Chapter 316

Two-Stage Least Squares

Introduction

This procedure calculates the two-stage least squares (2SLS) estimate. This method is used fit models that include *instrumental variables*. 2SLS includes four types of variable(s): *dependent*, *exogenous*, *endogenous*, and *instrument*. These are defined as follows:

Dependent Variable	This is the response (or Y) variable that is to be regressed on the exogenous and endogenous (but not the instrument) variables.
Exogenous Variables	These independent (X_{ex}) variables are included in both the first and second stage regression models. They are not correlated with the random error values in the second stage regression.
Endogenous Variables	Each endogenous (or V_{en}) variable becomes the dependent variable in the first stage regression equation. Each is regressed on all exogenous and instrument variables. The predicted values from these regressions replace the original values of the endogenous variables in the second stage regression model.
Instrument Variables	Each endogenous variable becomes the dependent variable in the first stage regression equation. Each is regressed on all exogenous and instrument (X_{iv}) variables. The predicted values from these regressions replace the original values of the endogenous variables in the second stage regression model.

2SLS is used in econometrics, statistics, and epidemiology to provide consistent estimates of a regression equation when controlled experiments are not possible. They are discussed in every modern econometrics text. We have used Kmenta (2011) for the outline and example to follow.

Two-Stage Least Squares

Technical Details

The 2SLS model is comprised of the following two linear regression models.

$$\boldsymbol{y} = \mathbf{X_{ex}}\boldsymbol{\beta_{ex}} + \mathbf{V_{en}}\boldsymbol{\beta_{en}} + \boldsymbol{e} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{e}$$

$$\mathbf{V_{en}} = \mathbf{X_{ex}}\boldsymbol{\Gamma_{ex}} + \mathbf{X_{iv}}\boldsymbol{\Gamma_{iv}} + \mathbf{E} = \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E}$$

where

n: sample size

$\boldsymbol{y}$: $n \times 1$ vector dependent variable

$\mathbf{X_{ex}}$: $n \times k_{\text{ex}}$ matrix of exogenous regressor variables

$\mathbf{X_{iv}}$: $n \times k_{\text{iv}}$ matrix of instrumental variables

$\mathbf{V_{en}}$: $n \times k_{\text{en}}$ matrix of endogenous regressor variables

$\boldsymbol{\beta_{en}}$: $k_{\text{en}} \times 1$ vector of endogenous regressor parameters

$\boldsymbol{\beta_{ex}}$: $k_{\text{ex}} \times 1$ vector of included exogenous parameters

$\boldsymbol{\beta} = \begin{bmatrix}\boldsymbol{\beta_{ex}} \\ \boldsymbol{\beta_{en}}\end{bmatrix}$: $(k_{\text{ex}} + k_{\text{en}}) \times 1$ vector of parameters

$\mathbf{X} = [\mathbf{X_{ex}} | \mathbf{V_{en}}]$

$\mathbf{Z} = [\mathbf{X_{ex}} | \mathbf{X_{iv}}]$

$\boldsymbol{\Gamma_{ex}}$: $k_{\text{ex}} \times k_{\text{en}}$ matrix of parameters

$\boldsymbol{\Gamma_{iv}}$: $k_{\text{iv}} \times k_{\text{en}}$ matrix of parameters

$\boldsymbol{\Gamma} = \begin{bmatrix}\boldsymbol{\Gamma_{ex}} \\ \boldsymbol{\Gamma_{iv}}\end{bmatrix}$: $(k_{\text{ex}} + k_{\text{iv}}) \times k_{\text{en}}$ matrix of parameters

$\boldsymbol{e}$: $n \times 1$ vector of errors

$\mathbf{E}$: $n \times k_{\text{en}}$ matrix of errors

The 2SLS estimator of $\boldsymbol{\beta}$ is $\boldsymbol{b}$ given by

$$\boldsymbol{b} = \left\{\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X}\right\}^{-1}\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'y}$$

$$\text{Var}(\boldsymbol{b}) = s^2\left\{\mathbf{X'Z}(\mathbf{Z'Z})^{-1}\mathbf{Z'X}\right\}^{-1}$$

where

$s^2 = E_{SS}/(n - (k_{\text{ex}} + k_{\text{en}}))$: mean squared error

$$E_{SS} = \sum_{i=1}^{n} u_i^2$$

Hausman's Test of Endogeneity

Cameron and Trivedi (2010) present a special version of Hausman's test may be used to test whether one or more explanatory variables are endogenous. For a single explanatory variable, the test is

$$T_{H,1} = \frac{b_{2SLS} - b_{OLS}}{Var(b_{2SLS}) - Var(b_{OLS})}$$

The test statistic is distributed as a chi-square with one degree of freedom under the null hypothesis that the regressor is exogenous.

NCSS also provides an overall test of endogeneity of all the designated endogenous variables. This is calculated as

$$T_{H,ken} = (\boldsymbol{b_{2SLS}} - \boldsymbol{b_{OLS}})'\left(V(\boldsymbol{b_{2SLS}}) - V(\boldsymbol{b_{OLS}})\right)^{-1}(\boldsymbol{b_{2SLS}} - \boldsymbol{b_{OLS}})$$

The test statistic is distributed as a chi-square with k_{en} degrees of freedom under the null hypothesis that the regressors are exogenous. Note that $\boldsymbol{b_{2SLS}}$ and $\boldsymbol{b_{OLS}}$ represent only those regression coefficients corresponding to endogenous variables.

Data Structure

The data for 2SLS are entered as numeric variables, one column for each variable. An example of data appropriate for this procedure is given in Kmenta (2011) page 687. In that dataset, Q is food consumption per head, P is the ratio of food prices and general consumer prices, D is disposable income, F is the ratio of preceding year's prices received by farmers for products and general consumer prices, and A is time in years. This dataset is called **Kmenta687**.

Kmenta687 Dataset (subset)

Q	P	D	F	A
98.485	100.323	87.4	98	1
99.187	104.264	97.6	99.1	2
102.163	103.435	96.7	99.1	3
101.504	104.506	98.2	98.1	4
104.240	98.001	99.8	110.8	5
103.243	99.456	100.5	108.2	6
103.993	101.066	103.2	105.6	7
99.900	104.763	107.8	109.8	8
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the variables used in the analysis.

Variables

Dependent Variable

Specify the column containing the Y (dependent, response, or predicted) variable. This is the variable to be predicted by the independent (X) variables.

The values in this column must be numeric. Any text values will be treated as missing values and ignored.

Exogenous Variables

Specify any columns containing the X's (independent or predictor) variables that are exogenous. Exogenous variables are outside the model or autonomous. They are not correlated with the random error component.

These variables should contain numeric values that are either continuous or binary. Any text values will be treated as missing values and ignored.

Endogenous Variables

Specify the columns containing the X's (independent or predictor) variables that are endogenous. Endogenous variables are correlated with the random error component. The first stage of the two-stage least squares algorithm regresses these variables on the exogenous and instrument variables. The predicted values of the endogenous variables are then used in the second stage as predictors of the dependent variable along with any exogenous variables.

These variables should contain numeric values that are either continuous or binary. Any text values will be treated as missing values and ignored.

Instrument Variables

Specify the columns containing instrument variables. Instrument variables are included with the exogenous variables as the independent variables in the first stage regression. Each endogenous variable is regressed on these variables. These variables are not included in the second-stage regression model.

These variables should contain numeric values that are either continuous or binary. Any text values will be treated as missing values and ignored.

Remove Intercept

Specifies whether to include the intercept terms in the regression equations. If checked, the intercept is not included. If not checked, the intercept is included.

Impact on Reports

Removing the intercept from the regression equation distorts many of the common regression measures such as R^2, mean square error, and t-tests. You should not use these measures when the intercept has been omitted.

Reports Tab

The following options control which reports and plots are displayed.

Select Reports

Run Summary ... Residuals

Each of these options specifies whether the indicated report is calculated and displayed. Note that since some of these reports provide results for each row, they may be too long for normal use when requested on large databases.

Report Options

Precision

Specifies the precision of numbers in the report when the *Decimal Places* are set to *General*. Single precision will display seven-place accuracy, while the double precision will display thirteen-place accuracy.

Variable Names

This option lets you select whether to display variable names, variable labels, or both.

Report Options – Decimal Places

Mean ... Y

Specify the number of digits after the decimal point to display on the output of values of this type. Note that this option in no way influences the accuracy with which the calculations are done.

Enter *General* to display all digits available. The number of digits displayed by this option is controlled by whether the *Precision* option is *Single* or *Double*.

Plots Tab

These options control the titles and style files used on each of the plots.

Select Plots

Residuals vs Yhat ... Probability Plot

Indicate whether to display these plots. Click the plot format button to change the plot settings.

Storage Tab

These options let you specify if, and where on the dataset, various statistics are stored.

Warning: Any data already in these columns are replaced by the new data. Be careful not to specify columns that contain important data.

Data Storage Options

Predicted Values, Residuals

Indicate the column in which these row-by-row values are stored. If the column contains data values, they will be replaced.

Example 1 – Two-Stage Least Squares (All Reports)

This section presents an example of how to run a Two-Stage Least Squares (2SLS) analysis of the Kmenta687 data. In this dataset, Q is the dependent variable, D is the exogenous variable, P is the endogenous variable, and A and F are instrument variables.

You may follow along here by making the appropriate entries or load the completed template **Example 1** by clicking on Open Example Template from the File menu of the Two-Stage Least Squares window.

1 Open the Kmenta687 dataset.

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Click on the file **Kmenta687.NCSS**.
- Click **Open**.

2 Open the Two-Stage Least Squares window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Two-Stage Least Squares** procedure.
- On the menus, select **File**, then **New Template**. This will fill the procedure with the default template.

3 Specify the variables.

- On the Two-Stage Least Squares window, select the **Variables tab**.
- Set the **Dependent Variable** box to **Q**.
- Set the **Exogenous Variables** box to **D.**
- Set the **Endogenous Variables** box to **P**.
- Set the **Instrument Variables** box to **A, F**.

4 Specify the reports.

- Select the **Reports tab**.
- Make sure all reports are checked.

5 Specify the plots.

- Select the **Plots tab**.
- Make sure all plots are checked.

6 Run the procedure.

- From the Run menu, select **Run Procedure**. Alternatively, just click the green Run button.

Run Summary

Item	Value	Item	Value
Dependent Variable	Q	Total Rows Processed	20
Number of Exogenous Variables	1	Unfiltered Rows	20
Number of Endogenous Variables	1	Unfiltered and Non-Missing Rows	20
Number of Instrument Variables	2	Intercept Included in Model	Yes
√MSE (2SLS)	1.966		
R^2 (OLS)	0.7638		

This report summarizes the 2SLS results. It presents the variables used, the number of rows used, etc. Note that the value of R2 is calculated from the regular regression of Y on the exogenous and endogenous variables.

Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Q	20	100.898	3.756	92.424	106.232
Intercept	20	1.000	0.000	1.000	1.000
D	20	97.535	11.830	75.100	127.100
P	20	100.019	5.926	86.498	113.490
A	20	10.500	5.916	1.000	20.000
F	20	96.625	12.709	68.600	110.800

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. This report is particularly useful for checking that the correct variables were selected.

Two-Stage Least Squares Estimation

Type	Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T Value b/Sb	P Value
Exogenous	Intercept	94.6333	7.9208	11.947	0.0000
Exogenous	D	0.3140	0.0469	6.689	0.0000
Endogenous	P	-0.2436	0.0965	-2.524	0.0218

Model
94.6333+0.3139918*D-0.2435565*P

This report presents the final results of the 2SLS estimation.

Comparison of Two-Stage Least Squares with Ordinary Least Squares

Type	Variable	2SLS Regression Coefficient b(2SLS,i)	OLS Regression Coefficient b(OLS,i)	2SLS Standard Error Sb(2SLS,i)	OLS Standard Error Sb(OLS,i)	Reg Coef Difference Z Value	P Value
Exo	Intercept	94.6333	99.8954	7.9208	7.5194		
Exo	D	0.3140	0.3346	0.0469	0.0454		
End	P	-0.2436	-0.3163	0.0965	0.0907	2.207	0.0273

Hausman's Combined Endogeneity Test

χ^2 Value	4.869
DF	1
P Value	0.0273

This report compares the 2SLS parameters with the OLS (ordinary least squares) parameters. A single degree-of-freedom Hausman z-test and associated p-value is provided to help assess whether each designated endogenous variable is in fact endogenous. A combined test is also provided that test whether all designated endogenous variables are indeed endogenous. If the p-value is small (less than 0.05), exogeneity is rejected and endogeneity is concluded.

First-Stage Ordinary Least Squares Estimation

Type	Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T Value b/Sb	P Value
Exogenous	Intercept	90.2678	3.2993	27.360	0.0000
Exogenous	D	0.6632	0.0414	16.011	0.0000
Instrument	A	-0.7370	0.0753	-9.792	0.0000
Instrument	F	-0.4884	0.0380	-12.847	0.0000
R^2	0.9434				

This series of reports presents the results of regressing each endogenous variable on the exogenous and instrument variables. It is important to pay particular attention to the R^2 value.

The report has the same definitions as in regular Multiple Regression.

Analysis of Variance Section

Term	DF	Sum of Squares	Mean Square	F-Ratio	P-Value
Intercept	1	203608.935	203608.935		
Model	2	202.385	101.193	26.172	0.0000
Error	17	65.729	3.866		
Total(Adjusted)	19	268.114			

This section reports the analysis of variance table. Note it based on the 2SLS estimates and so an R^2 value is not reported since it has no interpretation in this case.

Predicted Values and Residuals

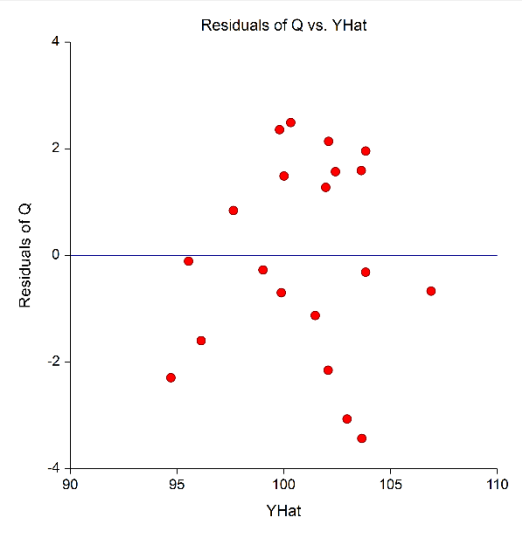
Predicted Values and Residuals

Row	Actual Q	Predicted Q	Residual
1	98.485	97.642	0.843
2	99.187	99.885	-0.698
3	102.163	99.804	2.359
4	101.504	100.014	1.490
5	104.240	102.101	2.139
6	103.243	101.966	1.277
7	103.993	102.422	1.571
8	99.900	102.966	-3.066
9	100.350	101.475	-1.125
10	102.820	100.328	2.492
11	95.435	95.543	-0.108
12	92.424	94.716	-2.292
13	94.535	96.133	-1.598
14	98.757	99.028	-0.271
15	105.797	103.839	1.958
16	100.225	103.655	-3.430
17	103.522	103.835	-0.313
18	99.929	102.080	-2.151
19	105.223	103.631	1.592
20	106.232	106.900	-0.668

This report shows the predicted values and residuals based on 2SLS. These are the values that are plotted below.

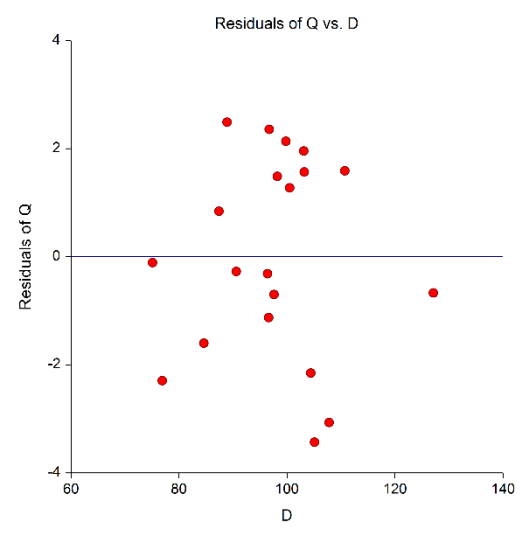
Residual vs Predicted Plot

This plot should always be examined. The preferred pattern to look for is a point cloud or a horizontal band. A wedge or bowtie pattern is an indicator of nonconstant variance, a violation of a critical assumption. A sloping or curved band signifies inadequate specification of the model. A sloping band with increasing or decreasing variability suggests nonconstant variance and inadequate specification of the model.



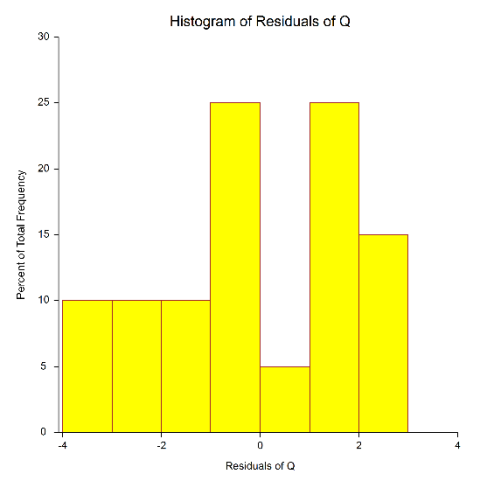
Residual vs Predictor(s) Plot

This is a scatter plot of the residuals versus each exogenous variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.



Histogram

The purpose of the histogram of the residuals is to evaluate whether they are normally distributed. Unless you have a large sample size, it is best not to rely on the histogram for visually evaluating the normality of the residuals. The better choice will be the normal probability plot.



Normal Probability Plot of Residuals

If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate outliers, curvature at both ends of the plot indicates long or short distributional tails, convex or concave curvature indicates a lack of symmetry, and gaps or plateaus or segmentation in the normal probability plot may require a closer examination of the data or model. Of course, use of this graphic tool with very small sample sizes is not recommended.