2-Stage Least-Squares Regression

Two Stage Least Squares regression may be used in the situation where the errors of independent and dependent variables are known (or likely) to be correlated. For example, the market price of a commodity and the demand for that commodity are non-recursive, that is, demand affects price and price affects demand. Prediction variables are "explanatory" variables to explain variability of the dependent variable. However, there may be other "instrumental" variables that predict one or more of these explanatory variables in which the errors are not correlated. If we first predict the explanatory variables with these instrumental variables and use the predicted values, we reduce the correlation of the errors with the dependent variable.

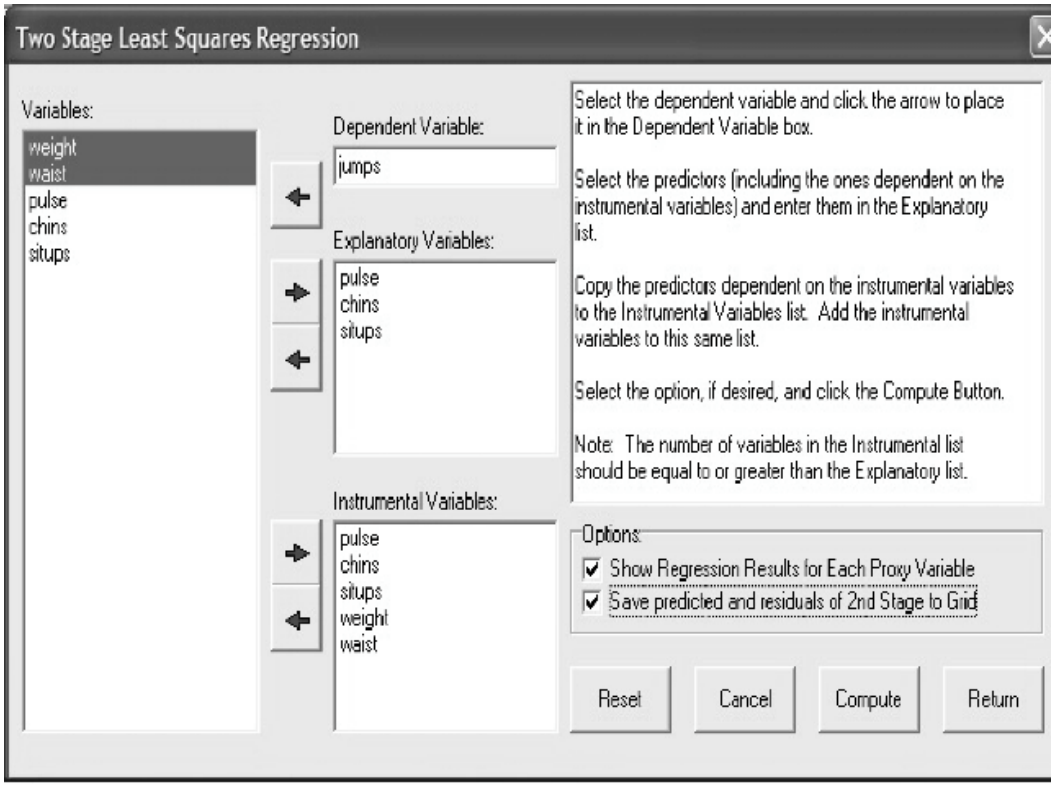
In this procedure, the user first selects the dependent variable of the study. Next, the explanatory variables (predictors) are entered. Finally, the instrumental variables AND the explanatory variables affected by these instrumental variables are entered into the instrumental variables list.

The two stages of this procedure are performed as follows:

Stage 1. The instrumental variables are identified as those in the instrumental list that are not in the explanatory list. The explanatory variables that are listed in both the explanatory and the instrumental lists are those for which predictions are to be obtained. These predicted scores are referred to as "proxy" scores. The predictions are obtained by regressing each explanatory variable listed in both lists with all of the remaining explanatory variables and instrumental variables. The predicted scores are obtained and stored in the data grid with a "P_" appended to the beginning of the original predictor name.

Stage 2. Once the predicted values are obtained, an OLS regression is performed with the dependent variable regressed on the proxy variables and the other explanatory variables not predicted in the previous stage.

In the following example, the cansas.LAZ file is analyzed. The dependent variable is the height of individual jumps. The explanatory (predictor) variables are pulse rate, no. of chinups and no. of situps the individual completes. These explanatory variables are thought to be related to the instrumental variables of weight and waist size. In the dialog box for the analysis, the option has been selected to show the regression for each of the explanatory variables that produces the predicted variables to be used in the final analysis. Results are shown below:




```
FILE: C:\Projects\LazStats\cansas.LAZ

Dependent = jumps
Explanatory Variables:
pulse
chins
situps
Instrumental Variables:
pulse
chins
situps
weight
waist
Proxy Variables:
P_pulse
P_chins
P_situps
```

Analysis for P_pulse

```
Dependent: pulse
Independent:
chins
situps
weight
waist

Means
Variables      chins      situps      weight       waist       pulse

                9.450     145.550     178.600      35.400      56.100

Standard Deviations
Variables      chins      situps      weight       waist       pulse

                5.286      62.567      24.691       3.202       7.210
```

No. of valid cases = 20

CORRELATION MATRIX

VARIABLE	chins	situps	weight	waist	pulse
chins	1.000	0.696	-0.390	-0.552	0.151
situps	0.696	1.000	-0.493	-0.646	0.225
weight	-0.390	-0.493	1.000	0.870	-0.366
waist	-0.552	-0.646	0.870	1.000	-0.353
pulse	0.151	0.225	-0.366	-0.353	1.000

Dependent variable: pulse

Variable	Beta	B	Std.Err.	t	Prob.>t	VIF	TOL
chins	-0.062	-0.084	0.468	-0.179	0.860	2.055	0.487
situps	0.059	0.007	0.043	0.158	0.876	2.409	0.415
weight	-0.235	-0.069	0.146	-0.471	0.644	4.360	0.229
waist	-0.144	-0.325	1.301	-0.249	0.806	5.832	0.171
Intercept	0.000	79.673	32.257	2.470	0.026		

SOURCE	DF	SS	MS	F	Prob.>F
Regression	4	139.176	34.794	0.615	0.6584
Residual	15	848.624	56.575		
Total	19	987.800			

R2 = 0.1409, F = 0.62, D.F. = 4 15, Prob>F = 0.6584
Adjusted R2 = -0.0882
Standard Error of Estimate = 7.52

Analysis for P chins

Dependent: chins
Independent:
pulse
situps
weight
waist

Means

Variables	pulse	situps	weight	waist	chins
	56.100	145.550	178.600	35.400	9.450

Standard Deviations

Variables	pulse	situps	weight	waist	chins
	7.210	62.567	24.691	3.202	5.286

No. of valid cases = 20

CORRELATION MATRIX

VARIABLE	pulse	situps	weight	waist	chins
pulse	1.000	0.225	-0.366	-0.353	0.151
situps	0.225	1.000	-0.493	-0.646	0.696
weight	-0.366	-0.493	1.000	0.870	-0.390
waist	-0.353	-0.646	0.870	1.000	-0.552
chins	0.151	0.696	-0.390	-0.552	1.000

Dependent variable: chins

Variable	Beta	B	Std.Err.	t	Prob.>t	VIF	TOL
pulse	-0.035	-0.026	0.142	-0.179	0.860	1.162	0.861
situps	0.557	0.047	0.020	2.323	0.035	1.775	0.564
weight	0.208	0.045	0.080	0.556	0.586	4.335	0.231
waist	-0.386	-0.638	0.700	-0.911	0.377	5.549	0.180
Intercept	0.000	18.641	20.533	0.908	0.378		

SOURCE	DF	SS	MS	F	Prob.>F
Regression	4	273.089	68.272	3.971	0.0216

Residual	15	257.861	17.191
Total	19	530.950	

R2 = 0.5143, F = 3.97, D.F. = 4 15, Prob>F = 0.0216
Adjusted R2 = 0.3848
Standard Error of Estimate = 4.15

Analysis for P situps

Dependent: situps
Independent:
pulse
chins
weight
waist

Means

Variables	pulse	chins	weight	waist	situps
	56.100	9.450	178.600	35.400	145.550

Standard Deviations

Variables	pulse	chins	weight	waist	situps
	7.210	5.286	24.691	3.202	62.567

No. of valid cases = 20

CORRELATION MATRIX

VARIABLE	pulse	chins	weight	waist	situps
pulse	1.000	0.151	-0.366	-0.353	0.225
chins	0.151	1.000	-0.390	-0.552	0.696
weight	-0.366	-0.390	1.000	0.870	-0.493
waist	-0.353	-0.552	0.870	1.000	-0.646
situps	0.225	0.696	-0.493	-0.646	1.000

Dependent variable: situps

Variable	Beta	B	Std.Err.	t	Prob.>t	VIF	TOL
pulse	0.028	0.246	1.555	0.158	0.876	1.162	0.861
chins	0.475	5.624	2.421	2.323	0.035	1.514	0.660
weight	0.112	0.284	0.883	0.322	0.752	4.394	0.228
waist	-0.471	-9.200	7.492	-1.228	0.238	5.322	0.188
Intercept	0.000	353.506	211.726	1.670	0.116		

SOURCE	DF	SS	MS	F	Prob.>F
Regression	4	43556.048	10889.012	5.299	0.0073
Residual	15	30820.902	2054.727		
Total	19	74376.950			

R2 = 0.5856, F = 5.30, D.F. = 4 15, Prob>F = 0.0073
Adjusted R2 = 0.4751
Standard Error of Estimate = 45.33

Second Stage (Final) Results

Means

Variables	P_pulse	P_chins	P_situps	jumps
	56.100	9.450	145.550	70.300

Standard Deviations

Variables	P_pulse	P_chins	P_situps	jumps
	2.706	3.791	47.879	51.277

No. of valid cases = 20

CORRELATION MATRIX
VARIABLE

	P_pulse	P_chins	P_situps	jumps
P_pulse	1.000	0.671	0.699	0.239
P_chins	0.671	1.000	0.847	0.555
P_situps	0.699	0.847	1.000	0.394
jumps	0.239	0.555	0.394	1.000

Dependent variable: jumps

Variable	Beta	B	Std.Err.	t	Prob.>t	VIF	TOL
P_pulse	-0.200	-3.794	5.460	-0.695	0.497	2.041	0.490
P_chins	0.841	11.381	5.249	2.168	0.046	3.701	0.270
P_situps	-0.179	-0.192	0.431	-0.445	0.662	3.979	0.251
Intercept	0.000	203.516	277.262	0.734	0.474		

SOURCE	DF	SS	MS	F	Prob.>F
Regression	3	17431.811	5810.604	2.858	0.0698
Residual	16	32526.389	2032.899		
Total	19	49958.200			

R2 = 0.3489, F = 2.86, D.F. = 3 16, Prob>F = 0.0698
Adjusted R2 = 0.2269
Standard Error of Estimate = 45.09