

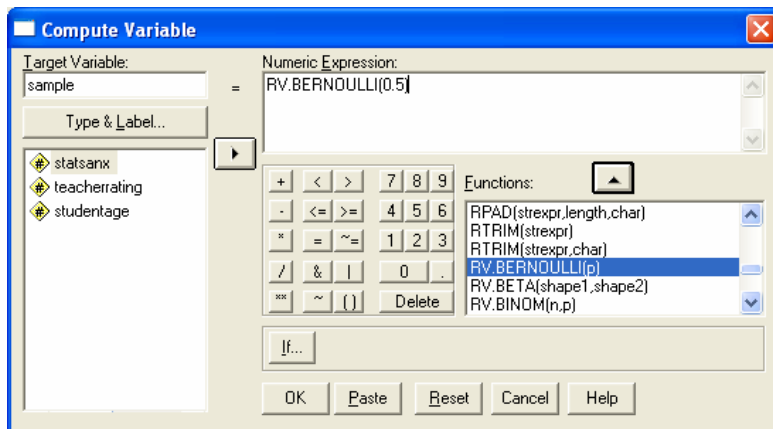
Multiple Regression Lab: Cross-validation

Note. Cross-validation is typically conducted with a large data set.

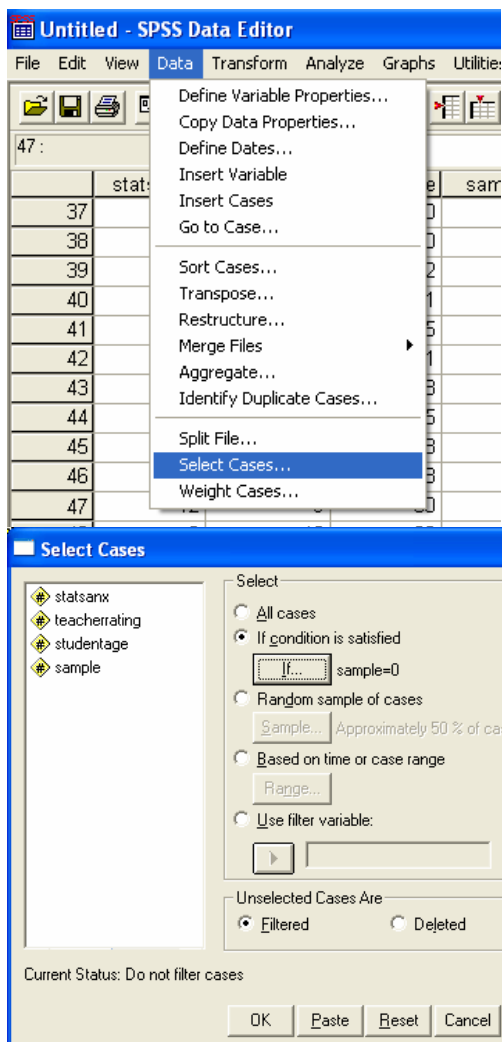
An educational researcher wanted to predict student's anxiety levels using information about teacher rating and student age.

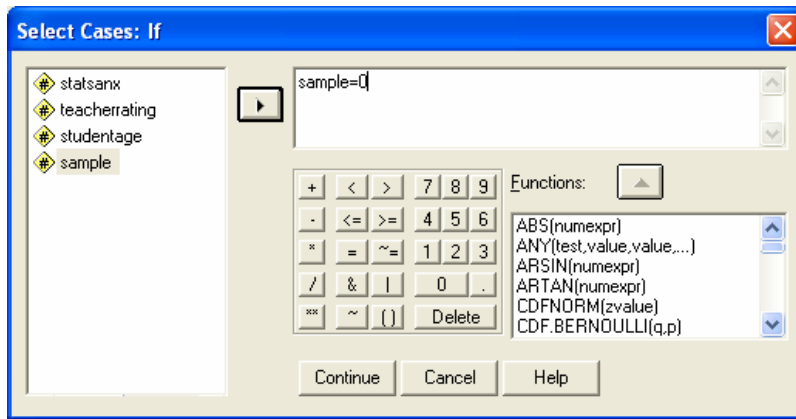
Statistics Anxiety	Teacher Rating	Student Age	Statistics Anxiety	Teacher Rating	Student Age
6	10	21	5	13	25
14	9	34	7	12	31
7	9	22	9	10	28
4	12	26	9	10	25
6	9	20	11	9	28
8	10	27	11	11	28
8	10	22	12	9	30
9	11	24	8	10	20
7	9	24	7	13	23
9	11	24	8	10	21
11	9	26	9	9	27
5	10	30	8	8	25
9	9	22	5	12	20
7	10	23	8	8	22
15	7	31	4	11	21
8	10	18	4	10	22
12	10	31	8	11	28
8	11	29	12	7	33
7	10	18	4	11	22
4	10	24	8	9	30
8	11	25			
10	11	27			
8	9	27			
13	6	32			
11	9	28			
10	8	28			
8	10	25			
7	11	21			
12	10	26			
7	10	29			
5	11	27			
12	10	30			
11	9	32			
9	11	28			
6	12	24			
7	10	25			
4	12	20			
3	11	20			
9	9	22			
6	11	21			

Create a new variable that will randomly split the data set into two parts. This approach will not guarantee an exact 50% split, but it causes no problems.

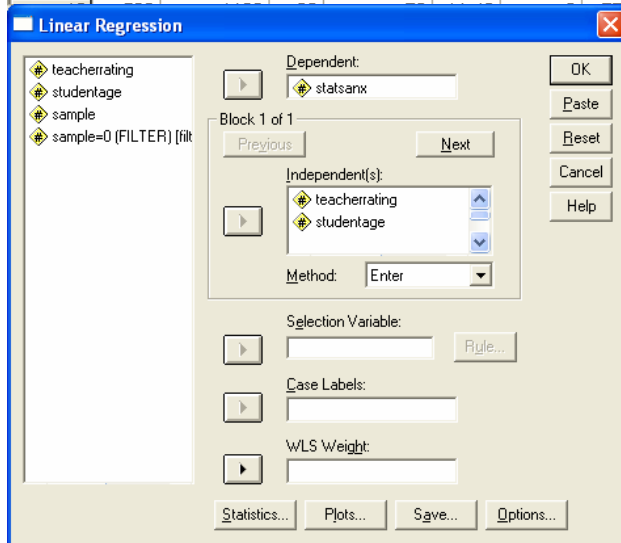
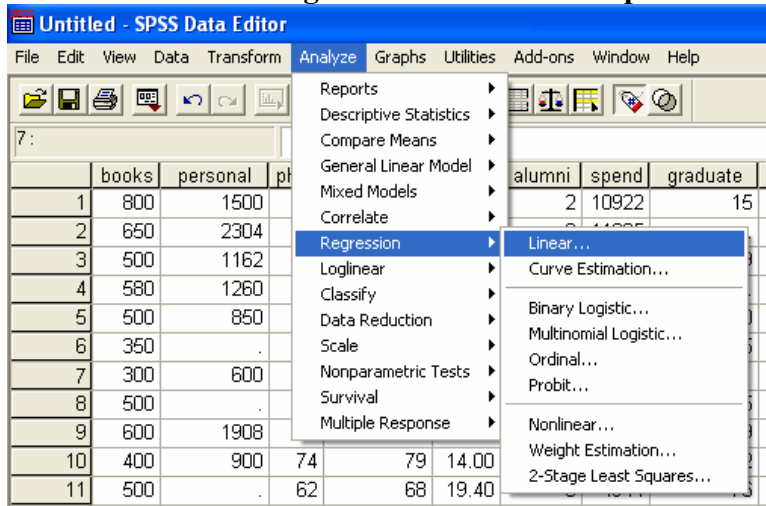


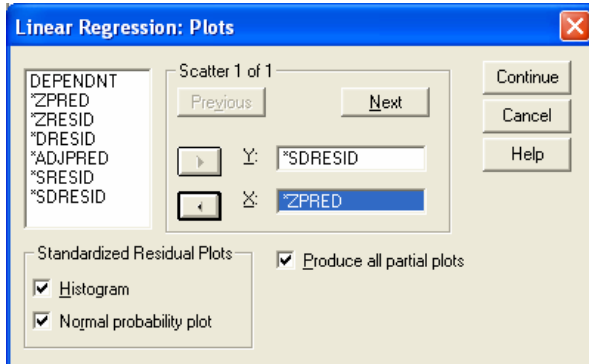
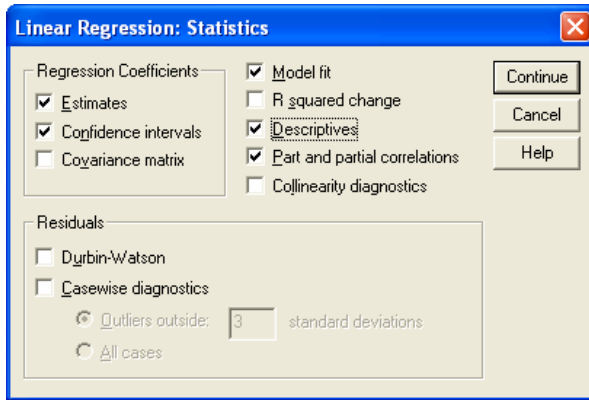
Select the Sample = 0 data to estimate the linear regression equation.





Estimate the linear regression model for Sample = 0.





Regression

Descriptive Statistics

	Mean	Std. Deviation	N
statsanx	7.62	2.155	26
teacherrating	10.35	1.198	26
studentage	24.88	3.756	26

Correlations

		statsanx	teacherrating	studentage
Pearson Correlation	statsanx	1.000	-.504	.409
	teacherrating	-.504	1.000	-.035
	studentage	.409	-.035	1.000
Sig. (1-tailed)	statsanx	.	.004	.019
	teacherrating	.004	.	.432
	studentage	.019	.432	.
N	statsanx	26	26	26
	teacherrating	26	26	26
	studentage	26	26	26

Variables Entered/Removed^a

Model	Variables Entered	Variables Removed	Method
1	studentage, teacherrating ^a	.	Enter

a. All requested variables entered.

b. Dependent Variable: statsanx

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.638 ^a	.408	.356	1.730

a. Predictors: (Constant), studentage, teacherrating

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	47.338	2	23.669	7.911	.002 ^a
	Residual	68.816	23	2.992		
	Total	116.154	25			

- a. Predictors: (Constant), studentage, teacherrating
- b. Dependent Variable: statsanx

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B		Correlations		
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part
1	(Constant)	11.1404	3.846		2.896	.008	3.184	19.097			
	teacherrating	-.8819	.289	-.490	-3.052	.006	-1.480	-.284	-.504	-.537	-.490
	studentage	.2250	.092	.392	2.441	.023	.034	.416	.409	.454	.392

- a. Dependent Variable: statsanx

Residual Statistics Summary – Omitted for brevity
Charts – Omitted for brevity

For Sample=0, the regression model is:
 Predicted
 Stats anxiety = 11.1404-0.8819*teacherrating+0.2250*studentage

Remove the Sample=0 restriction.

The screenshot shows the SPSS Data Editor interface. The 'Data' menu is open, and 'Select Cases...' is highlighted. The 'Select Cases' dialog box is displayed, showing the following options:

- Select:**
 - All cases
 - If condition is satisfied (with 'If...' button and 'sample=1' text)
 - Random sample of cases (with 'Sample...' button and 'Approximately 50 % of cases' text)
 - Based on time or case range (with 'Range...' button)
 - Use filter variable: (with a selection arrow and an empty text box)
- Unselected Cases Are:**
 - Filtered
 - Deleted

Current Status: Filter cases by values of filter_\$

Buttons: OK, Paste, Reset, Cancel, Help

Calculate predicted statistics anxiety scores for individuals in the other sample (Sample=1).

Untitled - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities

47 :

	statsanx	
37	4	1
38	3	1

Compute...
 Recode
 Visual Bander...
 Count...
 Rank Cases...
 Automatic Recode...

Compute Variable

Target Variable: predictedanx
 Numeric Expression: $11.1404 - 0.8819 * \text{teacherrating} + 0.2250 * \text{studentage}$

Type & Label...
 statsanx
 teacherrating
 studentage
 sample
 sample=1 (FILTER) [filt]

Functions:
 ABS(numexpr)
 ANY(test,value,value,...)
 ARSIN(numexpr)
 ARTAN(numexpr)
 CDFNORM(zvalue)
 CDF.BERNOULLI(q,p)

If... sample=1

OK Paste Reset Cancel Help

Compute Variable: If Cases

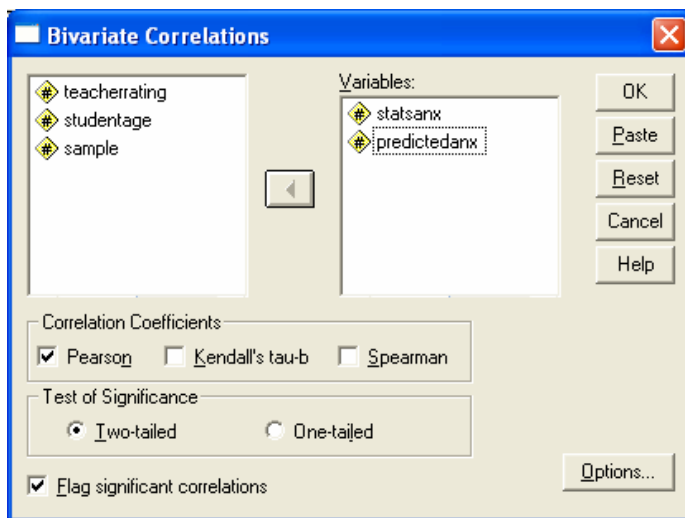
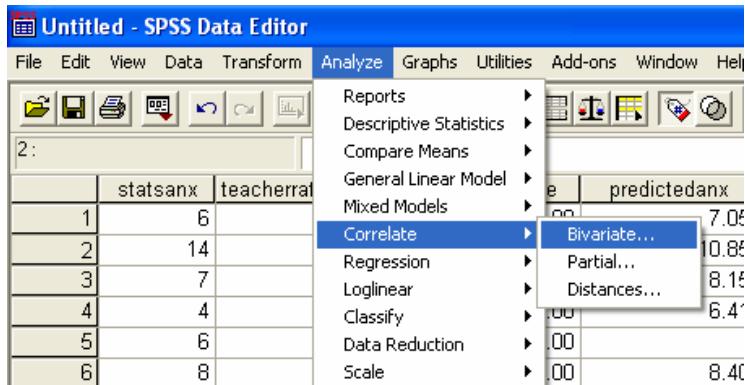
statsanx
 teacherrating
 studentage
 sample
 sample=1 (FILTER) [filt]

Include all cases
 Include if case satisfies condition:
 sample=1

Functions:
 ABS(numexpr)
 ANY(test,value,value,...)
 ARSIN(numexpr)
 ARTAN(numexpr)
 CDFNORM(zvalue)
 CDF.BERNOULLI(q,p)

Continue Cancel Help

Finally, estimate the correlation between the actual statistics anxiety scores of Sample=1 and the predicted statistics anxiety scores obtained from the regression model (i.e., predictedanx).



Correlations

Correlations

		statsanx	predictedanx
statsanx	Pearson Correlation	1	.764**
	Sig. (2-tailed)	.	.000
	N	60	34
predictedanx	Pearson Correlation	.764**	1
	Sig. (2-tailed)	.000	.
	N	34	34

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

$R^2 = .408$ from the Sample=0 linear regression model.

The squared correlation between predictedanx and statsanx was $.764^2 = .584$.

Since there was more than a .10 change (.408 vs. .584) we conclude the model would likely not be valid with other data. In fact, it was an anomaly that the squared correlation was higher than the R^2 for the original model; this would seldom happen in reality.

A much larger data set would be needed for a 'true' cross-validation study. If the model had been validated, we would have re-estimated the regression model using ALL of the data and used that model for future analyses.