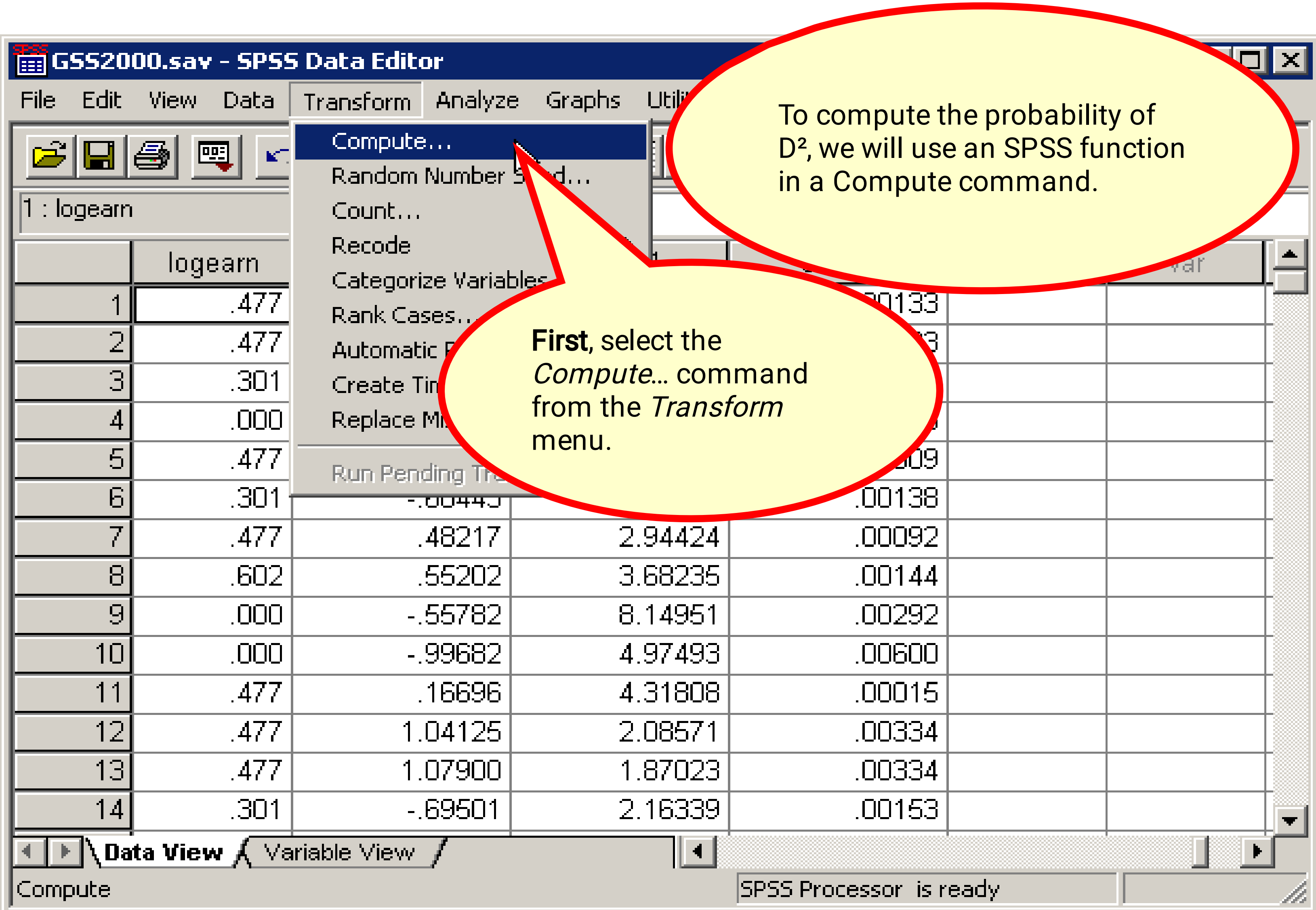
The variable containing Cook's distances for identifying influential cases has been named coo_1 by SPSS.


0.47712125471966

	logearn	sre_1	mah_1	coo_1	var	var
1	.477	.68478	1.84967	.001		
2	.477	.93831	4.54758	.00493		
3	.301	.	.	.		
4	.000	-1.22647	3.85623	.00735		
5	.477	.14664	3.09472	.00009		
6	.301	-.68443	1.95918	.00138		
7	.477	.48217	2.94424	.00092		
8	.602	.55202	3.68235	.00144		
9	.000	-.55782	8.14951	.00292		
10	.000	-.99682	4.97493	.00600		
11	.477	.16696	4.31808	.00015		
12	.477	1.04125	2.08571	.00334		
13	.477	1.07900	1.87023	.00334		
14	.301	-.69501	2.16339	.00153		

多变量线性回归

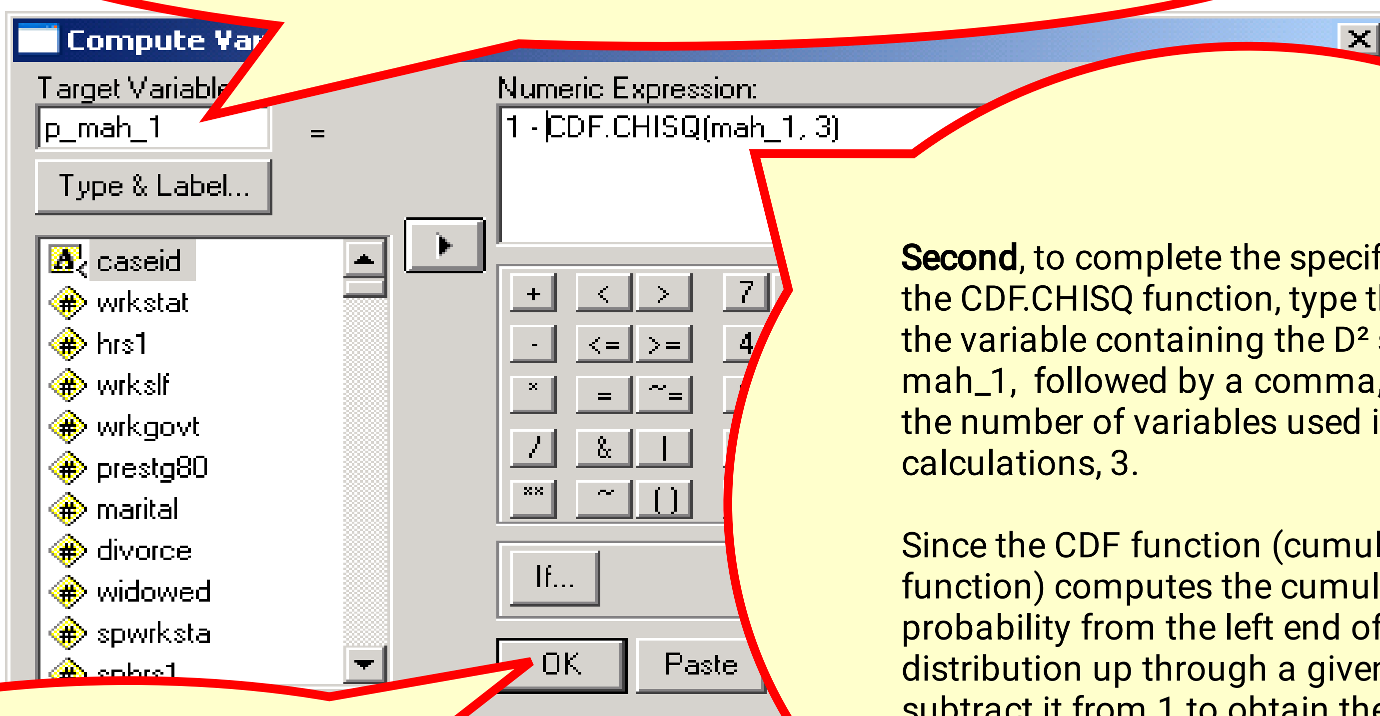
Computing the probability for Mahalanobis D²



多变量线性回归

Formula for probability for Mahalanobis D²

First, in the *target variable* text box, type the name " p_mah_1" as an acronym for the probability of the mah_1, the Mahalanobis D² score.



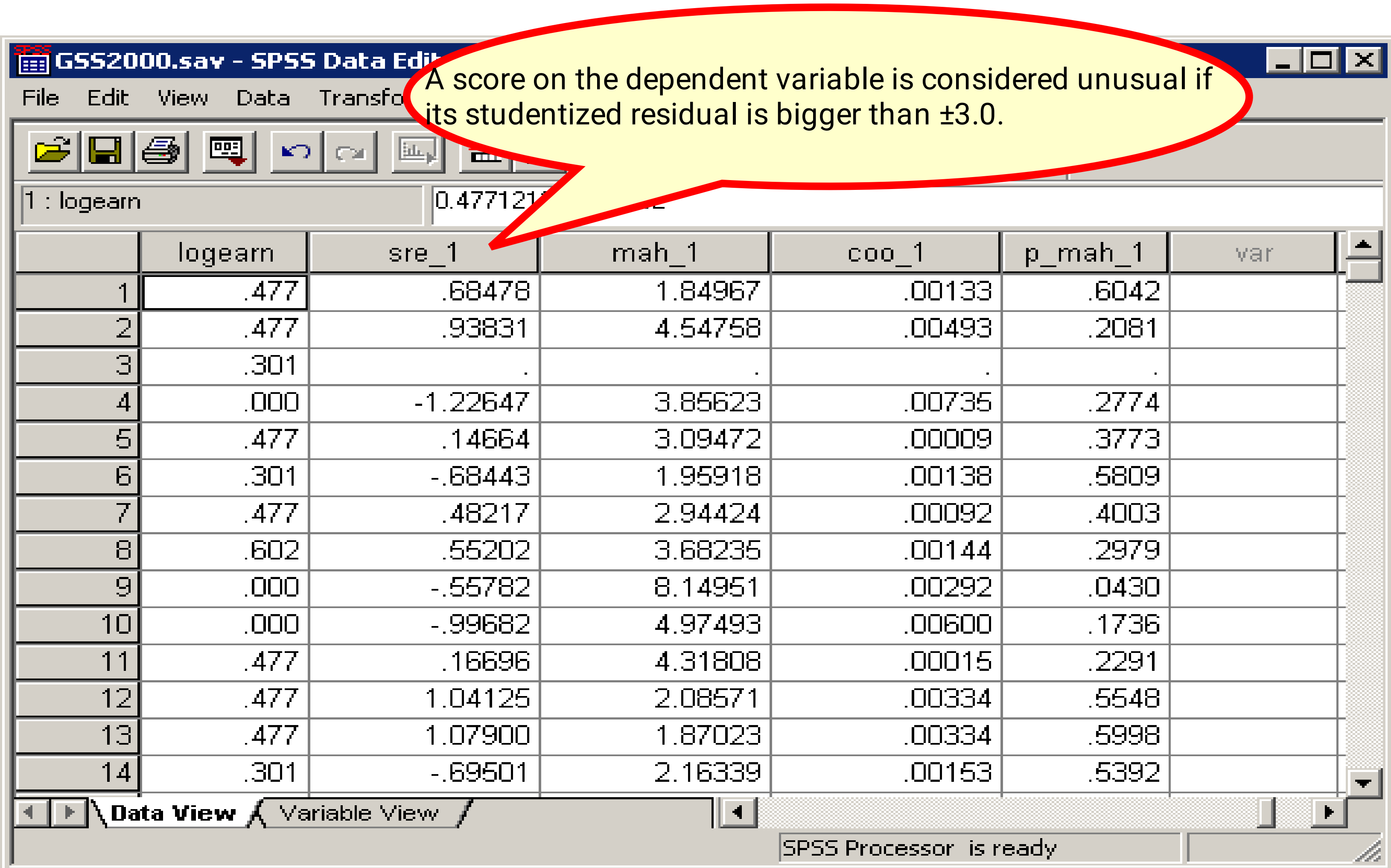
Second, to complete the specifications for the CDF.CHISQ function, type the name of the variable containing the D² scores, mah_1, followed by a comma, followed by the number of variables used in the calculations, 3.

Since the CDF function (cumulative density function) computes the cumulative probability from the left end of the distribution up through a given value, we subtract it from 1 to obtain the probability in the upper tail of the distribution.

Third, click on the *OK* button to signal completion of the computer variable dialog.

多变量线性回归

Univariate outliers

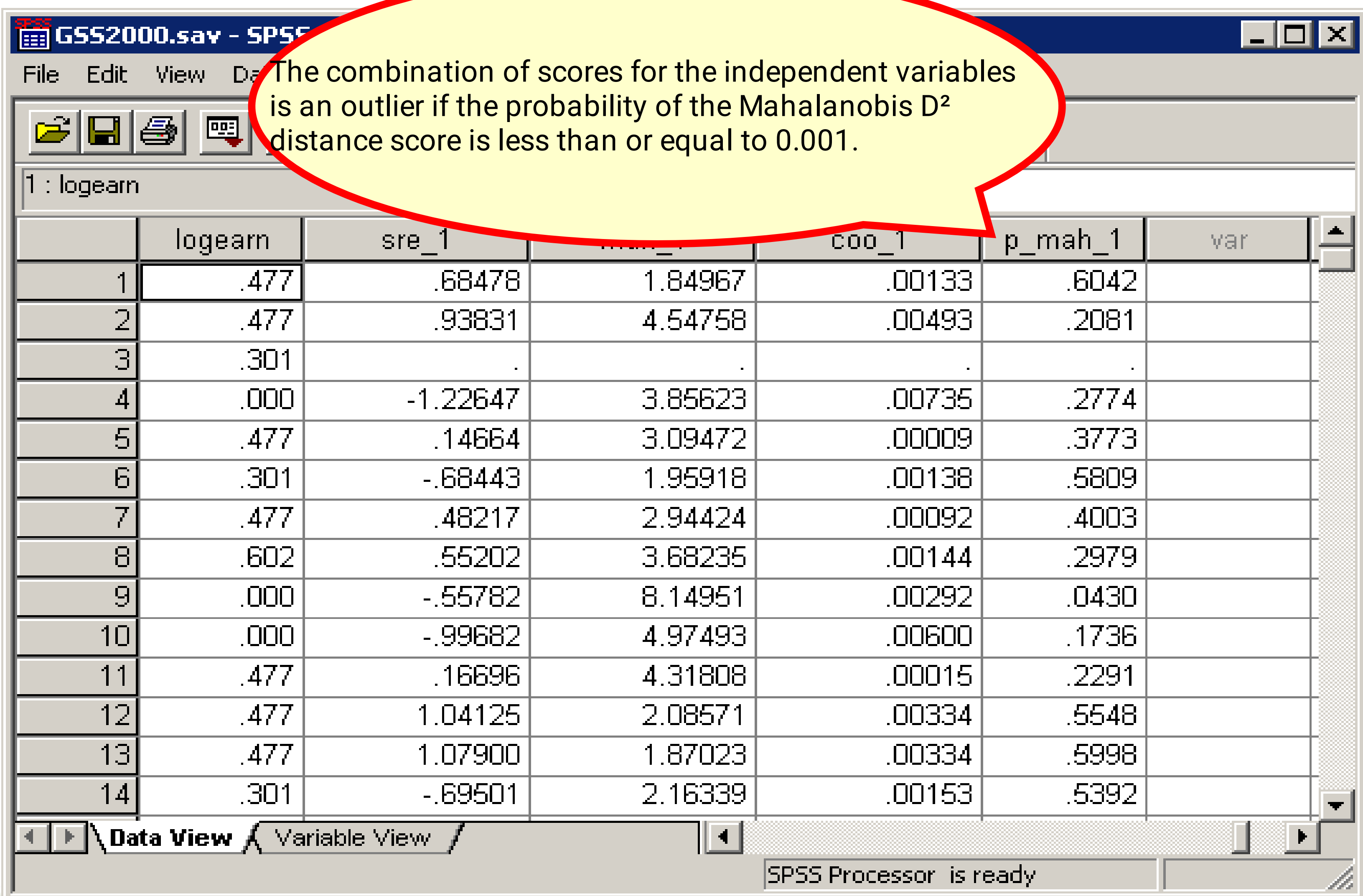


	logearn	sre_1	mah_1	coo_1	p_mah_1	var
1	.477	.68478	1.84967	.00133	.6042	
2	.477	.93831	4.54758	.00493	.2081	
3	.301	.	.	.	.	
4	.000	-1.22647	3.85623	.00735	.2774	
5	.477	.14664	3.09472	.00009	.3773	
6	.301	-.68443	1.95918	.00138	.5809	
7	.477	.48217	2.94424	.00092	.4003	
8	.602	.55202	3.68235	.00144	.2979	
9	.000	-.55782	8.14951	.00292	.0430	
10	.000	-.99682	4.97493	.00600	.1736	
11	.477	.16696	4.31808	.00015	.2291	
12	.477	1.04125	2.08571	.00334	.5548	
13	.477	1.07900	1.87023	.00334	.5998	
14	.301	-.69501	2.16339	.00153	.5392	

多变量线性回归

Multivariate outliers

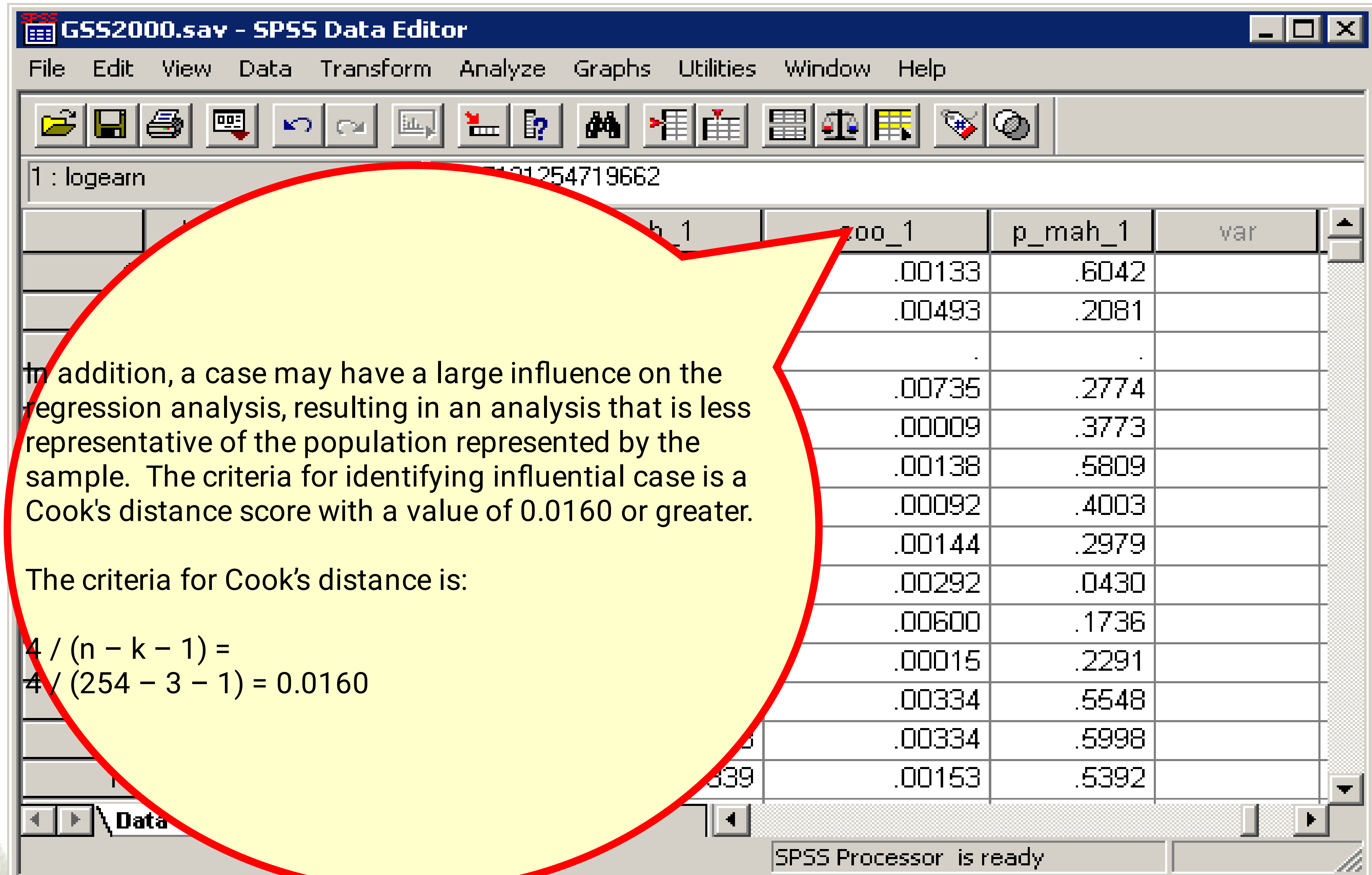
The combination of scores for the independent variables is an outlier if the probability of the Mahalanobis D² distance score is less than or equal to 0.001.



	logearn	sre_1	[illegible]	coo_1	p_mah_1	var
1	.477	.68478	1.84967	.00133	.6042	
2	.477	.93831	4.54758	.00493	.2081	
3	.301	.	.	.	.	
4	.000	-1.22647	3.85623	.00735	.2774	
5	.477	.14664	3.09472	.00009	.3773	
6	.301	-.68443	1.95918	.00138	.5809	
7	.477	.48217	2.94424	.00092	.4003	
8	.602	.55202	3.68235	.00144	.2979	
9	.000	-.55782	8.14951	.00292	.0430	
10	.000	-.99682	4.97493	.00600	.1736	
11	.477	.16696	4.31808	.00015	.2291	
12	.477	1.04125	2.08571	.00334	.5548	
13	.477	1.07900	1.87023	.00334	.5998	
14	.301	-.69501	2.16339	.00153	.5392	

多变量线性回归

Influential cases



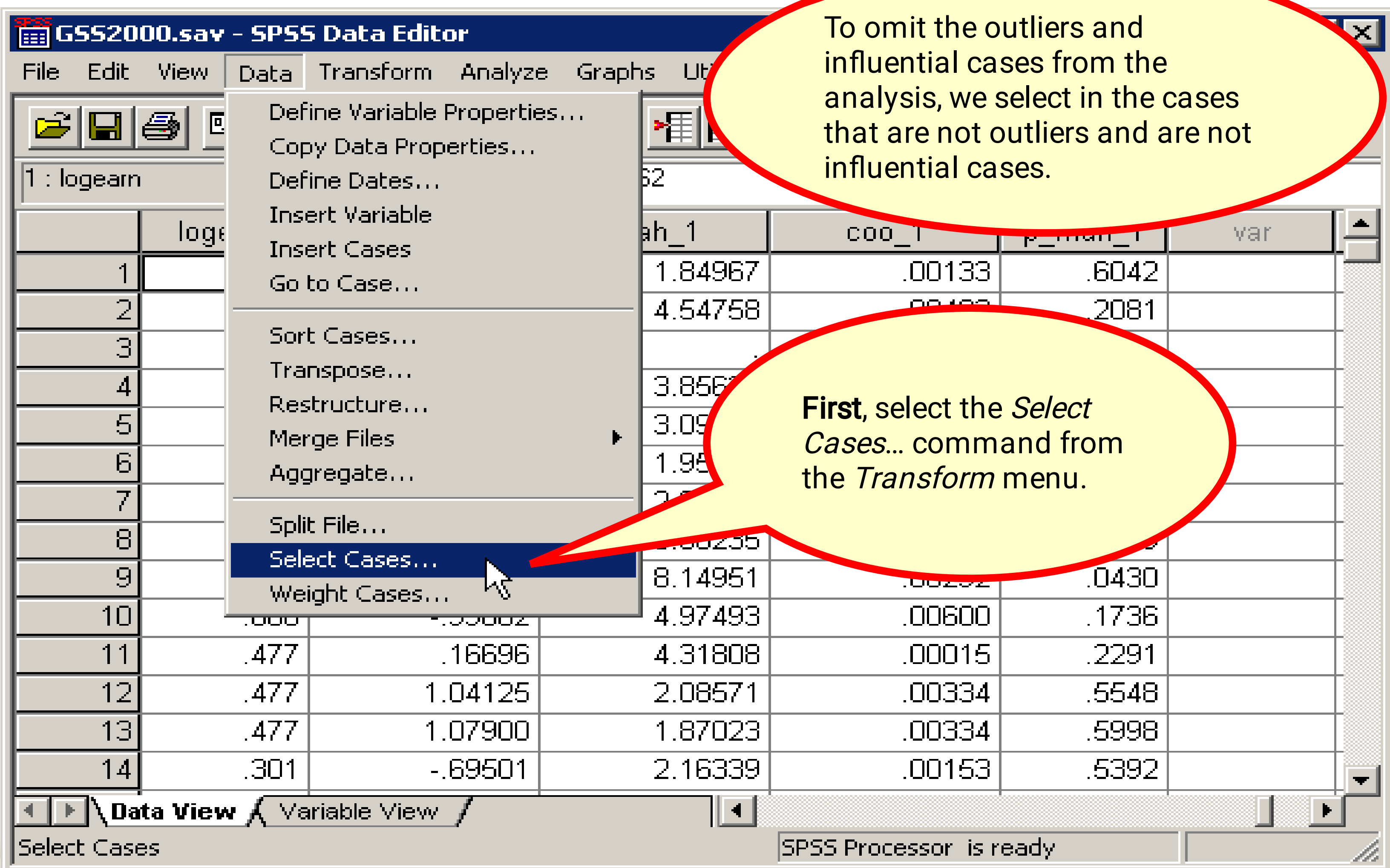
In addition, a case may have a large influence on the regression analysis, resulting in an analysis that is less representative of the population represented by the sample. The criteria for identifying influential case is a Cook's distance score with a value of 0.0160 or greater.

The criteria for Cook's distance is:

$4 / (n - k - 1) =$
$4 / (254 - 3 - 1) = 0.0160$

多变量线性回归

Omitting the outliers and influential cases

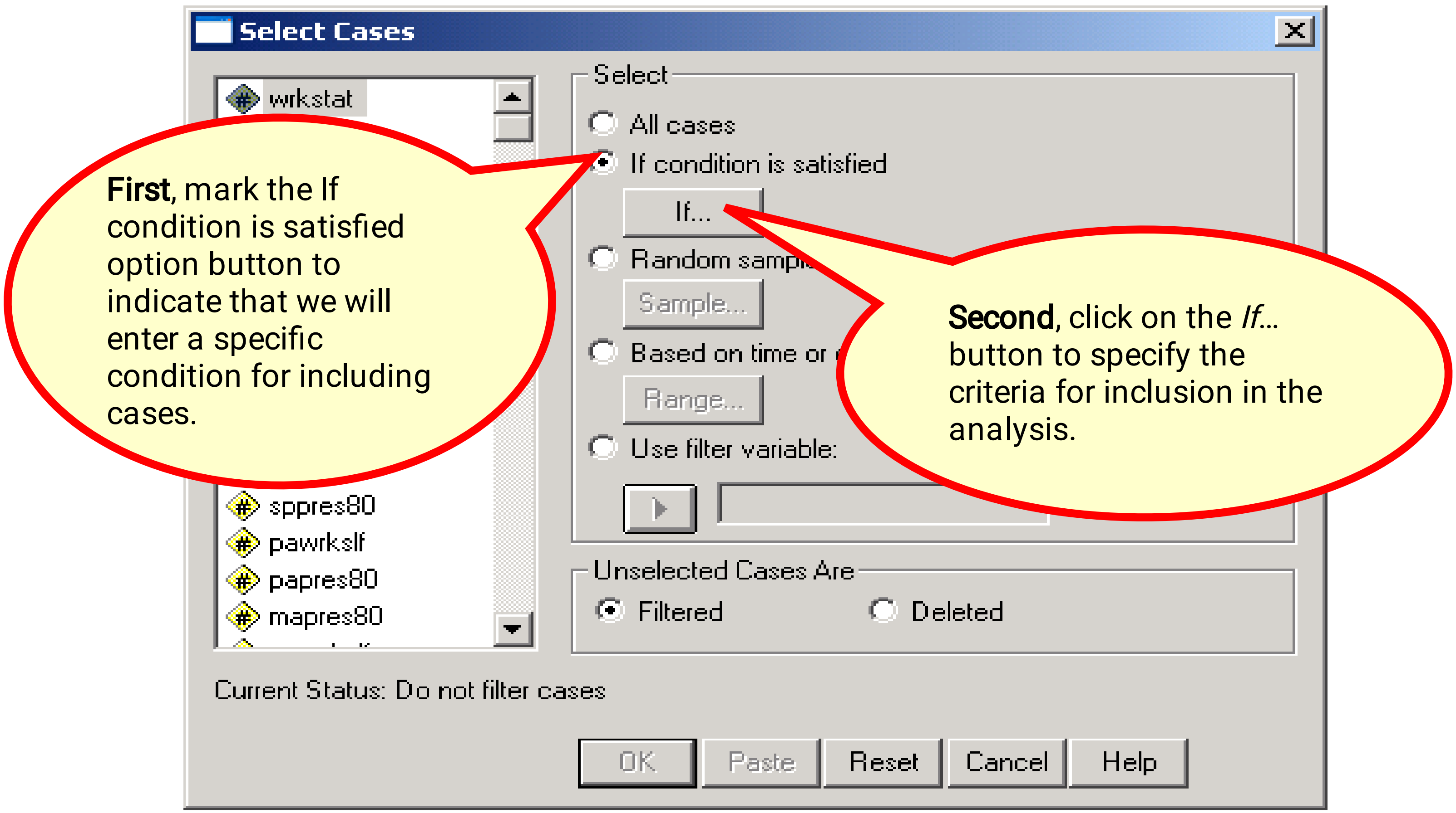


To omit the outliers and influential cases from the analysis, we select in the cases that are not outliers and are not influential cases.

First, select the *Select Cases...* command from the *Transform* menu.

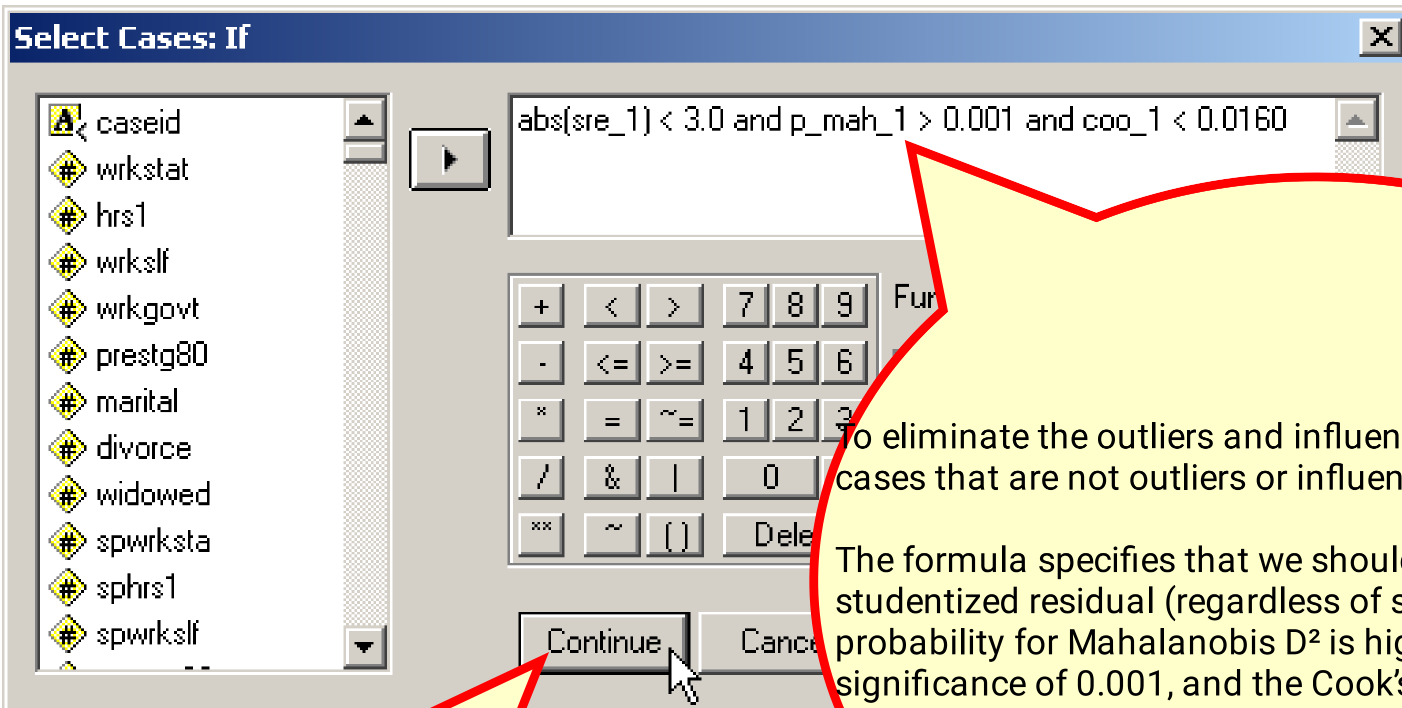
多变量线性回归

Specifying the condition to omit outliers



多变量线性回归

The formula for omitting outliers



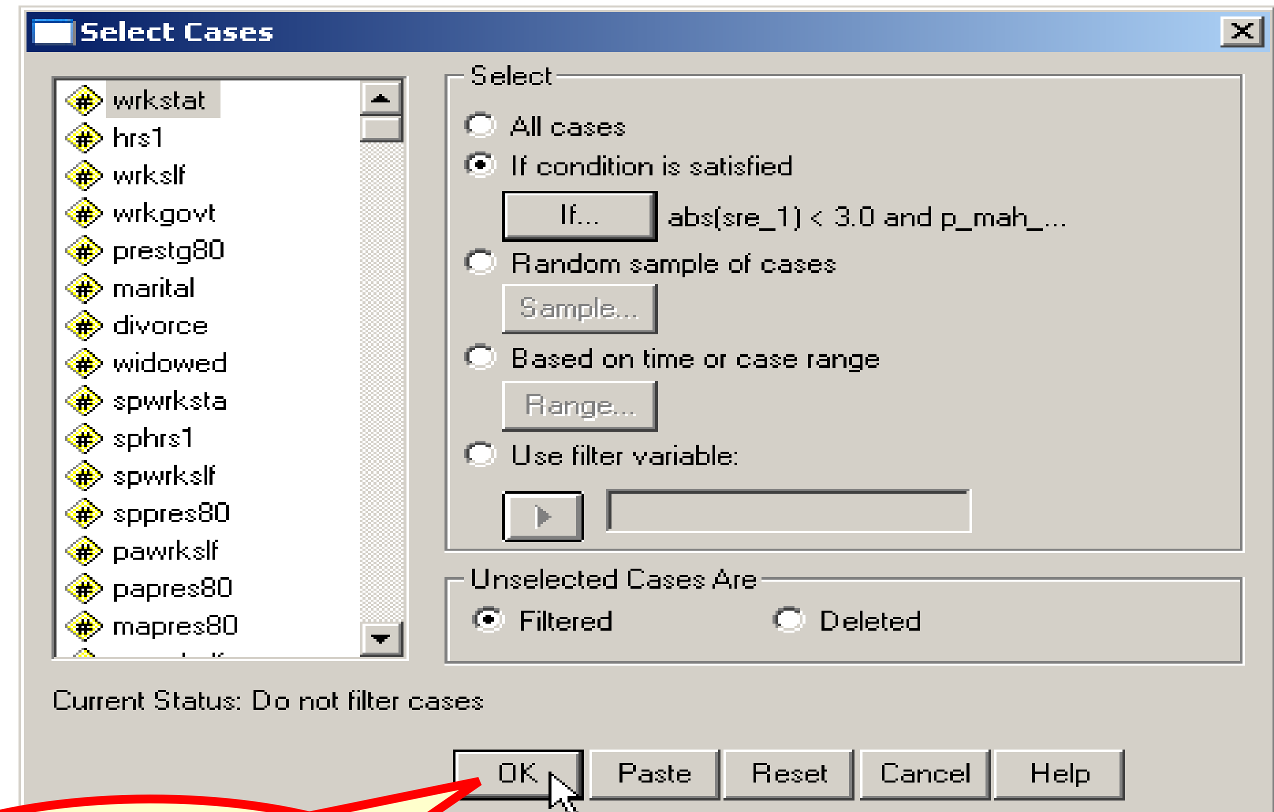
To eliminate the outliers and influential cases, we request the cases that are not outliers or influential cases.

The formula specifies that we should include cases if the studentized residual (regardless of sign) is less than 3, the probability for Mahalanobis D² is higher than the level of significance of 0.001, and the Cook's distance value is less than the critical value of 0.0160.

After typing in the formula, click on the *Continue* button to close the dialog box,

多变量线性回归

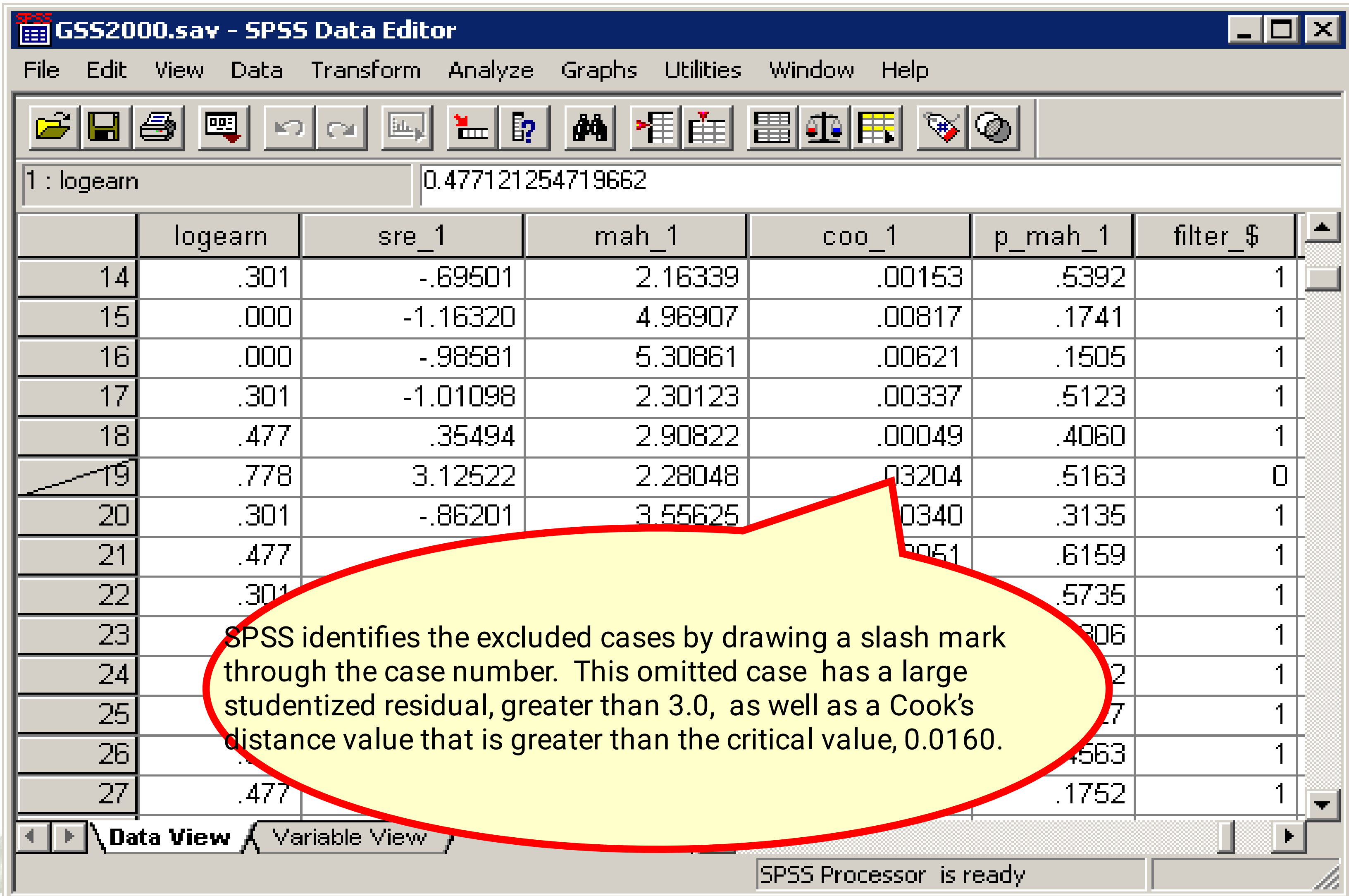
Completing the request for the selection



To complete the request, we click on the *OK* button.

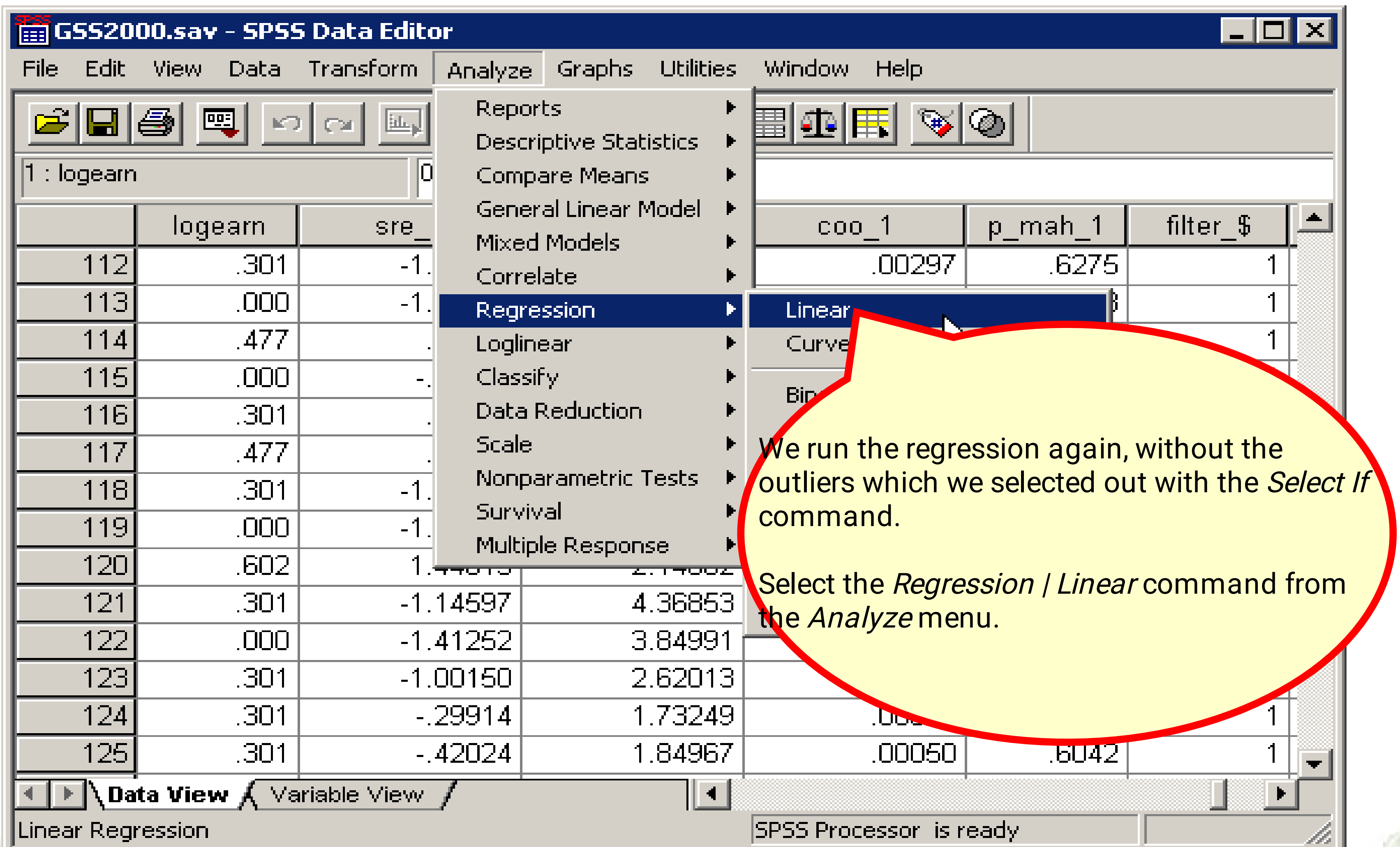
多变量线性回归

An omitted outlier and influential case



多变量线性回归

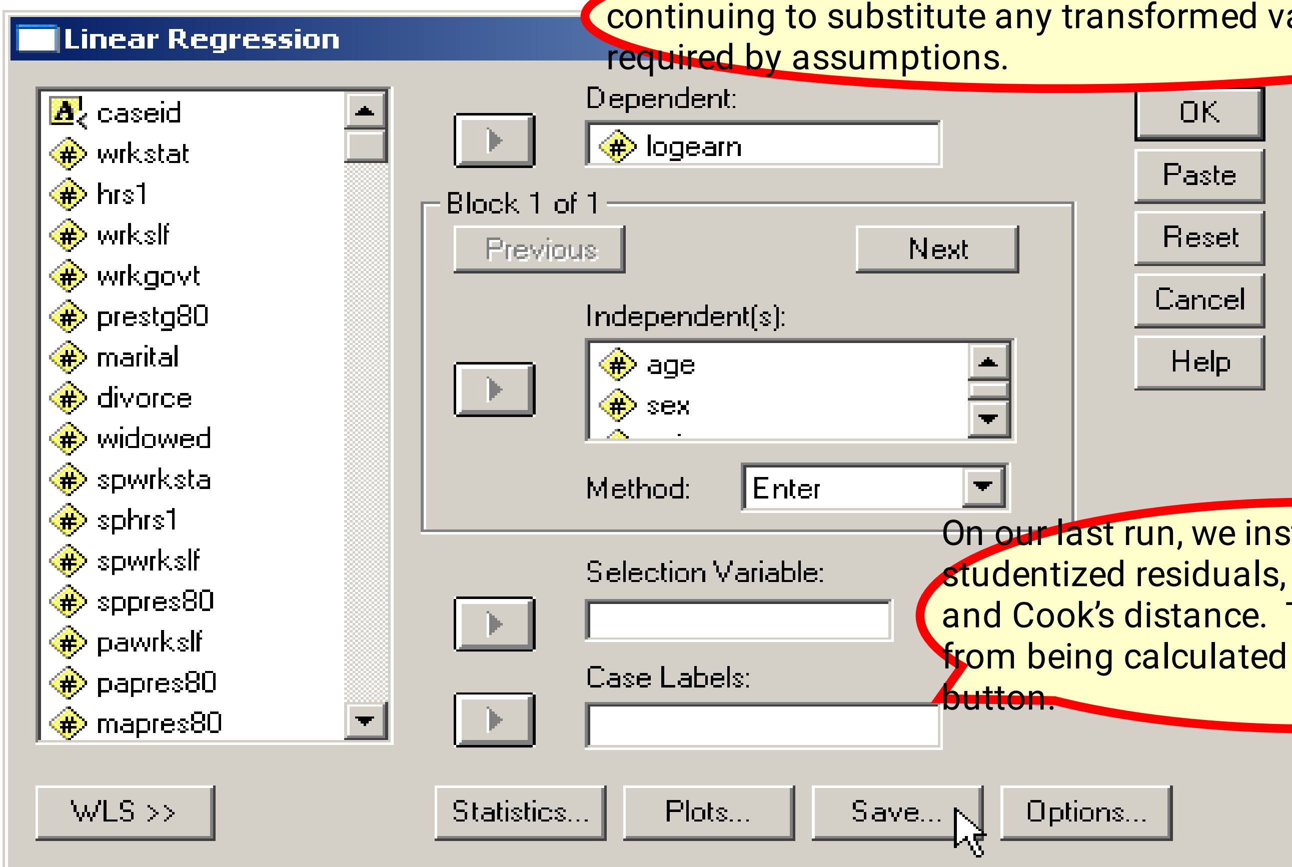
Running the regression omitting outliers



多变量线性回归

Opening the save options dialog

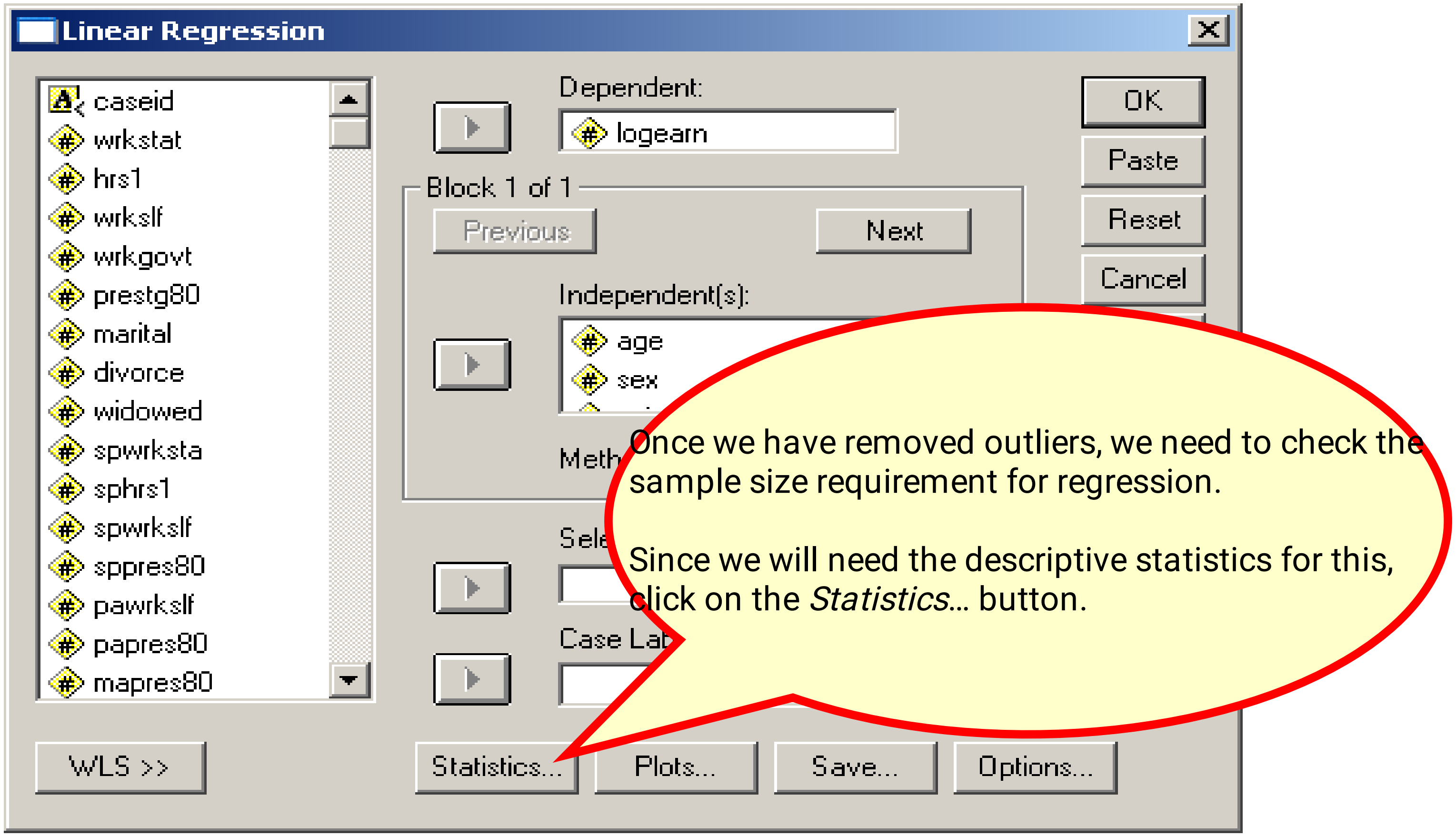
We specify the dependent and independent variables, continuing to substitute any transformed variables required by assumptions.



On our last run, we instructed SPSS to save studentized residuals, Mahalanobis distance, and Cook's distance. To prevent these values from being calculated again, click on the *Save...* button.

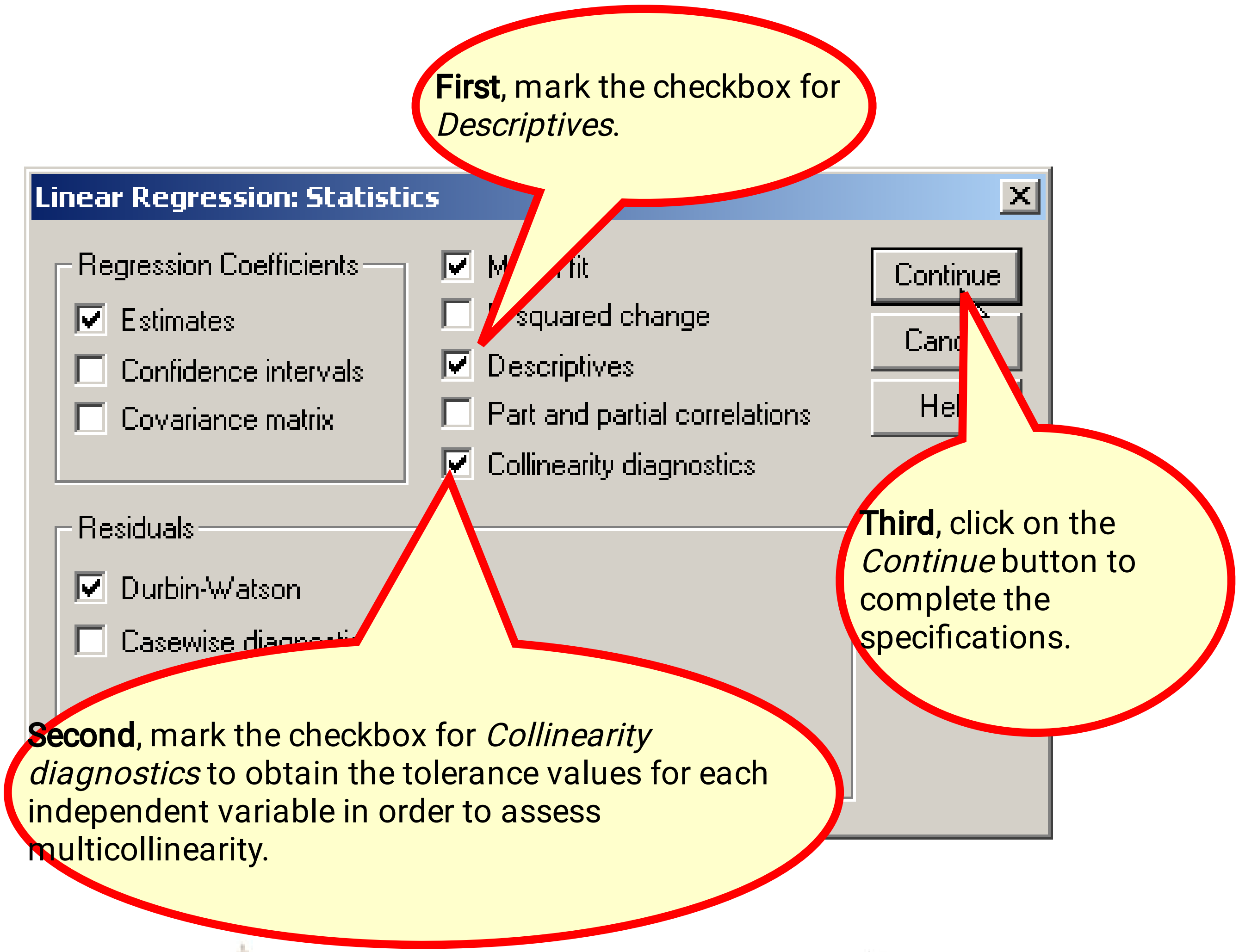
多变量线性回归

Opening the statistics options dialog



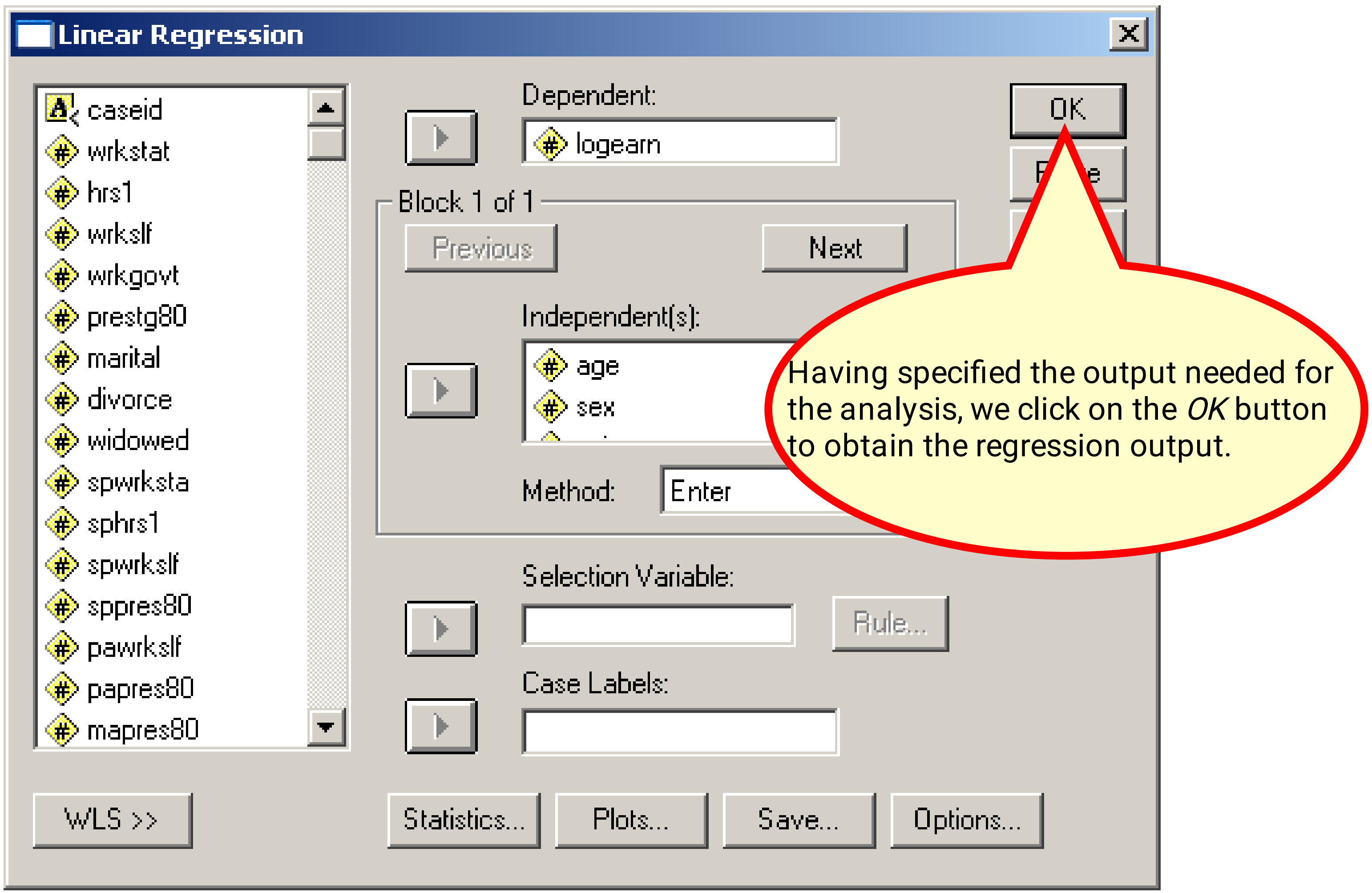
多变量线性回归

Requesting descriptive statistics



多变量线性回归

Requesting the output



多变量线性回归

Selection of model for interpretation

Prior to any transformations of variables to satisfy the assumptions of multiple regression and the removal of outliers and influential cases, the proportion of variance in the dependent variable explained by the independent variables (R^2) was 18.7%.

After substituting transformed variables and removing outliers and influential cases, the proportion of variance in the dependent variable explained by the independent variables (R^2) was 38.4%.

Model S…

Model	R	R Square	…ed …quare	Std. Error of the Estimate
1	.620[a]	.384	.377	.1457258

a. Predictors: (Constant), SEI, AGE, SEX

b. Dependent Variable: LOGEARN

Since the regression analysis using transformations and omitting outliers and influential cases explained at least two percent more variance than the regression analysis with all cases and no transformations, the regression analysis with transformed variables omitting outliers and influential cases was interpreted.

多变量线性回归

Sample size

The minimum ratio of valid cases to independent variables for multiple regression is 5 to 1. After removing 6 influential cases or outliers, there are 248 valid cases and 3 independent variables.

The ratio of cases to independent variables for this analysis is 82.67 to 1, which satisfies the minimum requirement. In addition, the ratio of 82.67 to 1 satisfies the preferred ratio of 15 to 1.

Descriptive Statistics

	Mean	Std. Deviation	N
LOGEARN	.354289	.1845814	248
AGE	46.70	16.677	248
SEX	1.57	.496	248
SEI	48.819	19.1071	248

多变量线性回归

Overall relationship between independent and dependent variables

The probability of the F statistic (50.759) for the overall regression relationship is <0.001, less than or equal to the level of significance of 0.05. We reject the null hypothesis that there is no relationship between the set of independent variables and the dependent variable ($R^2 = 0$). We support the research hypothesis that there is a statistically significant relationship between the set of independent variables and the dependent variable.

We support the research hypothesis that there is a statistically significant relationship between the set of independent variables and the dependent variable.

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.234	3	1.078	50.759	.000[a]
	Residual	5.182	244	.021		
	Total	8.415	247			

a. Predictors: (Constant), SEI, AGE, SEX

b. Dependent Variable: LOGEARN

多变量线性回归

Overall relationship between independent and dependent variables

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.620[a]	.384	.377	.1457258

...rs: (Constant), SEI, AGE, SEX

...ent Variable: LOGEARN

The Multiple R for the relationship between the set of independent variables and the dependent variable is 0.620, which would be characterized as strong using the rule of thumb than a correlation less than or equal to 0.20 is characterized as very weak; greater than 0.20 and less than or equal to 0.40 is weak; greater than 0.40 and less than or equal to 0.60 is moderate; greater than 0.60 and less than or equal to 0.80 is strong; and greater than 0.80 is very strong.

ANOVA[b]

Sum of Squares	df	Mean Square	F	Sig.
3.234	3	1.078	50.759	.000[a]
5.182	244	.021		
8.415	247			

...tant), SEI, AGE, SEX

...able: LOGEARN

多变量线性回归

Multicollinearity

Coefficients [a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	.626	.048		12.989	.000		
	AGE	-.007	.001	-.615	-12.237	.000	.999	1.001
	SEX	.024	.019	.065	1.284	.200	.997	1.003
	SEI	.000	.000	.018	.354	.724	.997	1.004

a. Dependent Variable: LOGEARN

Multicollinearity occurs when one independent variable is so strongly correlated with one or more other variables that its relationship to the dependent variable is likely to be misinterpreted. Its potential unique contribution to explaining the dependent variable is minimized by its strong relationship to other independent variables. Multicollinearity is indicated when the tolerance value for an independent variable is less than 0.10.

The tolerance values for all of the independent variables are larger than 0.10. Multicollinearity is not a problem in this regression analysis.

多变量logistic回归

- Y_i's is Binary (Dichotomous)
- Y_i's ~ Bernoulli(μ_i), where μ_i=E(Y_i)=P(Y_i=1)
- X_i's can be continue variables or category variables

$$\mu_i = E(Y_i) = P(Y_i = 1)$$

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

多变量logistic回归

Assumptions

- Logistic regression does not make any assumptions of normality, linearity, and homogeneity of variance for the independent variables
- When the variables satisfy the assumptions of normality, linearity, and homogeneity of variance, discriminant analysis is generally cited as the more effective statistical procedure for evaluating relationships with a non-metric dependent variable
- When the variables do not satisfy the assumptions of normality, linearity, and homogeneity of variance, logistic regression is the statistic of choice since it does not make these assumptions

多变量logistic回归

Sample size requirements

- The minimum number of cases per independent variable is 10, using a guideline provided by Hosmer and Lemeshow, authors of *Applied Logistic Regression*, one of the main resources for Logistic Regression.

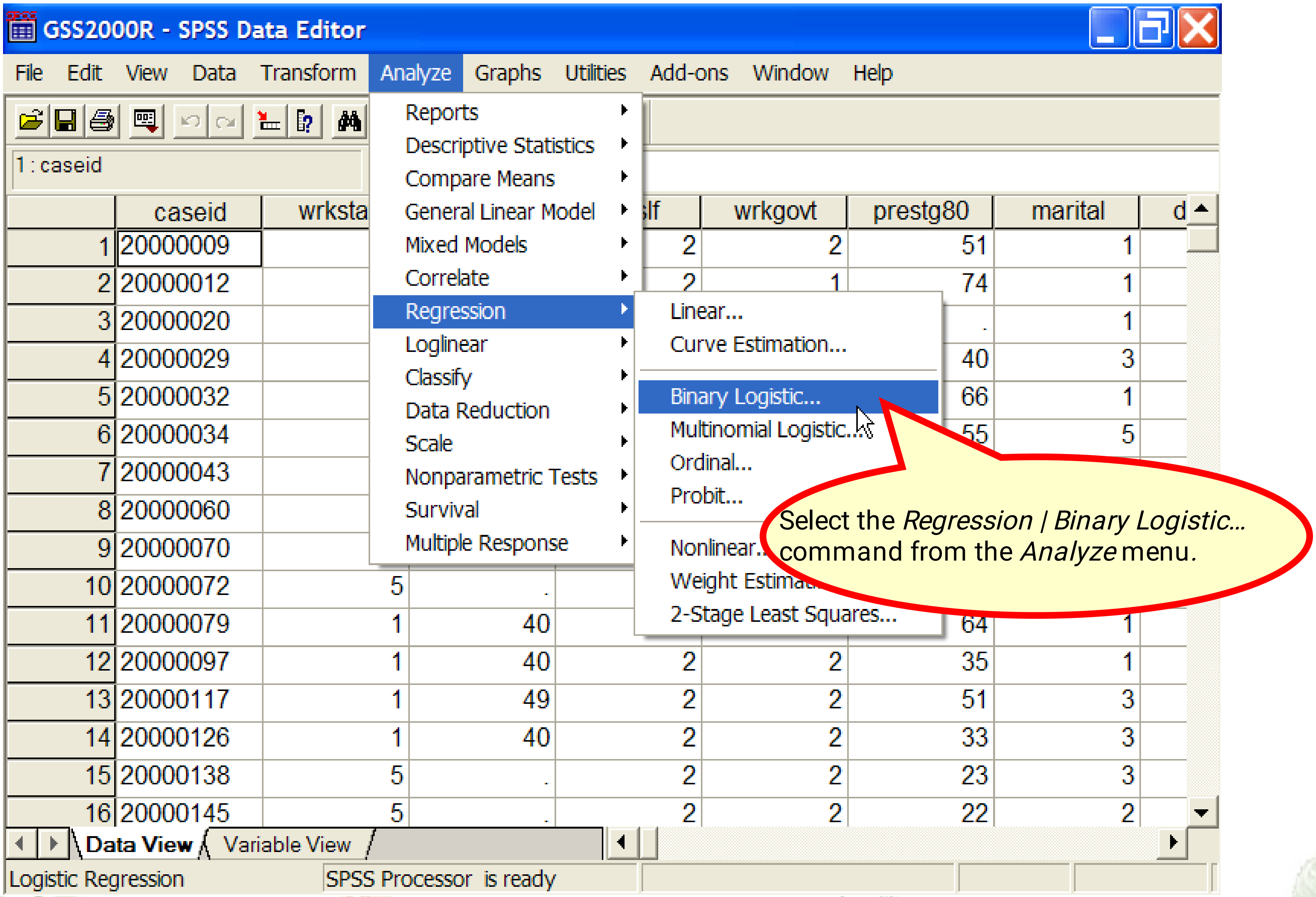
- For preferred case-to-variable ratios, we will use 20 to 1 for simultaneous and hierarchical logistic regression and 50 to 1 for stepwise logistic regression.

多变量logistic回归

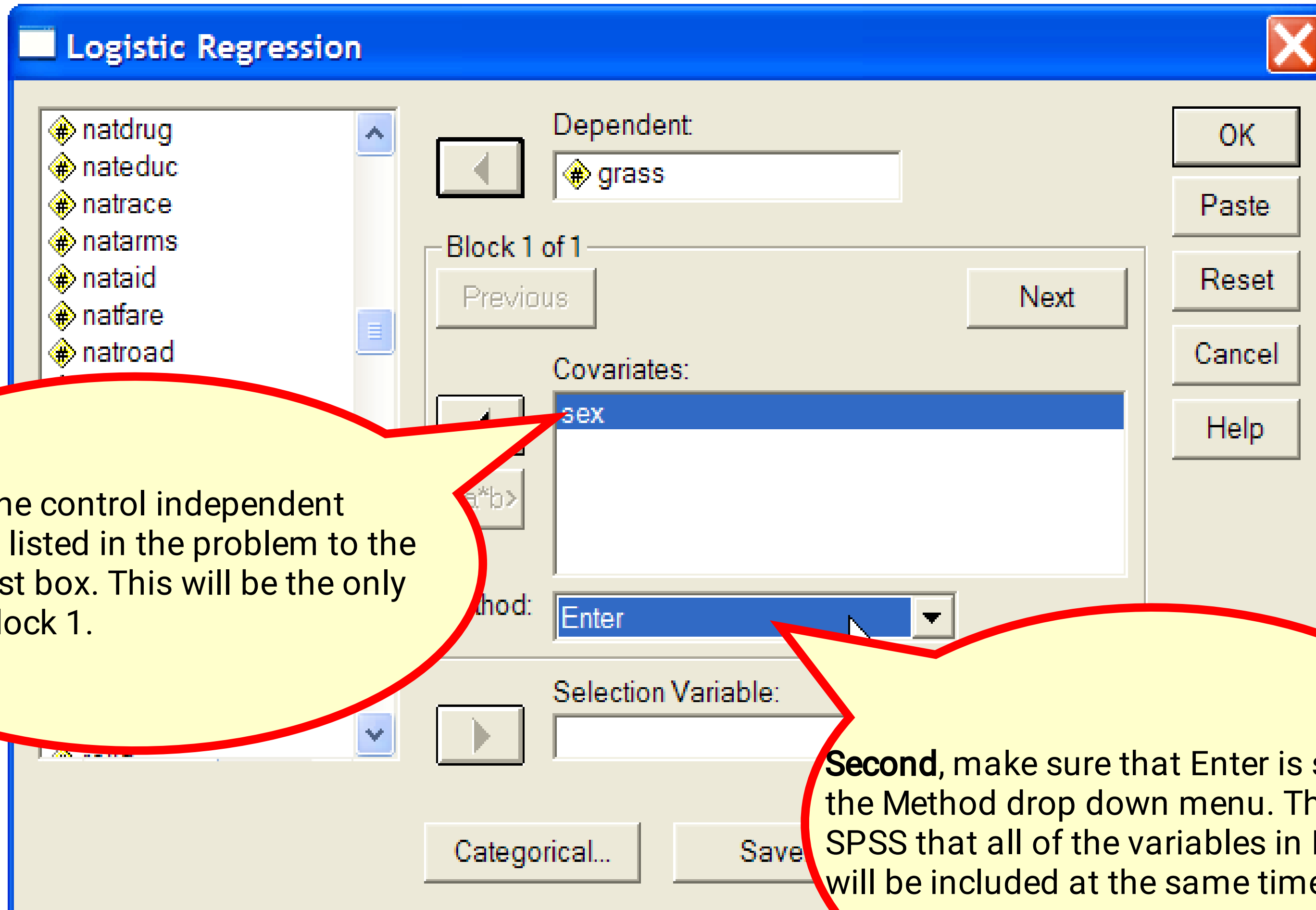
Strategy for Outliers and Influential Cases

- Our strategy for evaluating the impact of outliers and influential cases on our logistic regression model will parallel what we have done for multiple regression and discriminant analysis:
  - First, we run a baseline model including all cases
  - Second, we run a model excluding outliers (whose standardized residual is greater than 3.0 or less than 3.0) and influential cases (whose Cook's distance is greater than 1.0)
  - If the model excluding outliers and influential cases has a classification accuracy rate that is better than the baseline model, we will interpret the revised model. If the accuracy rate of the revised model without outliers and influential cases is less than 2% more accurate, we will interpret the baseline model

多变量logistic回归



多变量logistic回归



First, move the control independent variable, sex, listed in the problem to the *Covariates* list box. This will be the only variable in Block 1.

Second, make sure that Enter is selected in the Method drop down menu. This tells SPSS that all of the variables in Block 1 will be included at the same time.

软件缺陷预测：关键点

预处理

0: 数据预处理

1: 数据分布检查

2: Outlier识别

模型构建

3: 单变量分析

4: 多变量分析

模型评价

5: 模型验证

6: 性能评价

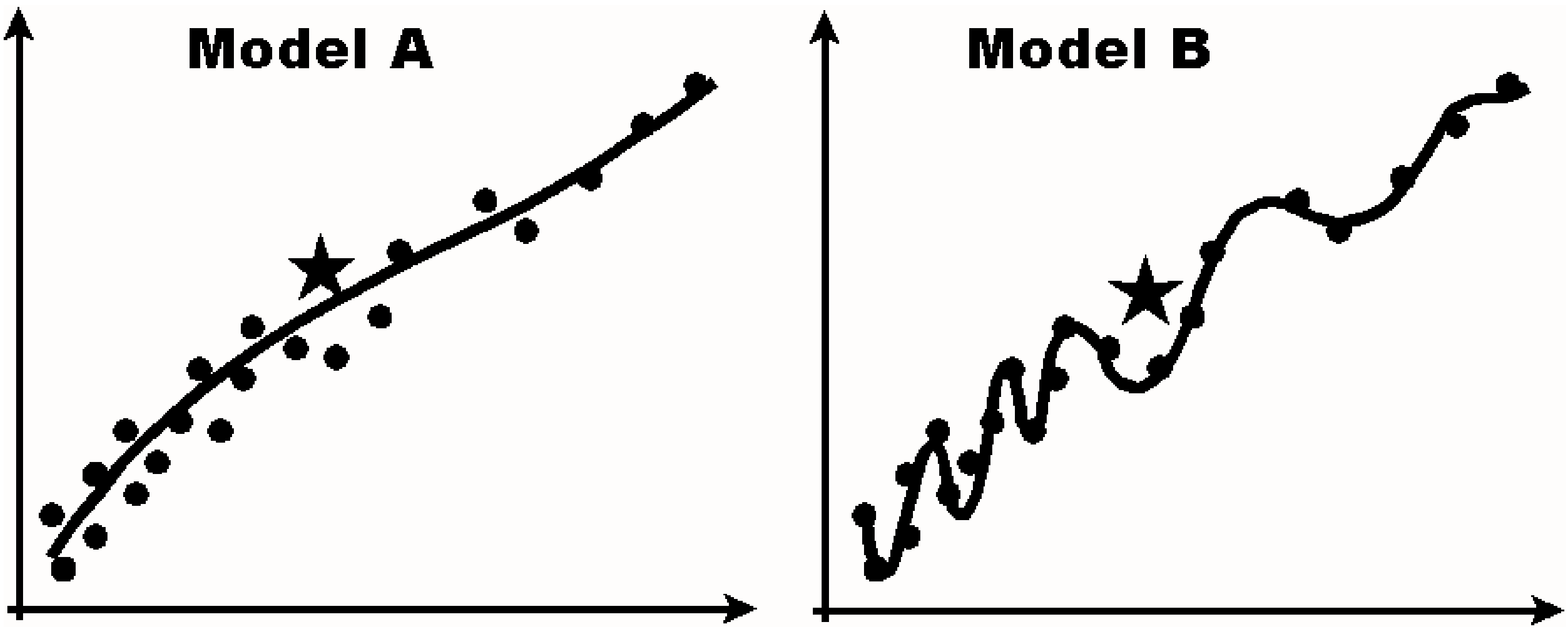
Split-sample

Leave-one-out cross-validation

K-fold cross-validation

模型评价

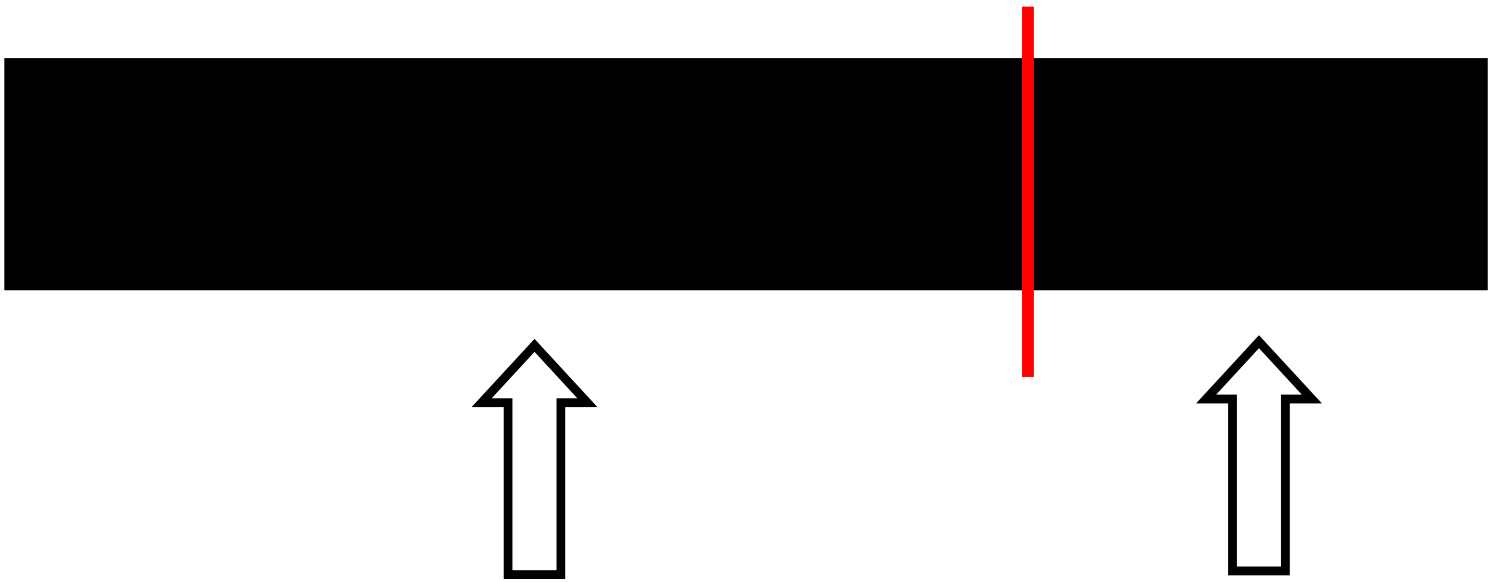
"Don't always prefer the model that describes the data best, but instead always prefer the model with the best predictive performance!"



Split-sample method

Main idea: predictive virtues can only be assessed for unseen data

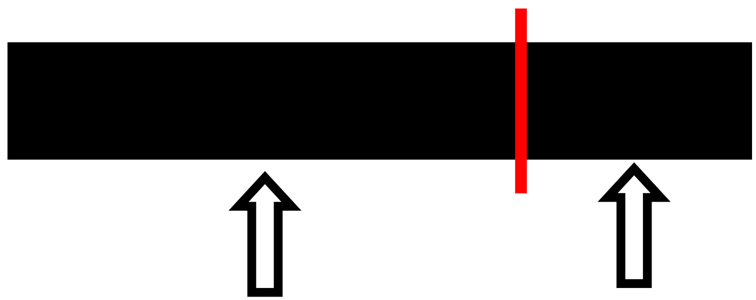
Data



Training set
Calibration set

Test set
Validation set

Split-sample method



Training set Test set

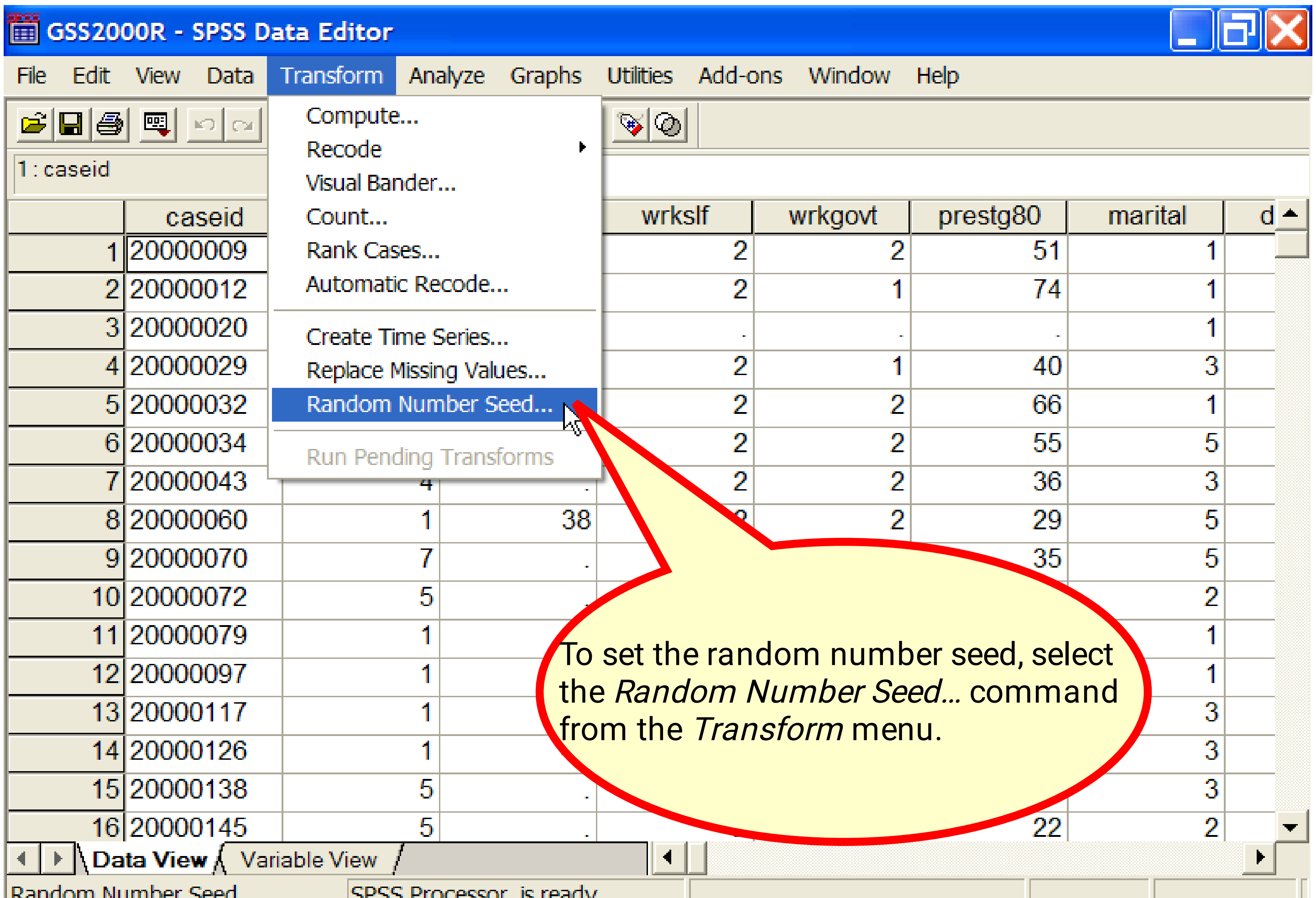
✓The training set and the test set do not change roles: there is no “crossing”

✓Only one part of the data is ever used for fitting

✓High variance

Split-sample method

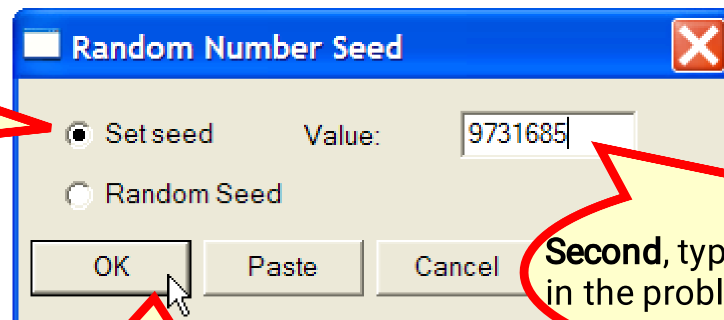
Example: set the random number seed



Split-sample method

Example: set the random number seed

First, click on the *Set seed to* option button to activate the text box.



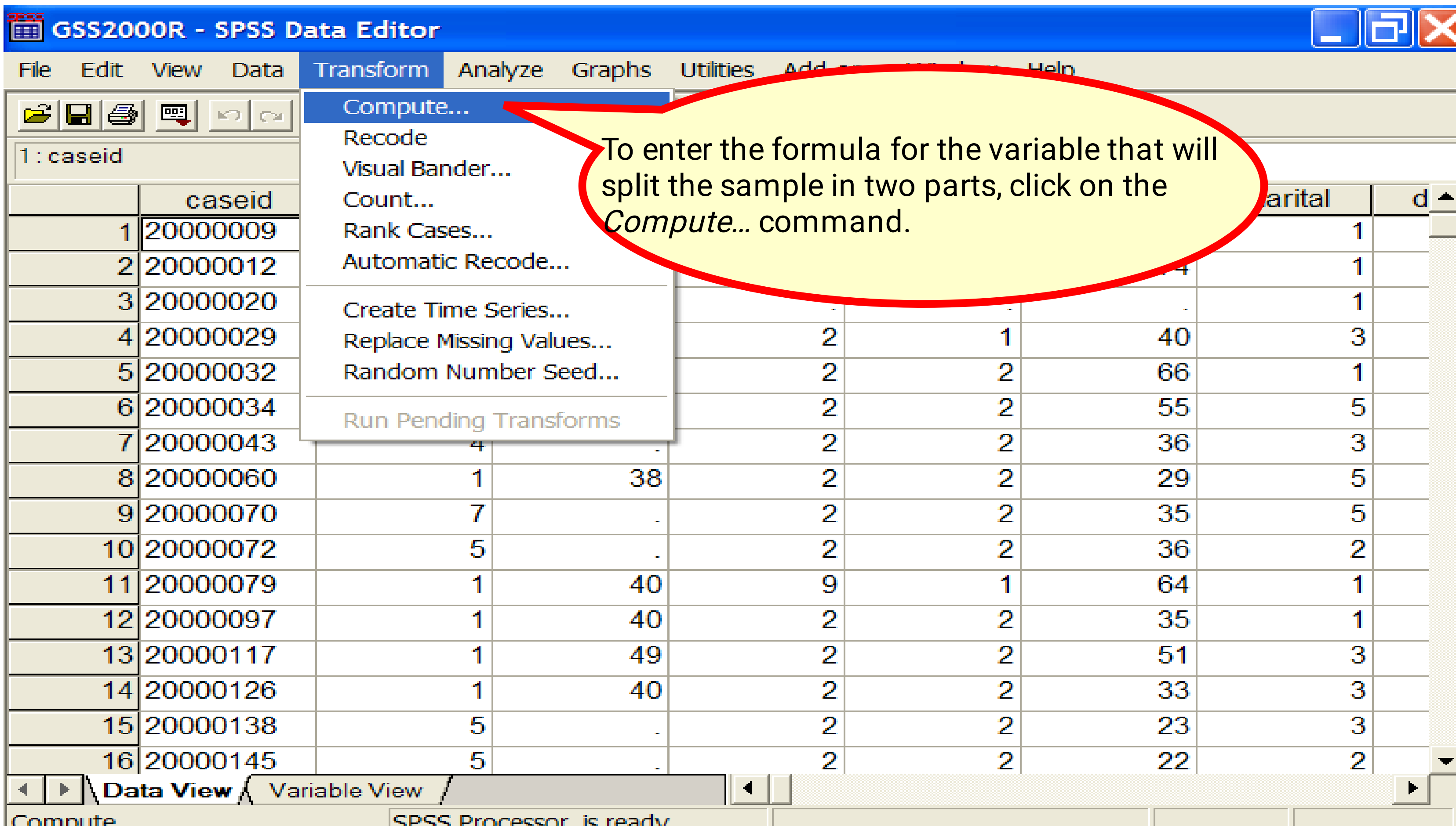
Second, type in the random seed stated in the problem.

Third, click on the OK button to complete the dialog box.

Note that SPSS does not provide you with any feedback about the change.

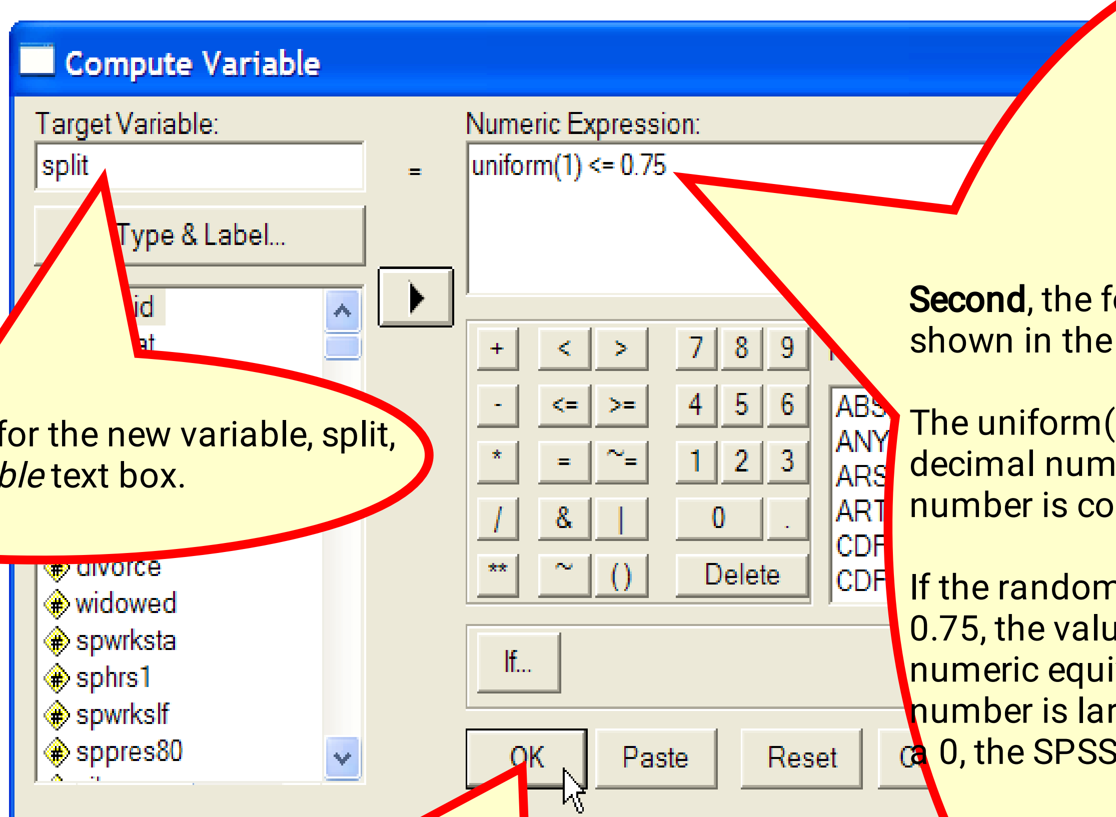
Split-sample method

Example: compute the split variable



Split-sample method

Example: The formula for the split variable



First, type the name for the new variable, split, into the *Target Variable* text box.

Second, the formula for the value of split is shown in the text box.

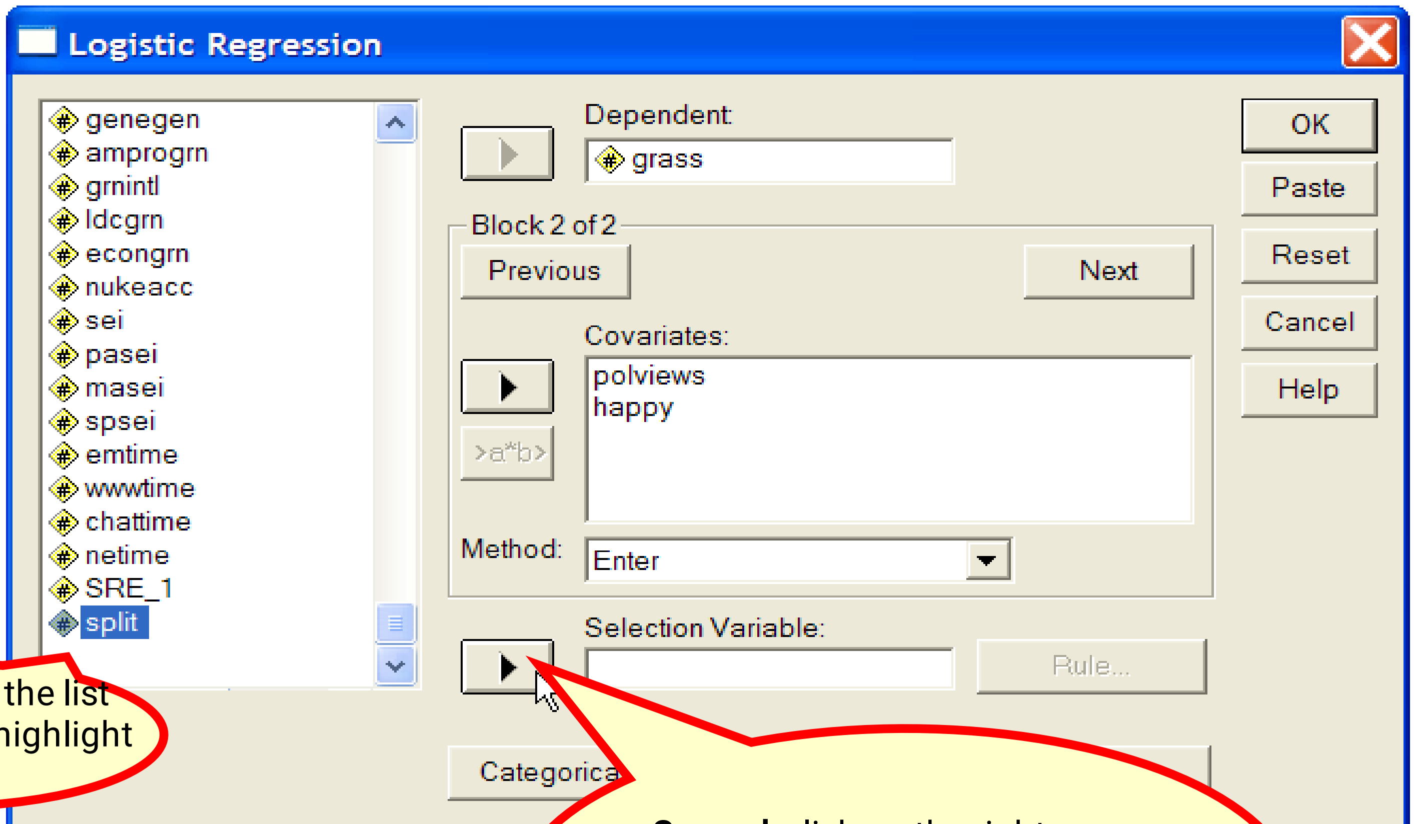
The uniform(1) function generates a random decimal number between 0 and 1. The random number is compared to the value 0. 75.

If the random number is less than or equal to 0.75, the value of the formula will be 1, the SPSS numeric equivalent to true. If the random number is larger than 0.75, the formula will return a 0, the SPSS numeric equivalent to false.

Third, click on the OK button to complete the dialog box.

Split-sample method

Example: Running the logistic regression with the training sample

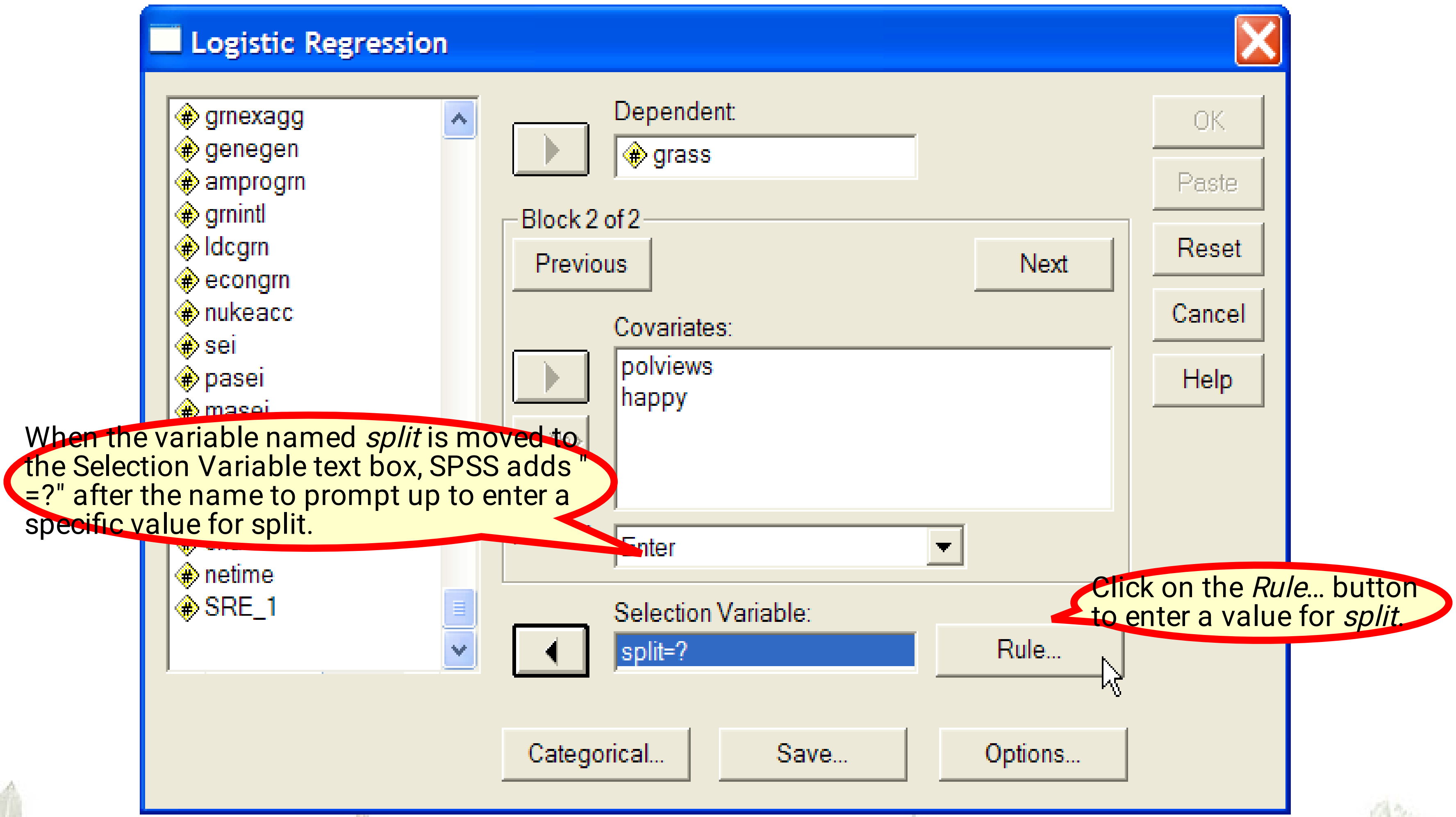


First, scroll down the list of variables and highlight the variable *split*.

Second, click on the right arrow button to move the split variable to the *Selection Variable* text box.

Split-sample method

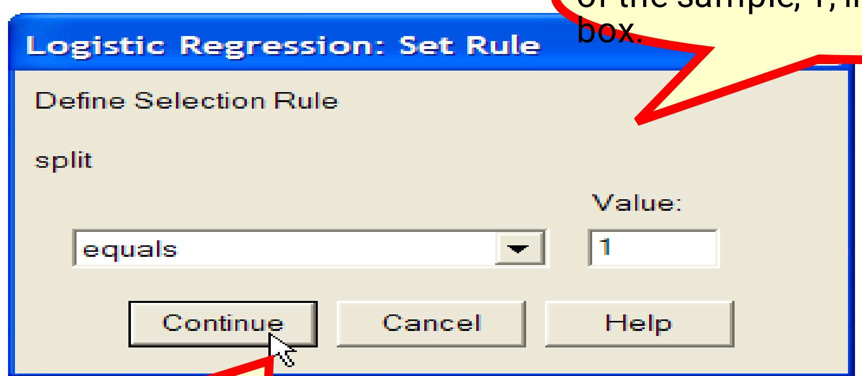
Example: Setting the value of split to select cases



Split-sample method

Completing the value selection

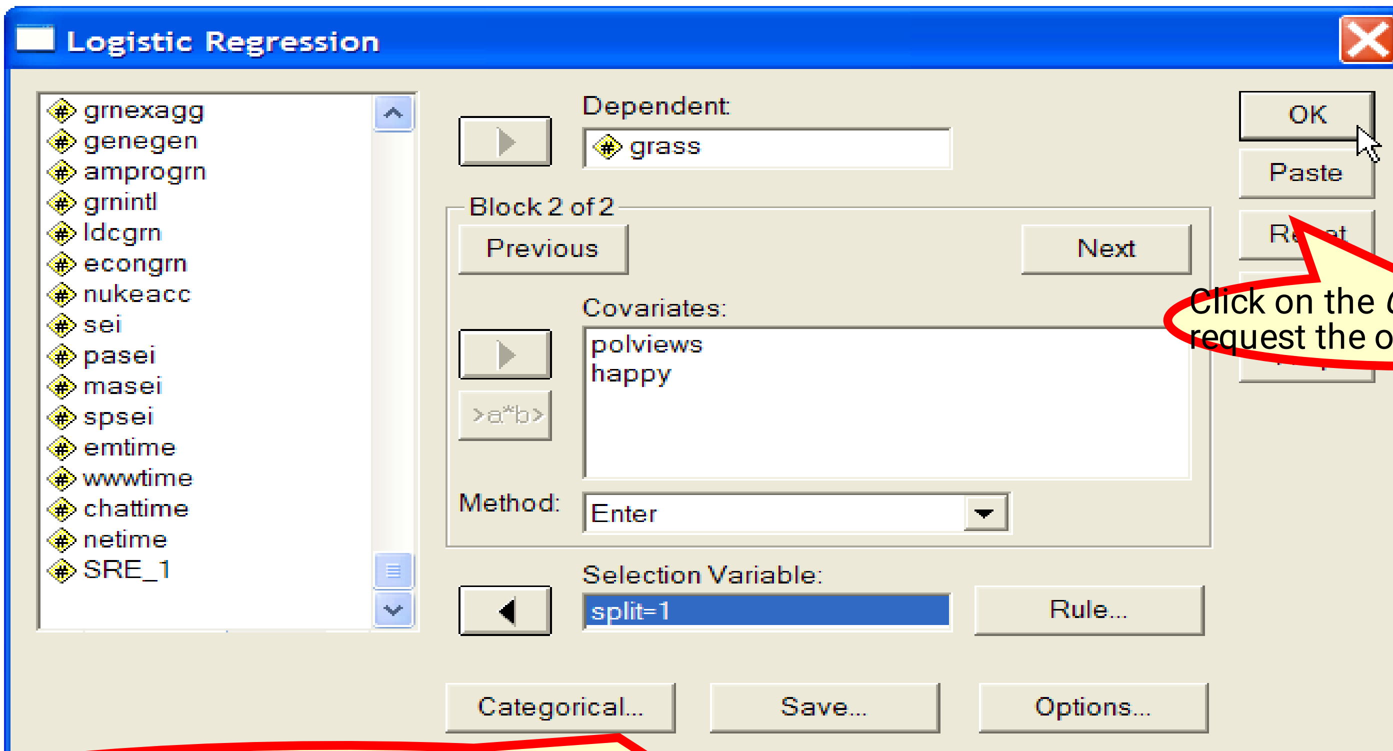
First, type the value for the first half of the sample, 1, into the *Value* text box.



Second, click on the *Continue* button to complete the value entry.

Split-sample method

Example: Requesting output



Click on the *OK* button to request the output.

When the value entry dialog box is closed, SPSS adds the value we entered after the equal sign. This specification now tells SPSS to include in the analysis only those cases that have a value of 1 for the split variable.

Split-sample method

Example: output

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	SEX	.820	.449	3.328	1	.068	2.270
	POLVIEWS	-.467	.162	8.285	1	.004	.627
	HAPPY	-1.771	.491	12.992	1	.000	.170
	Constant	3.750	1.409	7.081	1	.008	42.536

a. Variable(s) entered on step 1: POLVIEWS, HAPPY.

The relationship between "liberal or conservative political views" [polviews] and “support for legalization of marijuana" [grass] was statistically significant for the model using the full data set (p=0.008).

Similarly, the relationship in the cross-validation analysis was statistically significant. In the cross-validation analysis, the probability for the test of relationship between "liberal or conservative political views" [polviews and “support for legalization of marijuana" [grass] was p=0.004, which was less than or equal to the level of significance of 0.05 and statistically significant.

Split-sample method

Example: output

Classification Table[d]

Observed			Predicted: Selected Cases[a] SHOULD MARIJUANA BE MADE LEGAL: NOT LEGAL	Selected Cases[a] SHOULD MARIJUANA BE MADE LEGAL: LEGAL	Selected Cases[a] Percentage Correct	Unselected Cases[b,c] SHOULD MARIJUANA BE MADE LEGAL: NOT LEGAL	Unselected Cases[b,c] SHOULD MARIJUANA BE MADE LEGAL: LEGAL	Unselected Cases[b,c] Percentage Correct
Step 1	SHOULD MARIJUANA BE MADE LEGAL	NOT LEGAL	65	10	86.7	27	3	90.0
		LEGAL	29	14	32.6	12	3	20.0
	Overall Percentage				66.9			66.7

a. Selected cases SPLIT EQ 1

b. Unselected cases SPLIT NE 1

c. Some of the unselected cases are not classified due to either missing values in the independent variables or categorical variables with values out of the range of the selected cases.

d. The cut value is .500

The classification accuracy rate for the model using the training sample was 66.9%, compared to 66.7% for the validation sample. The shrinkage in classification accuracy for the validation analysis is the difference between the accuracy for the training sample (66.9%) and the accuracy for the validation sample (66.7%), which equals 0.2% in this analysis. The shrinkage was within the 2% criteria for minimal shrinkage, small enough to support a conclusion that the logistic regression model based on this analysis would be effective in predicting scores for cases other than those included in the calculation of the regression analysis.

Leave-one-out cross-validation

Data (n observations)

Test set = a single observation
Training set = all the rest

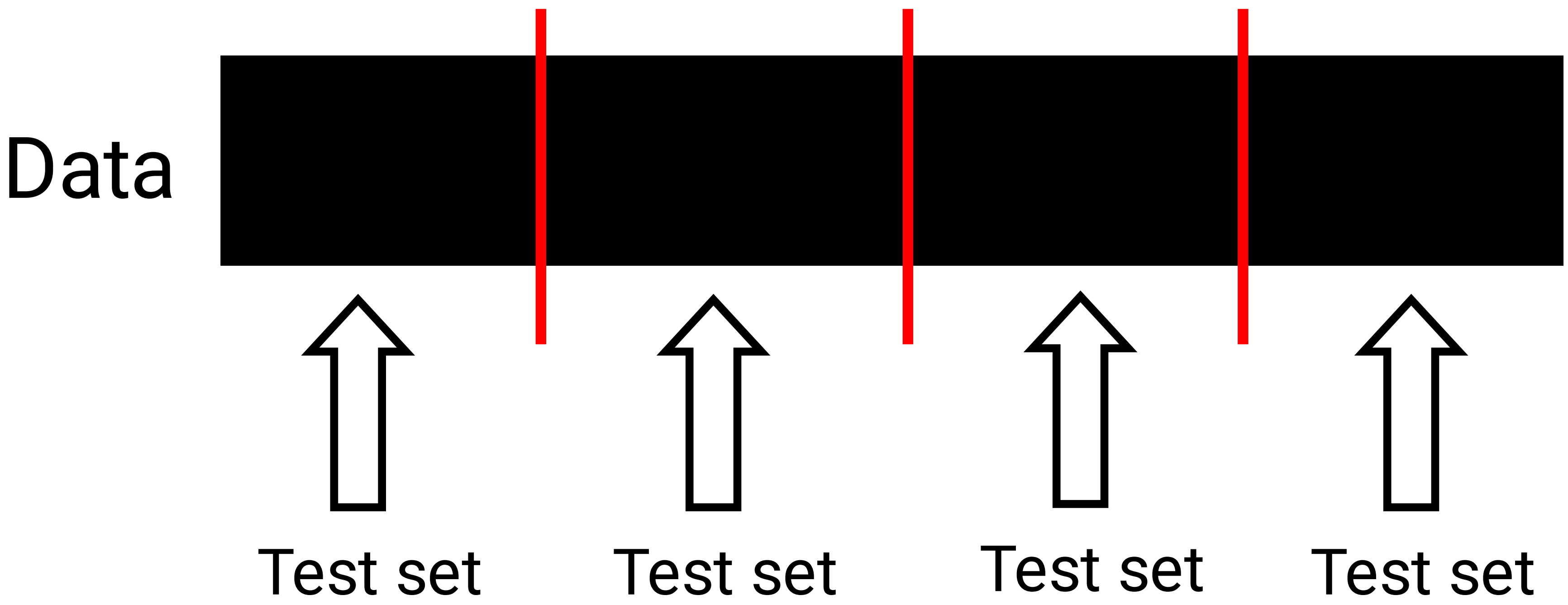
Prediction error is average performance on the n training sets.

Leave-one-out cross-validation

- All the data are used for fitting (but not at the same time, of course)
- Prediction is based on a large data set; This gives small prediction errors
- Problem: as n grows large, the method overfits (i.e., it does not converge on the correct model, in the case that there is one – that is, the method is not consistent)
- Sometimes, the method can have high variance

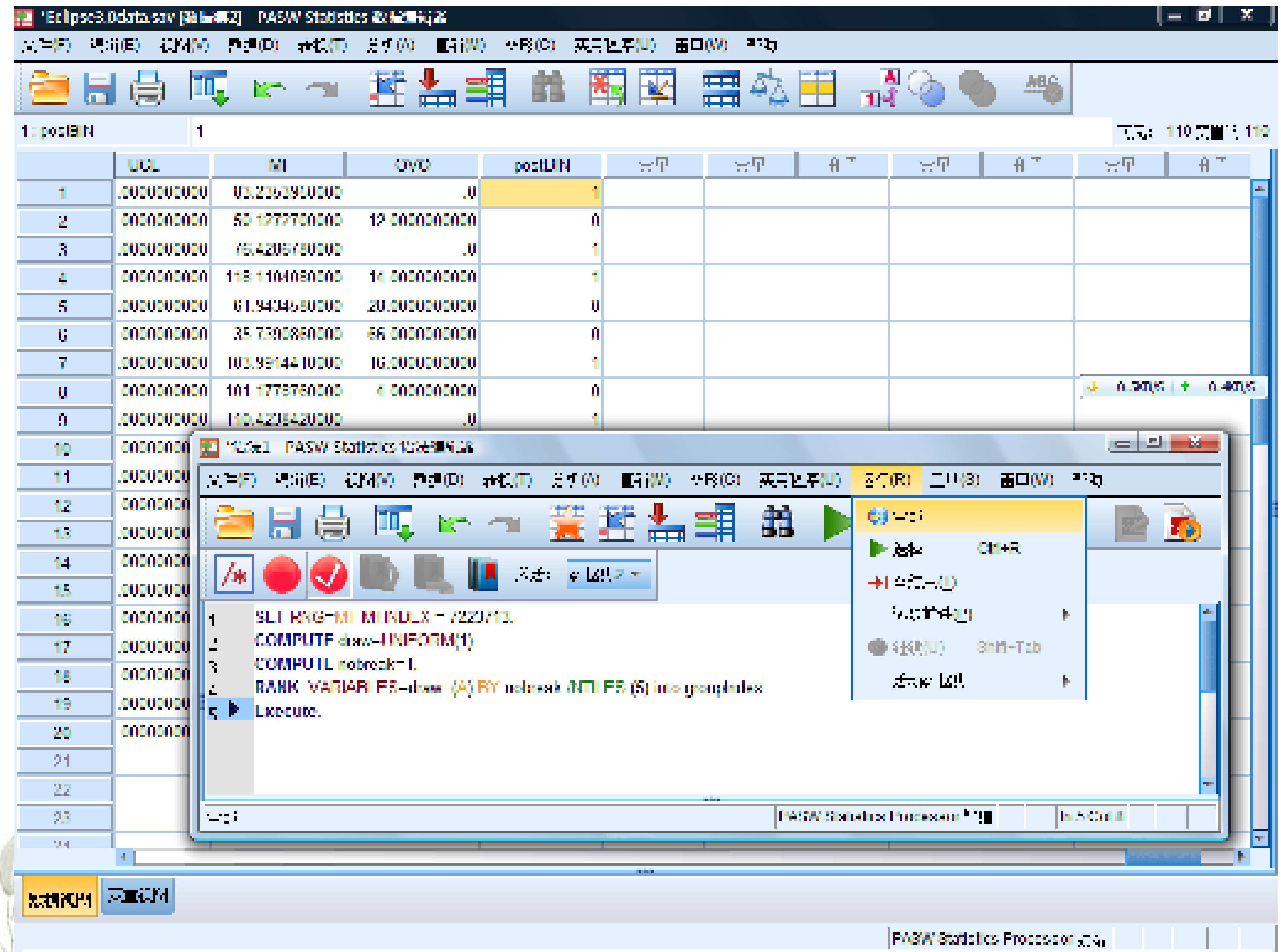
K-fold cross-validation

Successively setting apart a block of data.
(instead of a single observation)



K-fold cross-validation

Example: 5-fold cross-validation



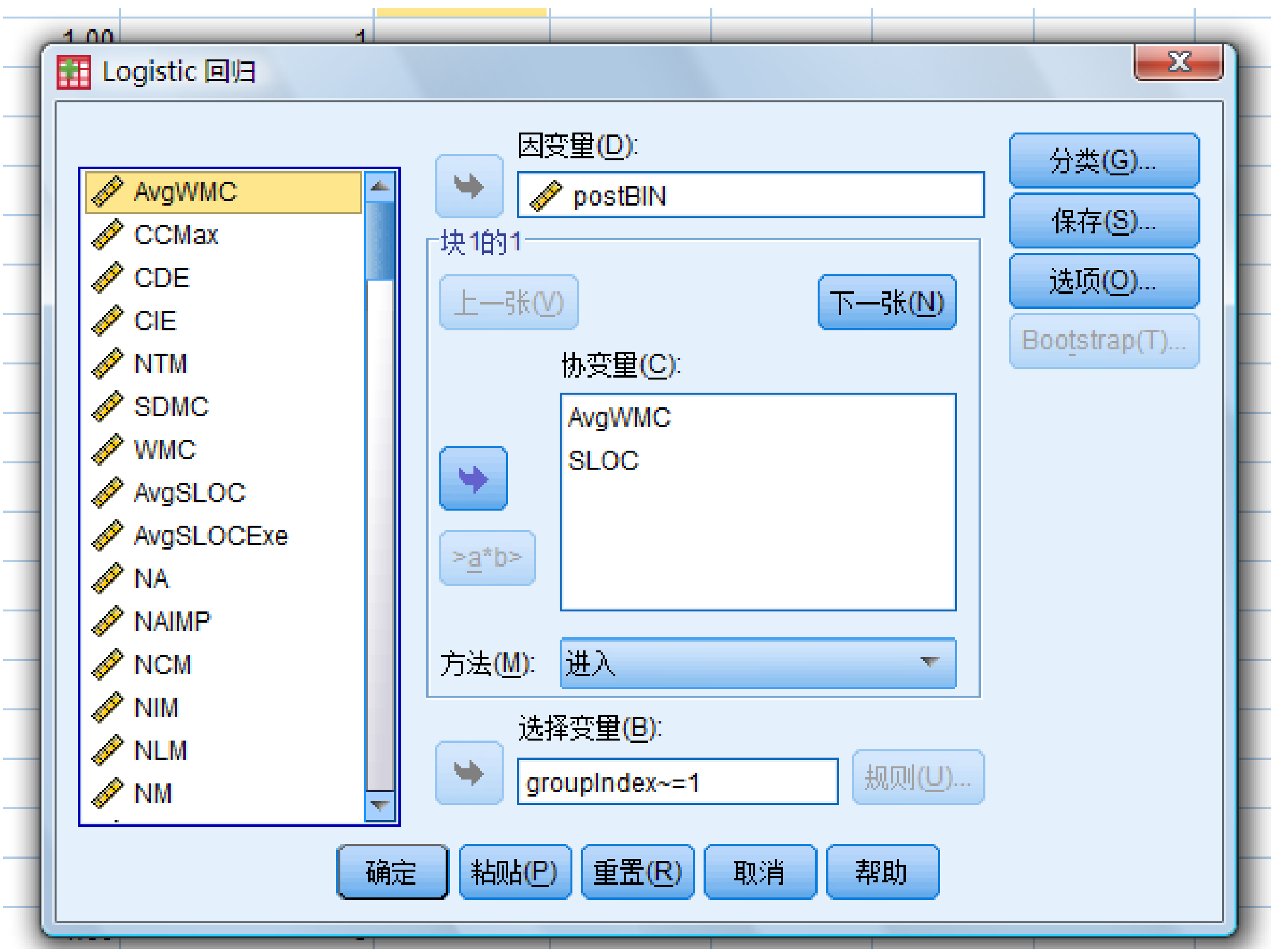
K-fold cross-validation

Example: 5-fold cross-validation

	UCL	MI	OVO	postBIN	draw	nobreak	groupIndex
1	.0000000000	83.2363950000	.0	1	.11	1.00	1
2	.0000000000	58.1272700000	12.0000000000	0	.72	1.00	4
3	.0000000000	76.4206780000	.0	1	.33	1.00	2
4	.0000000000	118.1104080000	14.0000000000	1	.22	1.00	2
5	.0000000000	61.9434680000	20.0000000000	0	.12	1.00	1
6	.0000000000	35.7390860000	66.0000000000	0	.69	1.00	3
7	.0000000000	103.9914410000	16.0000000000	1	.62	1.00	3
8	.0000000000	101.1778760000	4.0000000000	0	.50	1.00	3
9	.0000000000	110.4238420000	.0	1	.85	1.00	5
10	.0000000000	63.0803920000	8.0000000000	1	.67	1.00	4
11	.0000000000	141.5202680000	2.0000000000	1	.90	1.00	5
12	.0000000000	107.8798690000	0	1	.04	1.00	1
13	.0000000000	47.1204170000	6.0000000000	0	.63	1.00	4
14	.0000000000	63.9554480000	76.0000000000	0	.22	1.00	2
15	.0000000000	64.3014300000	6.0000000000	1	.05	1.00	1
16	.0000000000	77.3552740000	0	0	.78	1.00	5
17	.0000000000	91.1478020000	4.0000000000	1	.63	1.00	4
18	.0000000000	138.7651080000	6.0000000000	0	.92	1.00	5
19	.0000000000	95.5401130000	4.0000000000	1	.30	1.00	2
20	.0000000000	112.3486780000	0	0	.42	1.00	3

K-fold cross-validation

Example: 5-fold cross-validation



K-fold cross-validation

Example: 5-fold cross-validation

案例处理汇总

未加权的案例[a]		N	百分比
选定案例	包括在分析中	16	80.0
	缺失案例	0	.0
	总计	16	80.0
未选定的案例		4	20.0
总计		20	100.0

a. 如果权重有效，请参见分类表以获得案例总数。

分类表[c]

已观测			已预测					
			选定案例[a]			未选定的案例[b]		
			postBIN		百分比校正	postBIN		百分比校正
			0	1		0	1	
步骤 1	postBIN	0	4	4	50.0	0	1	.0
		1	3	5	62.5	1	2	66.7
	总计百分比				56.3			50.0

a. 已选定的案例 groupIndex NE 1

b. 未选定的案例 groupIndex EQ 1

c. 切割值为 .500

软件缺陷预测：关键点

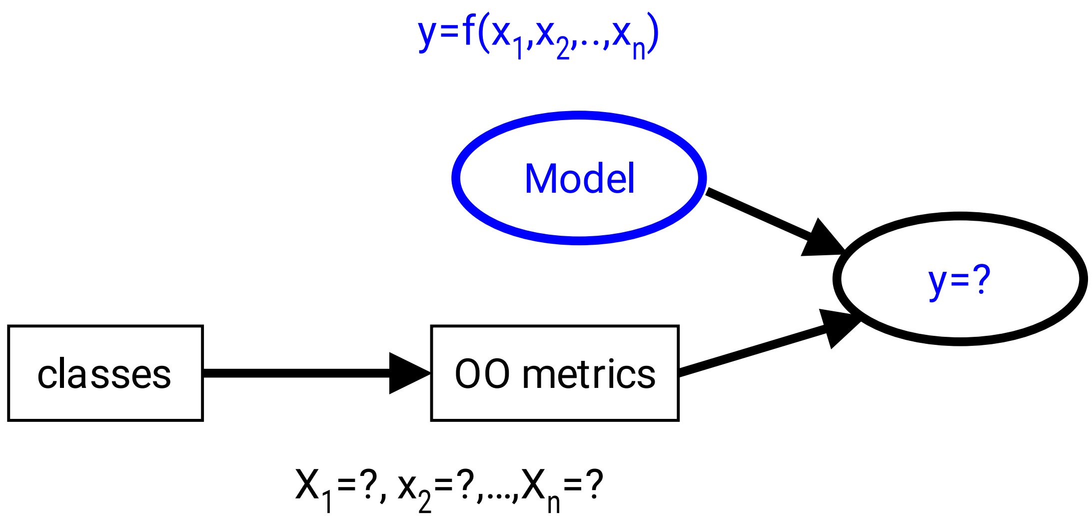
预处理
- 0: 数据预处理
- 1: 数据分布检查
- 2: Outlier识别

模型构建
- 3: 单变量分析
- 4: 多变量分析

模型评价
- 5: 模型验证
- 6: 性能评价

分类性能
排序性能
假设检验

Quantitative prediction



For each case in the data set: Magnitude of Relative Error (MRE)

$$MRE_i = \frac{|y_i - \hat{y}_i|}{y_i}$$

Quantitative prediction

For the model, compare actual and estimated quantity for n cases in the dataset:

Mean Magnitude of Relative Error (MMRE)

$$MMRE = \frac{1}{n}\sum_{i=1}^{n} MRE_i$$

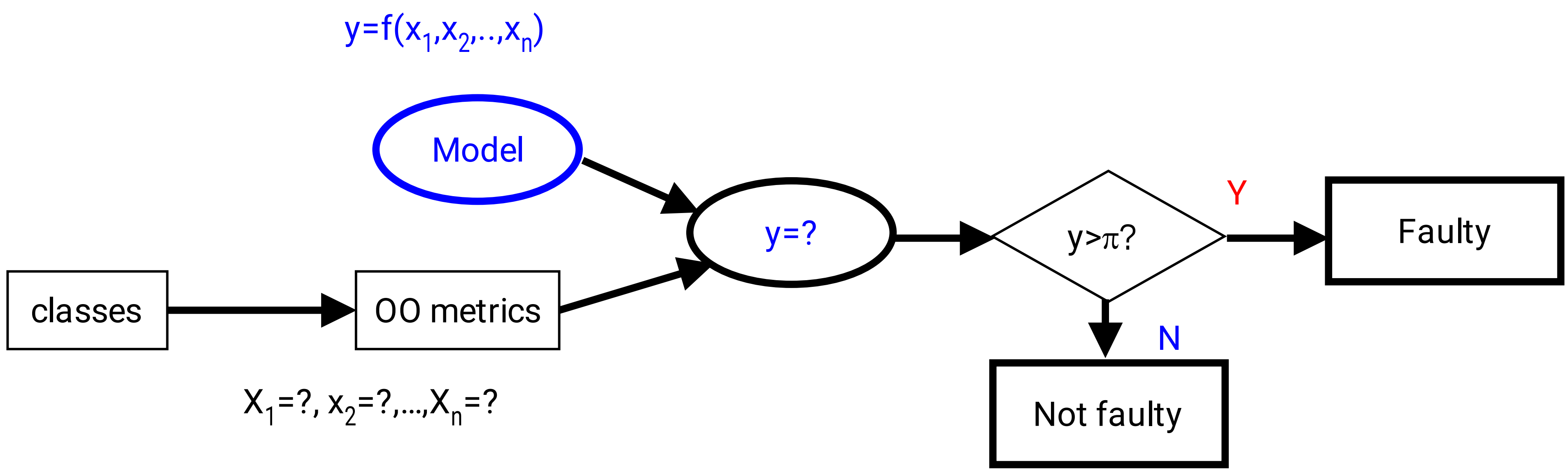
Prediction level

$$\mathrm{Pred}(q) = \frac{k}{n}$$

where k = the number cases in a set of n cases whose MRE <= *q*

good: $\mathrm{Pred}(0.25) \geq 0.75$

Classification prediction



Actual	Predicted	
	Fault($y>\pi$)	No fault($y\leq\pi$)
fault	**a**	**c**
No fault	**b**	**d**

a: # True Positives (TP)

b: # False Positives (FP)

c: # False Negatives (FN)

d: # True Negatives (TN)

Classification prediction

Actual	Predicted	
	Fault($y>\pi$)	No fault($y\leq\pi$)
fault	**a**	**c**
No fault	**b**	**d**

Disadvantage: depend on π

Sensitivity = a/(a+c) = TP/(TP+FN) = Recall

Specificity = d/(b+d) = TN/(FP+TN)

Precision = a/(a+b) = TP/(TP+FP)

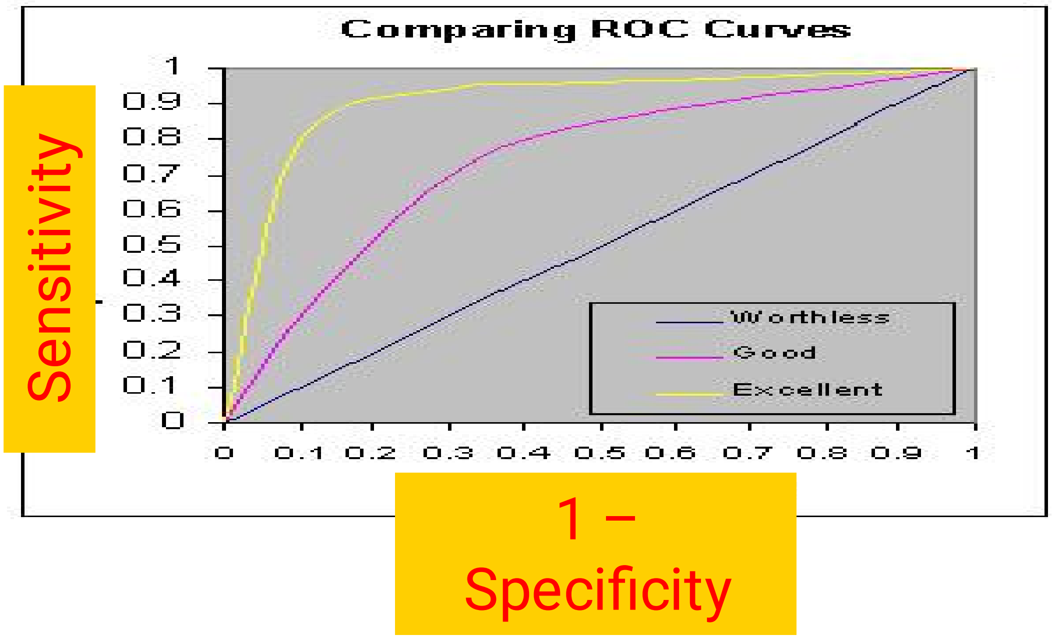
Accuracy = (a+d)/(a+b+c+d)

= (TP+TN)/(TP+FP+FN+TN)

F-measure = 2 * Recall * Precision / (Recall + Precision)

Classification prediction

AUC (area under ROC, Receiver Operating Characteristic curves)



poor:	[0.5, 0.7)
moderate:	[0.7, 0.9)
very good:	[0.9, 1.0]

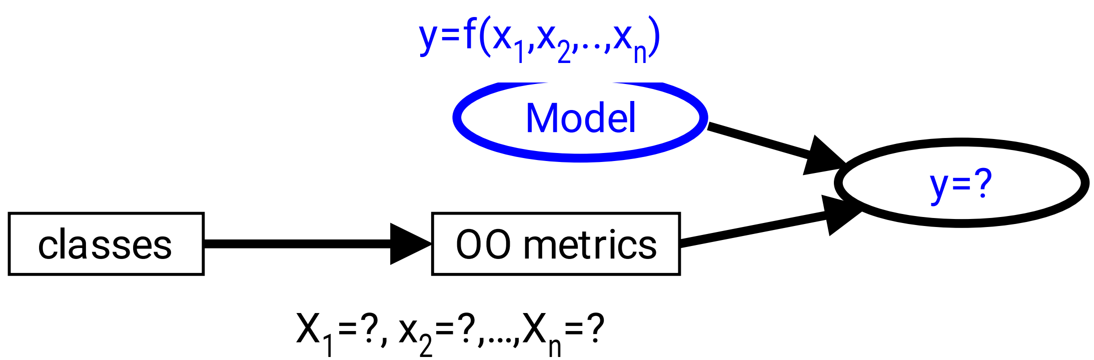
Advantages:

① Does not depend on the threshold π

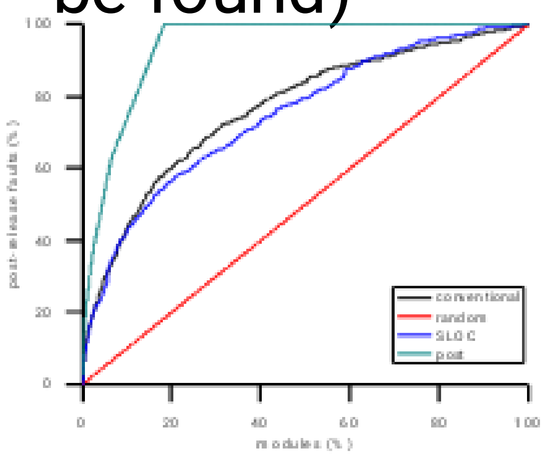
② Does not depend on the prior probabilities of positive and negative cases

③ can be interpreted as the probability that a randomly chosen positive observation is (correctly) rated or ranked with greater suspicion than a randomly chosen negative observation

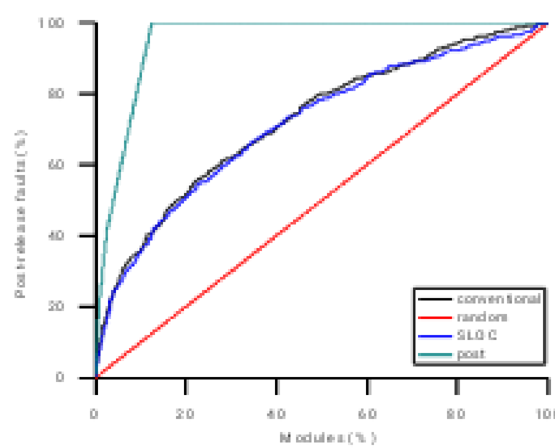
Rank prediction



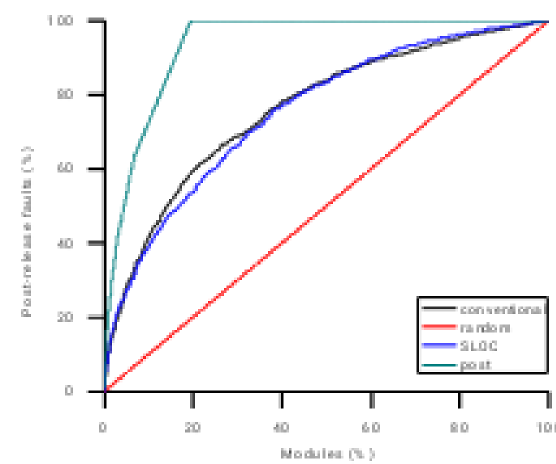
Alberg diagram: x is modules%, y is faults % (if top x% modules are selected to be tested/inspected, y% faults will be found)



Eclipse 2.0

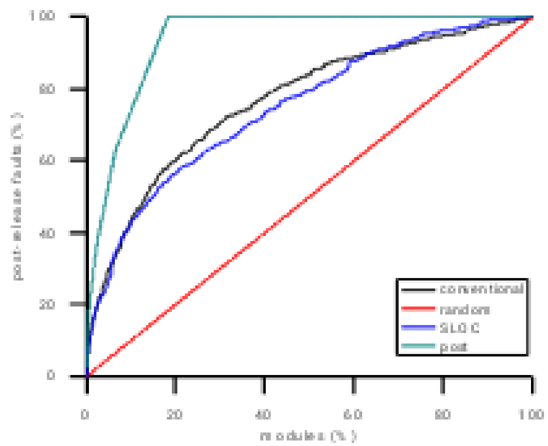


Eclipse 2.1



Eclipse 3.0

Rank prediction



Eclipse 2.0

Cost effectiveness:

$$CE_{\pi}(\text{model}) = \frac{Area_{\pi}(\text{model}) - Area_{\pi}(\text{Random})}{Area_{\pi}(\text{optimal}) - Area_{\pi}(\text{Random})}$$

Hypothesis testing

Given model A and model B, the problem is:

Model A is significantly better than model B?

If model A and model B are validated on the same data sets, we use:

- ✓ paired t-test
- ✓ Wilcoxon Signed-Rank Test (paired)

Hypothesis testing

Paired Sample t-test

$H_0: \mu_d = \mu_0$
$H_1: \mu_d \neq \mu_0$ (two-tailed).
μ_d: mean of population differences.
α: significant level (e.g., 0.05).
Test Statistic:

$$T_d = \frac{\bar{d} - \mu_d}{S_d/\sqrt{n}}, \quad t_d = \frac{\bar{d} - \mu_0}{S_d/\sqrt{n}}$$

$\bar{d}$: average of sample differences.

S_d: standard deviation of sample difference

n: number of pairs.

- Reject H_0 if $|t_d| > t_{\alpha/2, n-1}$.
- Power $= 1 - \beta$.
- $(1-\alpha)100\%$ Confidence Interval for μ_d:
  $\bar{d} - t_{\alpha/2} S/\sqrt{n} \leq \mu_d < \bar{d} + t_{\alpha/2} S/\sqrt{n}$
- p-$value = P_{H_0}(|\mathrm{T}| > t_d)$, $\mathrm{T} \sim t_{n-1}$.

Hypothesis testing

Wilcoxon Signed-Rank Test (paired)

- Null hypothesis: the population median from which both samples were drawn is the same.
- The sum of the ranks for the "positive" (up-regulated) values is calculated and compared against a precomputed table to a p-value.
  - Sorting the absolute values of the differences from smallest to largest.
  - Assigning ranks to the absolute values.
  - Find the sum of the ranks of the positive differences.
- If the null hypothesis is true, the sum of the ranks of the positive differences should be about the same as the sum of the ranks of the negative differences

Pair	Before	After	Diff.	Rank
1	89	73	16	15.5
2	83	77	6	7
3	80	58	22	17
4	72	77	−5	5
5	77	70	7	8
6	74	62	12	13.5
7	69	67	2	2
8	65	68	−3	3
9	60	44	16	15.5
10	55	50	5	5
11	54	46	8	9.5
12	50	38	12	13.5
13	42	47	−5	5
14	48	40	8	9.5
15	44	43	1	1
16	38	29	9	11
17	36	25	11	12

The Wilcoxon signed-rank Test:

$H_0 : \mu_1 = \mu_2$

$H_1 : \mu_1 \neq \mu_2$

$T = \min\{\sum_+ \text{Rank}, \sum_- \text{Rank}\}$

At $\alpha = 0.01$, two-tailed test,

reject H_0 if $T \neq 23$ when $N = 17$.

(Table)

(The zero difference is ignored when assigning ranks. $N_{new} = N_{old} - \#\{ties\}$)

$T = \min\{\sum_+ \text{Rank} = 140, \sum_- \text{Rank} = 13\}$

$= 13$

The obtained T=13 is less than the critical value 23, so we reject H_0.

```
p =

    0.0026

h =

     1

stats =

          zval: -3.0089
    signedrank: 13
```

Hypothesis testing

Assumptions of paired t-test

- For paired t-test, it is the distribution of the subtracted data that must be normal

Assumptions of Wilcoxon signed-rank test

- Do not assume that the data is normally distributed.
- Non-parametric methods are robust to outliers and noisy data

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 32, NO. 10, OCTOBER 2006 771

Empirical Analysis of Object-Oriented Design Metrics for Predicting High and Low Severity Faults

Yuming Zhou and Hareton Leung, *Member*, *IEEE Computer Society*

An In-Depth Study of the Potentially Confounding Effect of Class Size in Fault Prediction

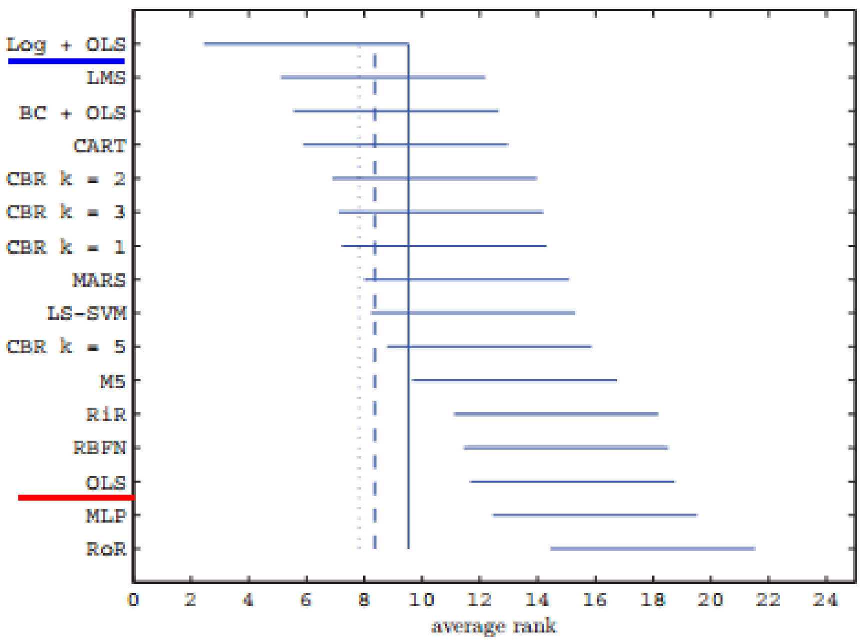
YUMING ZHOU and BAOWEN XU, Nanjing University
HARETON LEUNG, Hong Kong Polytechnic University
LIN CHEN, Nanjing University

开发工作量估算

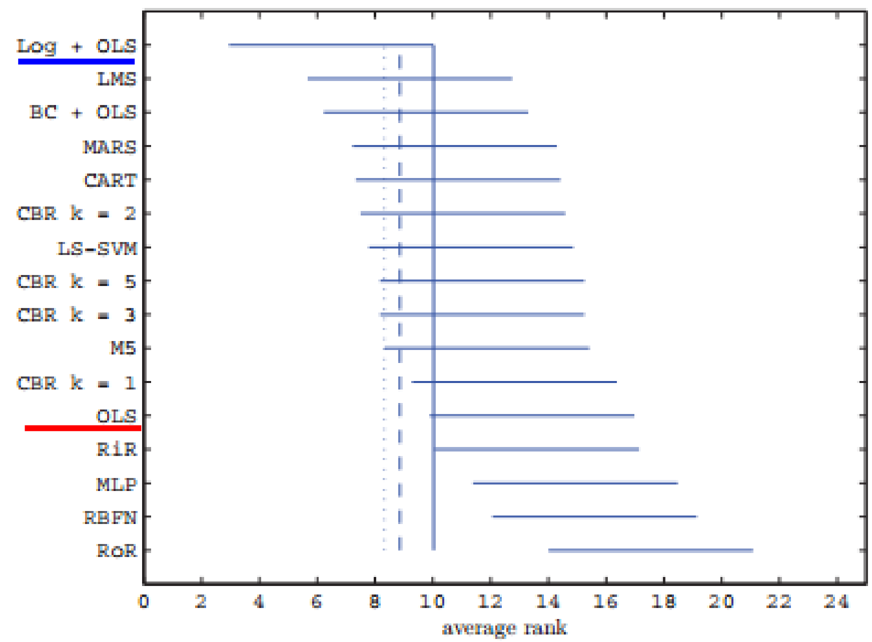
IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 38, NO. 2, MARCH/APRIL 2012 375

Data Mining Techniques for Software Effort Estimation: A Comparative Study

Karel Dejaeger, Wouter Verbeke, David Martens, and Bart Baesens



(a) plot of the Bonferroni-Dunn test for MdMRE



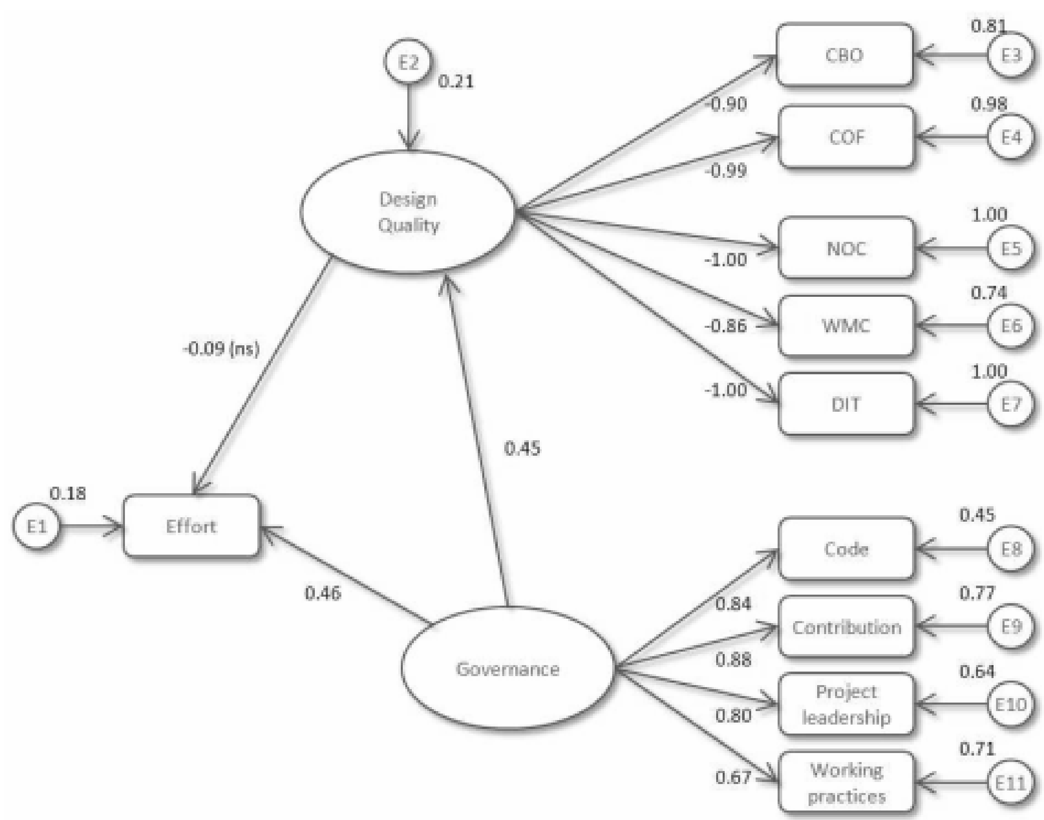
(b) plot of the Bonferroni-Dunn test for $Pred_{25}$

设计质量、开发工作量和管理的关系

IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 34, NO. 6, NOVEMBER/DECEMBER 2008 765

An Empirical Study on the Relationship among Software Design Quality, Development Effort, and Governance in Open Source Projects

Eugenio Capra, Chiara Francalanci, and Francesco Merlo



设计模式与代码缺陷的关系

904 IEEE TRANSACTIONS ON SOFTWARE ENGINEERING, VOL. 30, NO. 12, DECEMBER 2004

Defect Frequency and Design Patterns: An Empirical Study of Industrial Code

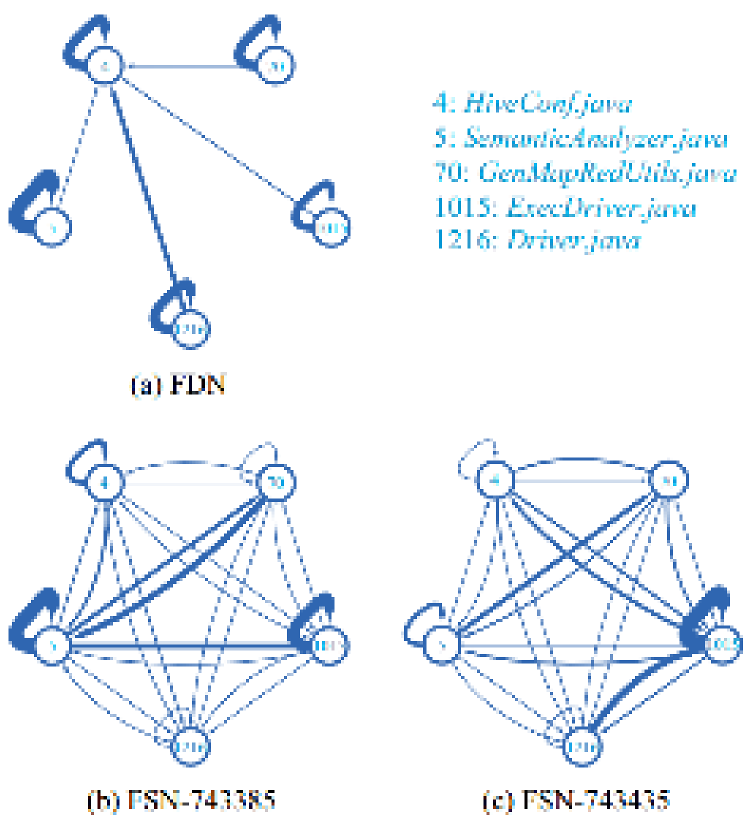
Marek Vokáč

Coefficient		P	Odds Ratio	95% CI Lower	95% CI Upper
Constant	β_0	0.000			
Week	β_W	0.000	**0.99**	0.99	0.99
Size (KLOC)	β_K	0.000	**1.69**	1.53	1.87
Factory	β_F	0.000	**0.63**	0.51	0.77
Singleton	β_S	0.141	1.35	0.91	2.02
Observer	β_O	0.000	**1.55**	1.26	1.91
Template Method	β_T	0.048	**0.72**	0.52	1.00
Decorator	β_D	0.154	0.49	0.18	1.31
Singleton + Observer	β_{SO}	0.000	**0.32**	0.21	0.48
Singleton × Size	β_{SK}	0.000	**13.18**	6.29	27.61
Observer × Size	β_{OK}	0.009	**1.21**	1.05	1.40

焦点切换模式与调用图的关系

Focus-Shifting Patterns of OSS Developers and Their Congruence with Call Graphs

Qi Xuan*†, Aaron Okano*, Premkumar T Devanbu*, Vladimir Filkov*
{qxuan@, adokano@, devanbu@cs., filkov@cs.}ucdavis.edu
*Department of Computer Science, University of California, Davis, CA 95616, USA
†Department of Automation, Zhejiang University of Technology, Hangzhou 310023, China



(a) FDN

(b) FSN-743385 (c) FSN-743435

开发者的代码熟悉程度建模

Degree-of-Knowledge: Modeling a Developer's Knowledge of Code

THOMAS FRITZ, University of Zurich
GAIL C. MURPHY, University of British Columbia
EMERSON MURPHY-HILL, North Carolina State University
JINGWEN OU, University of British Columbia
EMILY HILL, Montclair State University

5.3. Degree-of-Knowledge

We combine the DOA and DOI of a source-code element for a developer and over a period in time to provide an indicator of the developer's familiarity in that element. We use a linear combination as an initial starting point:

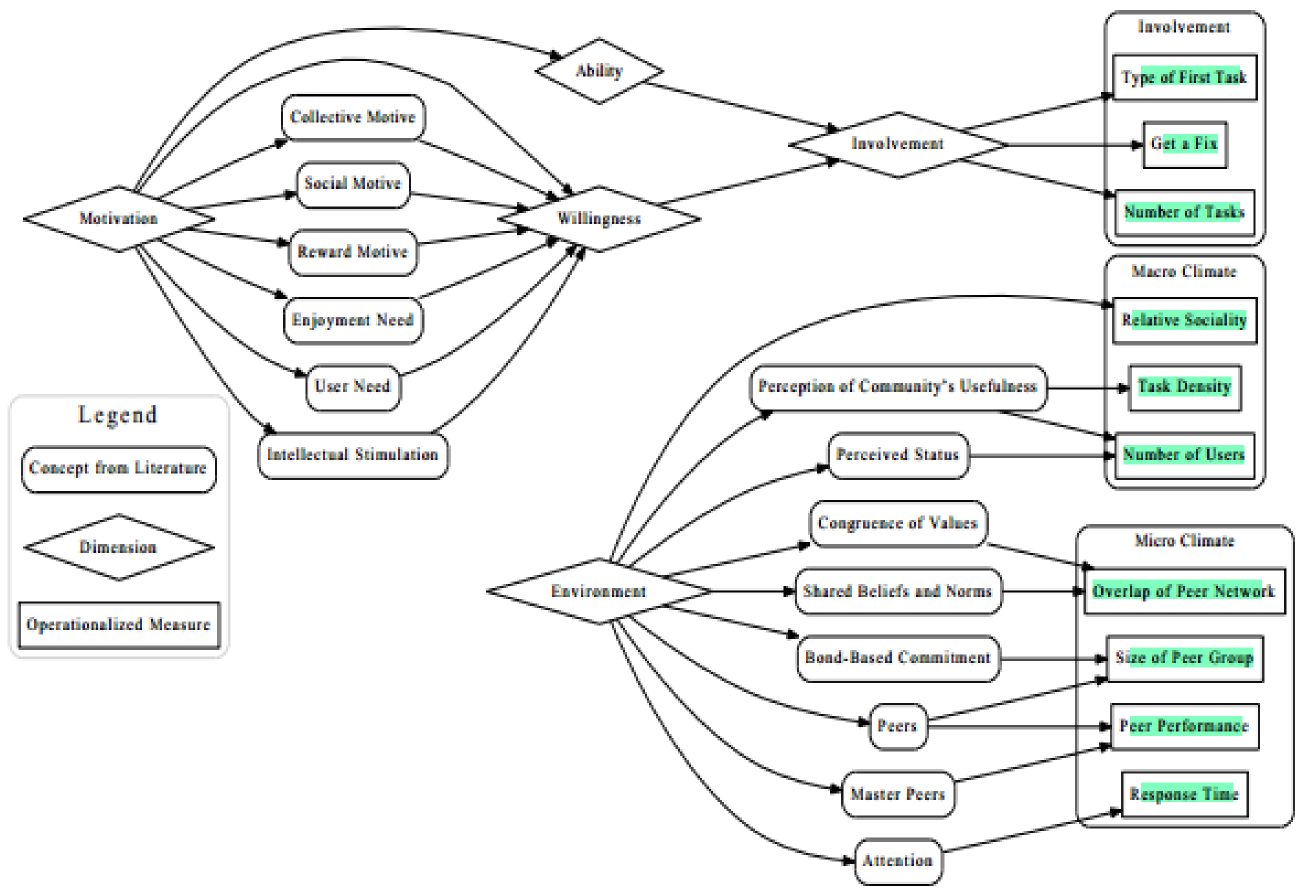
$$DOK = \alpha_{FA} * FA + \alpha_{DL} * DL + \alpha_{AC} * AC + \beta_{DOI} * DOI.$$

We discuss further this choice of using a linear combination later (Section 10).

谁将留在开源社区中？

Who Will Stay in the FLOSS Community? Modeling Participant's Initial Behavior

Minghui Zhou, *Member, ACM,* and Audris Mockus, *Member, IEEE*



Summary

Pre-processing

- 0: Data pre-processing
- 1: Data distribution exam.
  - Location
  - Dispersion
  - Skewness + Kurtosis
  - Box plot
- 2: Outlier identification
  - Univariate: standardized score
  - Multivariate: Mahalanobis D^2
  - Influential: Cook D

Model building

- 3: Univariate analysis
- 4: Multivariate analysis
  - Regression equation
  - Assumptions
  - Solution to violation

Model evaluation

- 5: Validation
  - Split-sample
  - K-fold cross-validation
- 6: Performance metrics
  - Classification
  - Ranking
  - Hypothesis testing

Thanks for your time and attention!