Chapter 317

Mediation Analysis

Introduction

This procedure performs mediation analysis using linear regression. Interest focuses on the interrelationship of three numeric variables Y, X, and M. This interrelationship can be adjusted for a number of other variables called covariates. Also, the analysis can also use one of two robust regression algorithms when the assumptions of ordinary least squares do not seem valid.

An in-depth discussion of mediation can be found in Hayes (2018) and MacKinnon (2008).

Mediation Model

A mediation model approximates the relationship between an independent variable (X) and a dependent variable (Y) when a mediator variable (M) is included. The mediation model assumes that X influences M which in turn influences Y. It also allows for an additional effect of X directly on Y over and above the effect that goes through M.

A popular method for testing for mediation is that of Baron and Kenny (1986). In this method, the following three linear regression models are fit.

(1) $Y = i_1 + c_1 X + bM + e_1$

(2) $Y = i_2 + c_2 X + e_2$

(3) $M = i_3 + aX + e_3$

The **indirect**, or **mediated**, effect is measured by the *ab* product estimated from equations 1 and 3. The regression coefficient c_2 from equation 2 is called the **total** effect. Similarly, the regression coefficient c_1 from equation 1 is called the **direct** effect.

Covariates

Often, additional independent variables are available. These variables may not be of direct interest in the mediation analysis, but their influence on the results is likely. These additional variables are called covariates. They may be specified as part of the analysis and they will be included in all three regressions. Both numeric and categorical covariates can be specified.

Testing the Mediated Effect

The total, direct, and indirect effects are all of interest in a mediation analysis. However, the main hypothesis to be tested is whether the indirect effect, *ab*, is significant. As shown in MacKinnon (2008), this may be done in two ways.

Large Sample Wald Test

A Wald test can be constructed as follows to test whether *ab* is zero.

$$z = \frac{ab}{s_{ab}}$$

where

$s_{ab} = \sqrt{(as_b)^2 + (bs_a)^2}$ (first-order standard error of Sobel (1982))

or

$s_{ab} = \sqrt{(as_b)^2 + (bs_a)^2 + (s_a s_b)^2}$ (second-order standard error of Baron and Kenny (1986))

The first-order standard error is used in several specialized software programs such as EQS, Mplus, and LISREL. MacKinnon (2008) page 74 notes that simulation studies have shown that the first-order equation 'performs better' than the second-order equation, so this is the estimator that we recommend. Both methods are available in **NCSS**.

Several authors have noted that the product *ab* is not normally distributed, so they criticize the use of the Wald test. Often, bootstrapping is the recommended alternative.

Bootstrapping

Bootstrapping was developed (see Efron and Tibshirani, 1993) to provide standard errors and confidence intervals in situations such as this in which the standard assumptions are not valid. The method is simple in concept, but it requires extensive computation time.

Assume that the sample of *N* subjects is actually the population and draw *B* samples (*B* is usually over 1000) of *N* from the original dataset, with replacement. For each bootstrap sample, compute and store the *ab* product.

The bootstrap sampling process has provided *B* estimates of the *ab*. The standard deviation of these *B* estimates is the bootstrap estimate of the standard error of *ab*. Using this estimate, a Wald-type z-test can be constructed.

The *bootstrap confidence interval* is found by arranging the *B* values in sorted order and selecting the appropriate percentiles from the list. For example, a 90% bootstrap confidence interval for the difference is given by fifth and ninety-fifth percentiles of the bootstrap *ab* values.

The main assumption made when using the bootstrap method is that the sample approximates the population fairly well. Because of this assumption, bootstrapping does not work well for small samples in which there is little likelihood that the sample is representative of the population.

Robust Regression

Regular multiple regression is optimum when all of its assumptions are valid. When some of these assumptions are invalid, least squares regression can perform poorly. Thorough residual analysis can point to these assumption breakdowns and allow you to work around these limitations. However, this residual analysis is time consuming and requires a great deal of training.

Robust regression provides an alternative to least squares regression that works with less restrictive assumptions. Specifically, it provides much better regression coefficient estimates when outliers are present in the data. Outliers violate the assumption of normally distributed residuals in least squares regression. They tend to distort the least squares coefficients by having more influence than they deserve. Typically, you would expect that the weight attached to each observation would be about 1/*N* in a dataset with *N* observations. However, outlying observations may receive a weight of 10, 20, or even 50 percent. This leads to serious distortions in the estimated coefficients.

Because of this distortion, these outliers are difficult to identify since their residuals are much smaller than they should be. When only one or two independent variables are used, these outlying points may be visually detected in various scatter plots. However, the complexity added by additional independent variables often hides the outliers from view in scatter plots. Robust regression down-weights the influence of outliers. This makes residuals of outlying observations larger and easier to spot. Robust regression is an iterative procedure that seeks to identify outliers and minimize their impact on the coefficient estimates.

The amount of weighting assigned to each observation in robust regression is controlled by a special curve called an *influence function*. There are two influence functions available in **NCSS**: Huber and Tukey.

Although robust regression can particularly benefit untrained users, careful consideration should be given to the results. Essentially, robust regression conducts its own residual analysis and down-weights or completely removes various observations. You should study the weights it assigns to each observation, determine which observations have been largely eliminated, and decide if you want these observations in your analysis.

Further details of robust regression can be found in the Robust Regression procedure chapter. If you find yourself using the technique often, we suggest that you study a text on regression analysis. Most texts have chapters on robust regression. A good introductory discussion of robust regression is found in Hamilton (1991). A more thorough discussion is found in Montgomery and Peck (1992).

Standard Errors of Robust Regression Coefficients

The standard errors, confidence intervals, and t-tests produced by the weighted least squares assume that the weights are ***fixed***. Of course, this assumption is violated in robust regression since the weights are calculated from the sample residuals, which are ***random***. **NCSS** can produce standard errors, confidence intervals, and t-tests that have been adjusted to account for the random nature of the weights. The method described next was given in Hamilton (1991).

Let $\phi(u)$ represent the derivative of the influence function $\psi(u)$. To find adjusted standard errors, etc., take the following steps:

1. Calculate a and λ using

$$a = \frac{\sum_i \phi(u_i)}{N}, \quad \lambda = 1 + \frac{(p+1)(1-a)}{Na}$$

where

for Huber estimation

$$\phi(u) = \begin{cases} 1 & \text{if } |u| \le c \\ 0 & \text{if } |u| > c \end{cases}$$

for Tukey's biweight estimation

$$\phi(u) = \begin{cases} \left[1 - \frac{u^2}{c^2}\right]\left[1 - 5\frac{u^2}{c^2}\right] & \text{if } |u| \le c \\ 0 & \text{if } |u| > c \end{cases}$$

2. Define a set of pseudo values of y_i using

$$\tilde{y}_i = \hat{y}_i + \frac{\lambda s}{a}\psi(u_i)$$

3. Regress $\widetilde{\mathbf{Y}}$ on $\boldsymbol{X}$. The standard errors, t-tests, and confidence intervals from this regression are asymptotically correct for the robust regression.

This method is not without criticism. The main criticism is that the results depend on the choices of the MAD scale factor (default = 0.6745) and the tuning constant, *c*. Changing these values may cause large changes in the resulting tests and confidence intervals.

Data Structure

The data are entered in three or more columns. An example of data appropriate for this procedure is shown below. These data are from a hypothetical study of the relationship of several variables with a person's water consumption. The dataset includes the columns Temp (average daily temperature at 2 p.m. in May), Thirst (an index of a person's thirst on a scale of 1 to 10), Age (subject's age), Adults (number of adults in the household), and Water (water consumption for May). The data are contained in the *Mediation* dataset. The first few rows of this dataset are shown below.

Mediation Dataset

Temp	Thirst	Age	Adults	Water
68	9	75	3	84
36	5	40	2	33
34	4	48	2	31
20	3	32	2	21
53	7	18	2	51
45	6	36	2	41
33	4	20	2	34
62	8	24	1	57
66	8	32	1	58
69	9	76	2	23
53	7	52	2	47
36	5	19	2	37
61	8	44	2	55
56	7	28	1	52
70	9	40	2	63

Missing Values

Rows with missing values in any columns being analyzed are ignored in all three regressions. This is often called *row-wise deletion*.

Example 1 – Mediation Analysis (OLS Solution)

This section presents an example of how to run a mediation analysis of the data presented earlier in this chapter. The data are in the *Mediation* dataset. In this example, it is supposed that the amount of water consumption (Y) is directly related to the temperature (X). The mediator is an index of how thirsty each subject was. This mediator is contained in the column named Thirst. Thus, for this example set X = Temp, M = Thirst, and Y = Water.

Setup

To run this example, complete the following steps:

1 Open the Mediation example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Mediation** and click **OK**.

2 Specify the Mediation Analysis procedure options

- Find and open the **Mediation Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables, Model Tab	
Y (Dependent Variable)	**Water**
X (Independent Variable)	**Temp**
M (Mediator Variable)	**Thirst**
Terms	**1-Way**
Robust, Bootstrap Tab	
Estimation Method	**Ordinary Least Squares**
Bootstrap Calculations	**Checked**
Samples	**3000**
Reports Tab	
All Reports	**Checked**
Plots Tab	
All Plots	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Y (Dependent Variable)	Water	Number Processed	100
X (Independent Variable)	Temp	Number Used in Estimation	100
M (Mediator Variable)	Thirst	Number Filtered Out	0
Number of Covariates	0	Number with X's Missing	0
Weight Variable	None	Number with Weight Missing	0
Robust Method	None (OLS)	Number with Y Missing	0
Bootstrap Samples	3000	Sum of Robust Weights	100
Bootstrap Random Seed	4655484		
Bootstrap C.L. Type	Reflection		

Run Summary Details

Regression Model	R^2	S from MSE	Completion Status
Y = X M	0.5482	10.78697	Normal Completion
Y = X	0.5475	10.74047	Normal Completion
M = X	0.4311	1.515404	Normal Completion

These reports summarize the mediation analysis results. They present the estimation method used, the variables used, the number of rows used, and the R^2 of each of the three models.

Descriptive Statistics

Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Temp	100	50.28	14.83449	17	93
Thirst	100	6.38	1.998889	1	10
Water	100	46.71	15.88551	15	86

For each variable, the count, arithmetic mean, standard deviation, minimum, and maximum are computed. Note that these statistics use the robust weights if robust regression was used. This report is particularly useful for checking that the correct variables were selected.

Correlation Matrix

Correlation Matrix

	Temp	Thirst	Water
Temp	1.0000	0.6565	0.7399
Thirst	0.6565	1.0000	0.4654
Water	0.7399	0.4654	1.0000

Pearson correlations are given for all variables. Outliers, nonnormality, nonconstant variance, and nonlinearities can all impact these correlations. Note that these correlations may differ from pair-wise correlations generated by the correlation matrix program because of the different ways the two programs treat rows with missing values. The method used here is row-wise deletion.

These correlation coefficients show which independent variables are highly correlated with the dependent variable and with each other. Independent variables that are highly correlated with one another may cause collinearity problems.

Direct, Indirect, and Total Effects

Direct, Indirect, and Total Effects

Estimation:	Ordinary least squares
Sb(Indirect):	First-order
Bootstrap:	Number of samples = 3000, Random seed = 4655484, Confidence limit type = Reflection
Y:	Water
X:	Temp
M:	Thirst
Covariates:	0

			Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
Type of Effect	**Regression Coefficient b(i)**	**Standard Error Sb(i)**	**Test Statistic**	**P-Value**	**Lower**	**Upper**
Total	0.7923434	0.07276677	10.889	0.0000	0.6479401	0.9367467
Direct (X → Y)	0.8175452	0.0968889	8.438	0.0000	0.6252475	1.009843
Indirect (X → M → Y)						
Normal Theory	-0.02520181	0.06367939	-0.396	0.6923	-0.1500111	0.0996075
Bootstrap	-0.02520181	0.07635163	-0.330	0.7413	-0.09416629	0.08528873

This report shows the main results of the mediation analysis. The estimated effects are copied from the three regression reports given later. The bootstrap results are copied from the Bootstrap Report given next.

One focus of the mediation analysis is whether the Indirect Effect is statistically significant. In this example it is not.

Bootstrap Report for Indirect Effect (AB)

Bootstrap Report for Indirect Effect (AB)

Estimation Results			Bootstrap Confidence Interval Limits		
Parameter	**Estimate**		**Confidence Level**	**Lower**	**Upper**
Indirect Effect (AB)					
Original Value	-0.02520181	|	95%	-0.09416629	0.08528873
Bootstrap Mean	-0.03120935				
Bias (BM - OV)	-0.006007541				
Bias Corrected Value	-0.01919427				
Standard Error	0.07635163				

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000,
User-Entered Random Seed = 4655484.

Original Value

This is the parameter estimate obtained from the two regressions without bootstrapping.

Bootstrap Mean

This is the average of the parameter estimates of the bootstrap samples.

Bias (BM - OV)

This is an estimate of the bias in the original estimate. It is computed by subtracting the original value from the bootstrap mean.

Bias Corrected Value

This is an estimated of the parameter that has been corrected for its bias. The correction is made by subtracting the estimated bias from the original parameter estimate.

Standard Error

This is the bootstrap method's estimate of the standard error of the parameter estimate. It is simply the standard deviation of the parameter estimate computed from the bootstrap estimates.

Confidence Level

This is the confidence coefficient of the bootstrap confidence interval given to the right.

Bootstrap Confidence Limits - Lower and Upper

These are the limits of the bootstrap confidence interval with the confidence coefficient given to the left. These limits are computed using the confidence interval method (percentile or reflection) designated on the Bootstrap panel.

Note that to be accurate, these intervals must be based on over a thousand bootstrap samples and the original sample must be representative of the population.

Regression Coefficients

Regression Coefficients of Y = X M

Estimation: Ordinary least squares
Y: Water
X: Temp
M: Thirst
Covariates: 0
Dependent: Water

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
			T-Statistic	P-Value	Lower	Upper
Intercept	7.421309	4.073763	1.822	0.0716	-0.6639827	15.5066
Temp	0.8175452	0.0968889	8.438	0.0000	0.6252475	1.009843
Thirst	-0.2848719	0.7190485	-0.396	0.6928	-1.711984	1.14224

R-Squared

$R^2 = 0.5482$

Regression Coefficients of Y = X

Estimation: Ordinary least squares
Y: Water
X: Temp
M: Thirst
Covariates: 0
Dependent: Water

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
			T-Statistic	P-Value	Lower	Upper
Intercept	6.870974	3.813104	1.802	0.0746	-0.6960065	14.43795
Temp	0.7923434	0.07276677	10.889	0.0000	0.6479401	0.9367467

R-Squared

$R^2 = 0.5475$

Regression Coefficients of M = X

Estimation:	Ordinary least squares
Y:	Water
X:	Temp
M:	Thirst
Covariates:	0
Dependent:	Thirst

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	T-Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
			T-Statistic	P-Value	Lower	Upper
Intercept	1.93187	0.5380015	3.591	0.0005	0.864224	2.999517
Temp	0.08846717	0.01026687	8.617	0.0000	0.06809291	0.1088414

R-Squared

$R^2 = 0.4311$

This report gives the coefficients, standard errors, and significance tests for the three individual regressions that were run for the mediation analysis.

Independent Variable

The names of the independent variables are listed here. The intercept is the value of the *Y* intercept.

Regression Coefficient b(i)

The regression coefficients are the least squares (or robust) estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in that particular X when the remaining *X's* are held constant. These coefficients are often called partial-regression coefficients since the effect of the other *X's* is removed.

Standard Error Sb(i)

The standard error of the regression coefficient, s_{b_i}, is the standard deviation of the estimate. It is used in hypothesis tests or confidence limits. Note that when robust fitting is used, these values depend on the option S_{IND}.

T-Statistic for the T-Test of H0: β(i) = 0

This is the t-test value for testing the hypothesis that $\beta_i = 0$ versus the alternative that $\beta_i \neq 0$ after removing the influence of all other *X's*. This *t*-value has *N-p*-1 degrees of freedom.

P-Value for the T-Test of H0: β(i) = 0

This is the *p*-value for the significance test of the regression coefficient. The *p*-value is the probability that this *t*-statistic will take on a value at least as extreme as the actually observed value, assuming that the null hypothesis is true (i.e., the regression estimate is equal to zero). If the *p*-value is less than alpha, say 0.05, the null hypothesis of equality is rejected. This *p*-value is for a two-tail test.

Lower and Upper 95% Confidence Interval Limits for β(i)

These are the lower and upper values of a $100(1-\alpha)\%$ confidence interval estimate for β_i based on a *t*-distribution with *N*-*p*-1 degrees of freedom. This interval estimate assumes that the residuals for the regression model are normally distributed.

The formulas for the lower and upper confidence limits are:

$$b_i \pm t_{1-\frac{\alpha}{2},N-p-1} s_{b_i}$$

Normality Tests

Normality Tests of Residuals from Y = X M

Test Name	Test Statistic to Test H0: Normal	Prob Level	Reject H0 at 20%?
Shapiro Wilk	0.609	0.0000	Yes
Anderson Darling	15.873	0.0000	Yes
D'Agostino Skewness	4.339	0.0000	Yes
D'Agostino Kurtosis	5.264	0.0000	Yes
D'Agostino Omnibus	46.537	0.0000	Yes

Normality Tests of Residuals from Y = X

Test Name	Test Statistic to Test H0: Normal	Prob Level	Reject H0 at 20%?
Shapiro Wilk	0.605	0.0000	Yes
Anderson Darling	16.067	0.0000	Yes
D'Agostino Skewness	4.282	0.0000	Yes
D'Agostino Kurtosis	5.276	0.0000	Yes
D'Agostino Omnibus	46.176	0.0000	Yes

Normality Tests of Residuals from M = X

Test Name	Test Statistic to Test H0: Normal	Prob Level	Reject H0 at 20%?
Shapiro Wilk	0.608	0.0000	Yes
Anderson Darling	10.581	0.0000	Yes
D'Agostino Skewness	-7.237	0.0000	Yes
D'Agostino Kurtosis	6.383	0.0000	Yes
D'Agostino Omnibus	93.114	0.0000	Yes

This report gives the results of applying several normality tests to the residuals from each of the three regressions. The Shapiro-Wilk test is probably the most popular, so it is given first. These tests are discussed in detail in the Normality Test section of the Descriptive Statistics procedure.

Analysis of Variance Reports

ANOVA for Model: Y = X M

Source	DF	R^2	Sum of Squares	Mean Square	F-Ratio	P-Value
Intercept	1		218182.4	218182.4		
Model	2	0.5482	13695.79	6847.895	58.852	0.0000
Temp	1	0.3316	8284.658	8284.658	71.199	0.0000
Thirst	1	0.0007	18.26341	18.26341	0.157	0.6928
Error	97	0.4518	11286.8	116.3588		
Total(Adjusted)	99	1.0000	24982.59	252.3494		

ANOVA for Model: Y = X

Source	DF	R^2	Sum of Squares	Mean Square	F-Ratio	P-Value
Intercept	1		218182.4	218182.4		
Model	1	0.5475	13677.53	13677.53	118.566	0.0000
Temp	1	0.5475	13677.53	13677.53	118.566	0.0000
Error	98	0.4525	11305.06	115.3578		
Total(Adjusted)	99	1.0000	24982.59	252.3494		

ANOVA for Model: M = X

Source	DF	R^2	Sum of Squares	Mean Square	F-Ratio	P-Value
Intercept	1		4070.44	4070.44		
Model	1	0.4311	170.5081	170.5081	74.249	0.0000
Temp	1	0.4311	170.5081	170.5081	74.249	0.0000
Error	98	0.5689	225.0519	2.296448		
Total(Adjusted)	99	1.0000	395.56	3.995556		

These ANOVA tables provide a line for each term in the model. They are especially useful when you have categorical covariates.

Source

This is the term from the design model.

Note that the name may become very long, especially for interaction terms. These long names may misalign the report. You can force the rest of the items to be printed on the next line by using the Skip Line After option in the Format tab. This should create a better-looking report when the names are extra-long.

DF

This is the number of degrees of freedom that the model is degrees of freedom is reduced when this term is removed from the model. This is the numerator degrees of freedom of the *F-test*.

R^2

This is the amount that R^2 is reduced when this term is removed from the regression model.

Sum of Squares

This is the amount that the model sum of squares that are reduced when this term is removed from the model.

Mean Square

The mean square is the sum of squares divided by the degrees of freedom.

F-Ratio

This is the *F*-statistic for testing the null hypothesis that all β_i associated with this term are zero. This *F-statistic* has *DF* and *N-p*-1 degrees of freedom.

P-Value

This is the *p*-value for the above *F*-test. The *p*-value is the probability that the test statistic will take on a value at least as extreme as the observed value, assuming that the null hypothesis is true. If the *p*-value is less than α, say 0.05, the null hypothesis is rejected. If the *p*-value is greater than α, the null hypothesis is accepted.

R^2 Reports

R² for Model: Y = X M

Independent Variable (IV)	Total R² for this IV and IV's Above	Increase in R² if this IV is Included with IV's Above	Decrease in R² if this IV was Removed	R² if this IV was Fit Alone	Partial R² if Adjusted for All Other IV's
Temp	0.5475	0.5475	0.3316	0.5475	0.4233
Thirst	0.5482	0.0007	0.0007	0.2166	0.0016

R² for Model: Y = X

Independent Variable (IV)	Total R² for this IV and IV's Above	Increase in R² if this IV is Included with IV's Above	Decrease in R² if this IV was Removed	R² if this IV was Fit Alone	Partial R² if Adjusted for All Other IV's
Temp	0.5475	0.5475	0.5475	0.5475	0.5475

R² for Model: M = X

Independent Variable (IV)	Total R² for this IV and IV's Above	Increase in R² if this IV is Included with IV's Above	Decrease in R² if this IV was Removed	R² if this IV was Fit Alone	Partial R² if Adjusted for All Other IV's
Temp	0.4311	0.4311	0.4311	0.4311	0.4311

R^2 reflects the percent of variation in *Y* explained by the independent variables in the model. A value of R^2 near zero indicates a complete lack of fit between *Y* and the *Xs,* while a value near one indicates a perfect fit.

In this section, various types of R^2 values are given for each regression to provide insight into the variation in the dependent variable explained either by the independent variables added in order (i.e., sequential) or by the independent variables added last. This information is valuable in an analysis of which variables are most important.

Independent Variable (IV)

This is the name of the independent variable reported on in this row.

Total R^2 for This IV and Those Above

This is the R^2 value that would result from fitting a regression with this independent variable and those listed above it. The IV's below it are ignored.

Increase in R^2 if this IV is Included with IV's Above

This is the amount that this IV adds to R^2 when it is added to a regression model that includes those IV's listed above it in the report.

Decrease in R^2 if this IV was Removed

This is the amount that R^2 would be reduced if this IV were removed from the model. Large values here indicate important independent variables, while small values indicate insignificant variables.

One of the main problems in interpreting these values is that each assumes all other variables are already in the equation. This means that if two variables both represent the same underlying information, they will each seem to be insignificant after considering the other. If you remove both, you will lose the information that either one could have brought to the model.

R^2 if This IV was Fit Alone

This is the R^2 that would be obtained if the dependent variable were only regressed against this one independent variable. Of course, a large R^2 value here indicates an important independent variable that can stand alone.

Partial R^2 if Adjusted for All Other IV's

The is the square of the partial correlation coefficient. The partial R^2 reflects the percent of variation in the dependent variable explained by one independent variable controlling for the effects of the rest of the independent variables. Large values for this partial R^2 indicate important independent variables.

Multicollinearity

Multicollinearity for Model: Y = X M

Independent Variable	Variance Inflation Factor	R² versus Other IV's	Tolerance	Diagonal of X'X Inverse
Temp	1.7576	0.4311	0.5689	8.067686E-05
Thirst	1.7576	0.4311	0.5689	0.004443419

Multicollinearity for Model: Y = X

Independent Variable	Variance Inflation Factor	R² versus Other IV's	Tolerance	Diagonal of X'X Inverse
Temp	1.0000	0.0000	1.0000	4.59007E-05

Multicollinearity for Model: M = X

Independent Variable	Variance Inflation Factor	R² versus Other IV's	Tolerance	Diagonal of X'X Inverse
Temp	1.0000	0.0000	1.0000	4.59007E-05

These reports provide information useful in assessing the amount of multicollinearity in each regression. The last two reports are only useful when covariates are included in the analysis.

Variance Inflation Factor

The variance inflation factor (*VIF*) is a measure of multicollinearity. It is the reciprocal of $1 - R_X^2$, where R_X^2 is the R^2 obtained when this variable is regressed on the remaining IV's. A *VIF* of 10 or more for large data sets indicates a collinearity problem since the R_X^2 with the remaining *IV*'s is 90 percent. For small data sets, even *VIF's* of 5 or more can signify collinearity. Variables with a high *VIF* are candidates for exclusion from the model.

$$VIF_i = \frac{1}{1 - R_i^2}$$

R² versus Other IV's

R_X^2 is the R^2 obtained when this variable is regressed on the remaining independent variables. A high R_X^2 indicates a lot of overlap in explaining the variation among the remaining independent variables.

Tolerance

Tolerance is just $1 - R_X^2$, the denominator of the variance inflation factor.

Diagonal of X'X Inverse

The *X'X* inverse is an important matrix in regression. This is the j^{th} row and j^{th} column element of this matrix.

Y, X, M, and Residuals

Y, X, M, and Residuals

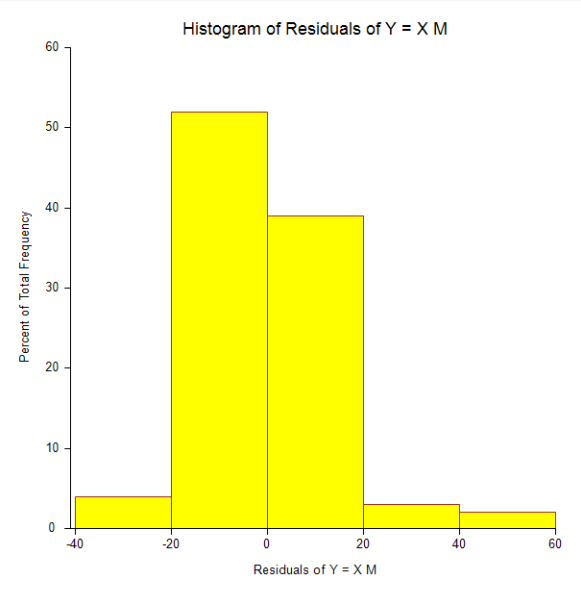
Row	Water Y	Temp X	Thirst M	Residuals Y = X M	Residuals Y = X	Residuals M = X
1	84	68	9	23.5495	23.2497	1.0524
2	33	36	5	-2.4286	-2.3953	-0.1167
3	31	34	4	-3.0784	-2.8106	-0.9398
4	21	20	3	-1.9176	-1.7178	-0.7012
5	51	53	7	2.2429	2.1348	0.3794
6	41	45	6	-1.5016	-1.5264	0.0871
7	34	33	4	0.7392	0.9817	-0.8513
8	57	62	8	1.1699	1.0037	0.5832
9	58	66	8	-1.1003	-1.1656	0.2293
10	23	69	9	-38.2681	-38.5427	0.9639
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

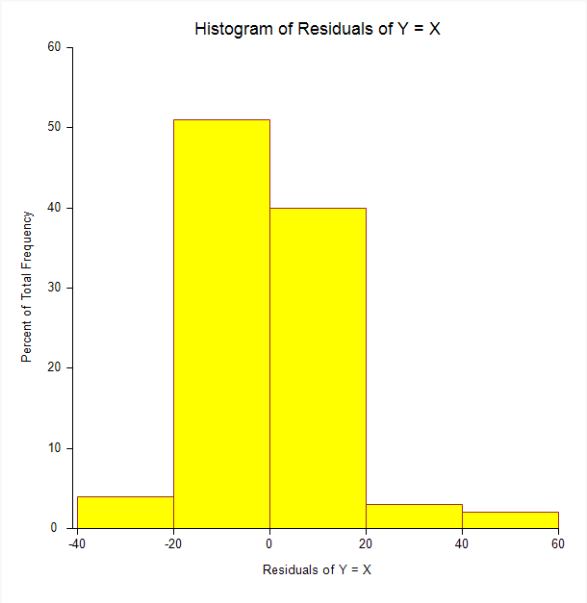
This report gives the values of Y, X, and M for each row followed by the residuals from the three regression models. It allows you to quickly see any rows with large residuals in at least one of the regressions.

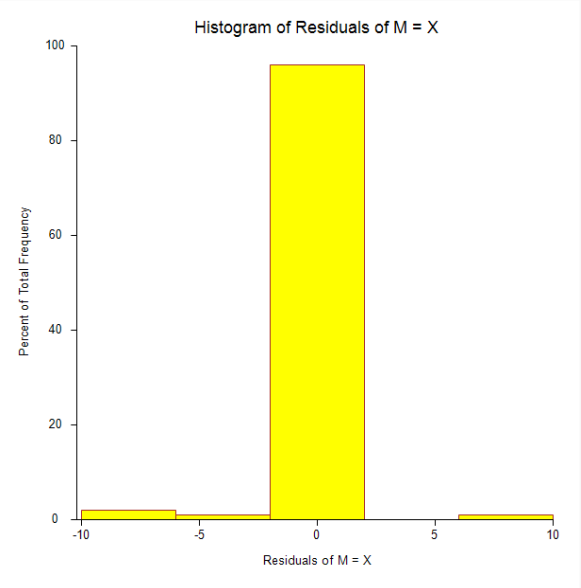
Histograms of Residuals

The purpose of these histograms 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 would be the normal probability plot.

Histograms of Residuals





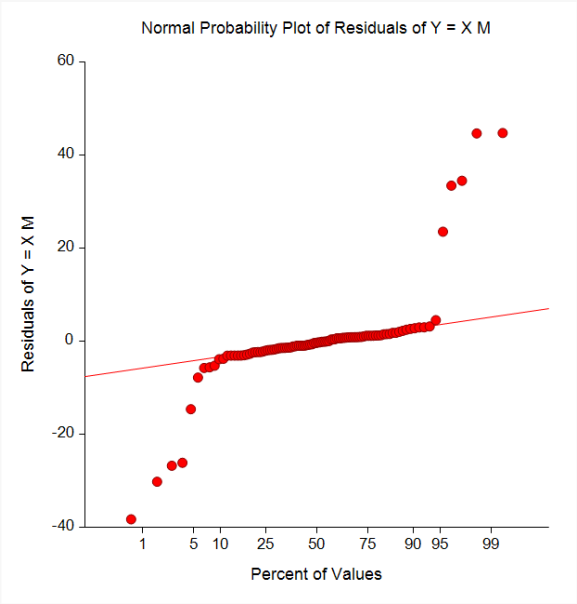


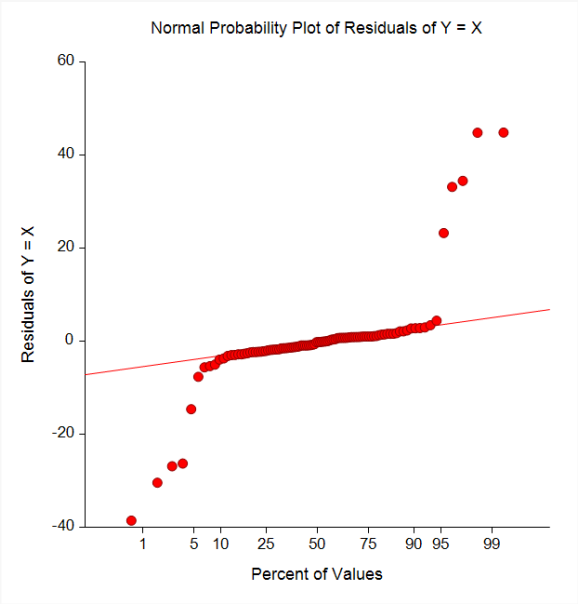
Probability Plots of Residuals

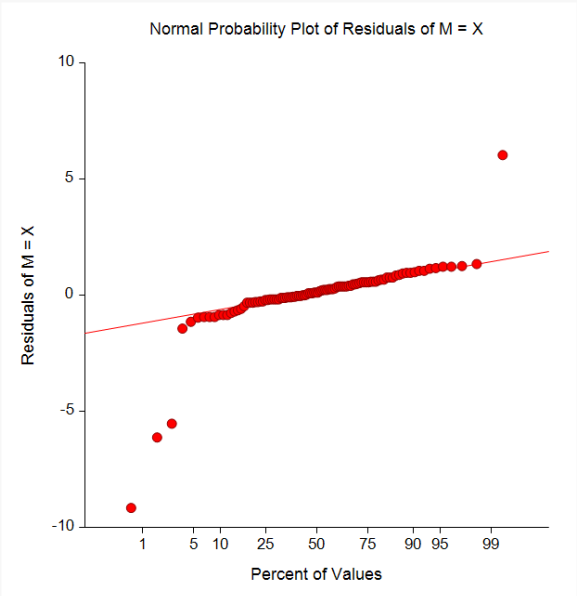
If the residuals are normally distributed, the data points of the normal probability plot will fall along a straight line through the origin with a slope of 1.0. Major deviations from this ideal picture reflect departures from normality. Stragglers at either end of the normal probability plot indicate 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.

If the residuals are not normally distributed, then the t-tests on regression coefficients and any interval estimates are not valid. This is a critical assumption to check.

Probability Plots of Residuals



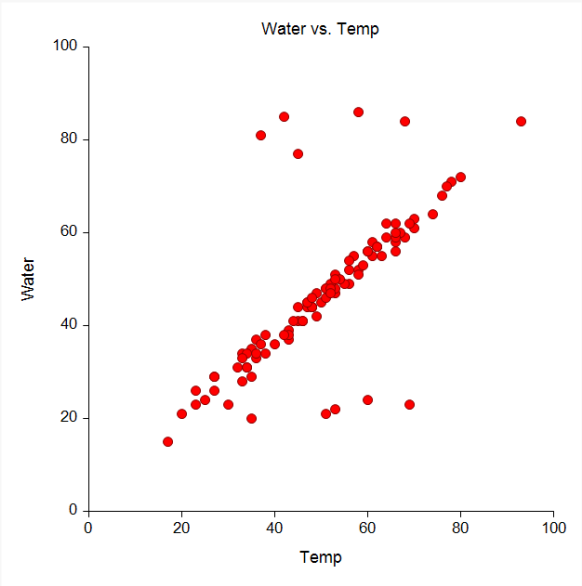


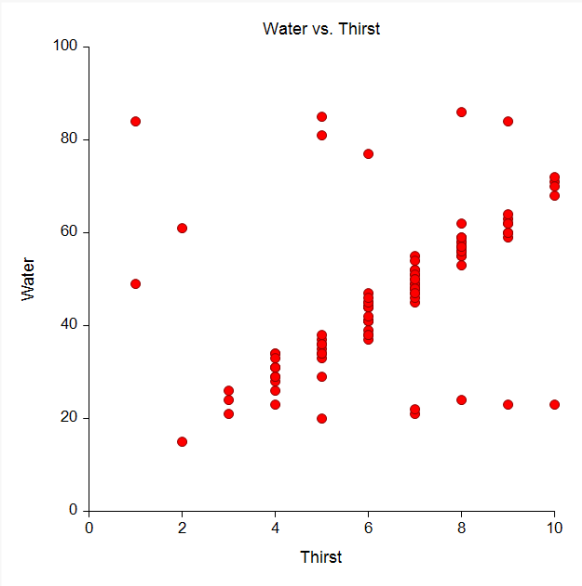


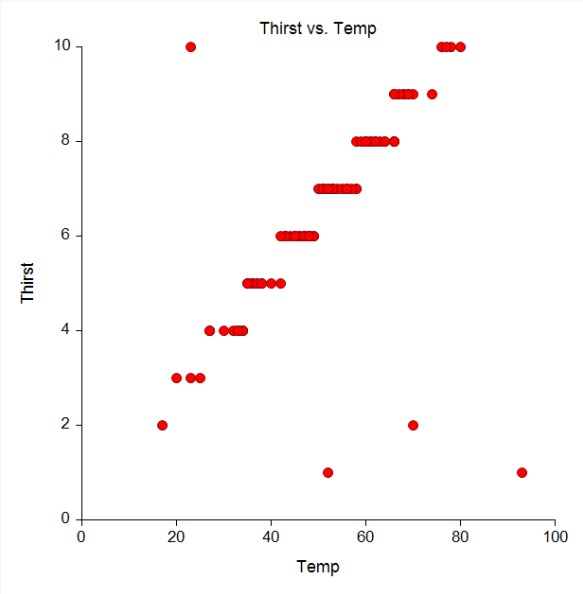
Scatter Plots of Y versus each Independent Variable

Actually, a regression analysis should always begin with a plot of Y versus each independent variable. These plots often show outliers, curvilinear relationships, and other anomalies.

Scatter Plots



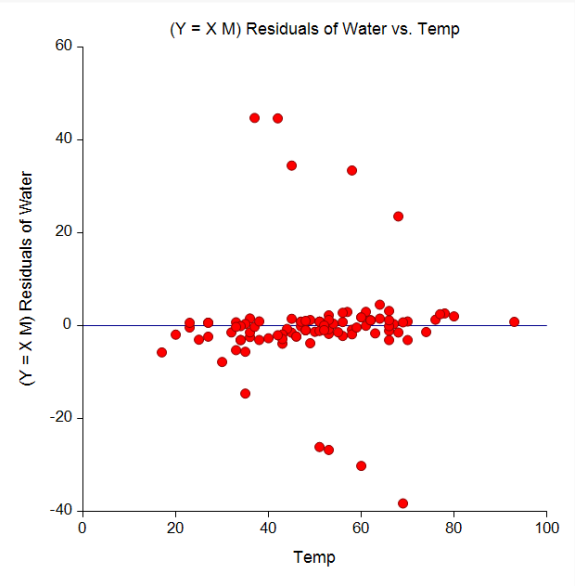


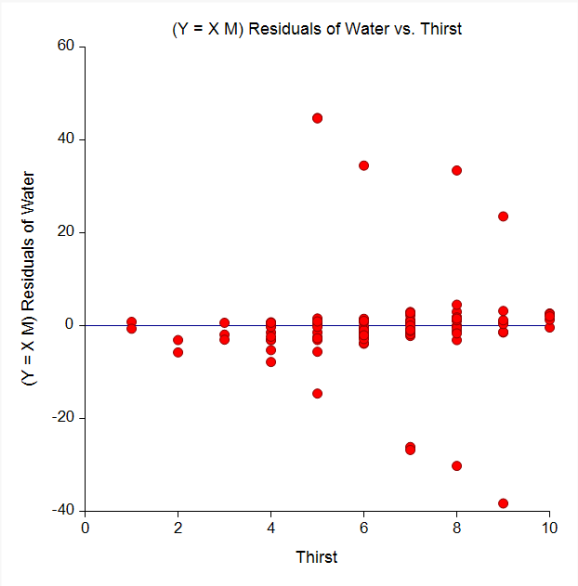


Scatter Plots of Residuals versus each Independent Variable

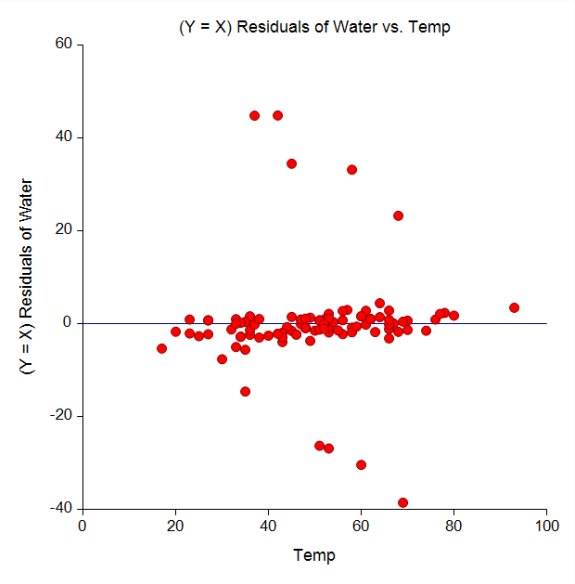
No regression analysis is complete without viewing the residuals plotted against each independent variable. These plots often show outliers, curvilinear relationships, and other anomalies.

(Y = X M) Residual Plots

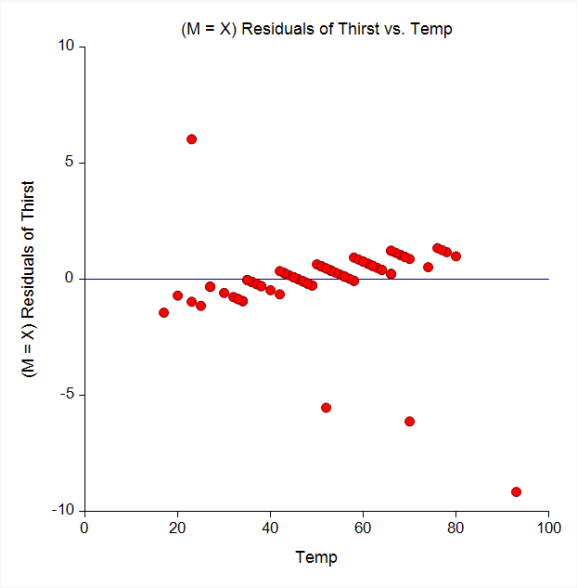




(Y = X) Residual Plots



(M = X) Residual Plots



These plots show the presence of a view outliers in each plot. This suggests the robust regression could be useful in this case.

Example 2 – Mediation Analysis (Robust Regression Solution)

This section presents an example of how to run a mediation analysis using robust regression. The residual plots in Example 1 showed the presence of outliers in the data. This suggests that the data in Example 1 should be reanalyzed using robust regression.

The data are in the *Mediation* dataset. In this example, it is supposed that the amount of water consumption (Y) is directly related to the temperature (X). The mediator is an index of how thirsty each subject was. This mediator is contained in the column named Thirst. Thus, for this example set X = Temp, M = Thirst, and Y = Water.

Setup

To run this example, complete the following steps:

1 Open the Mediation example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Mediation** and click **OK**.

2 Specify the Mediation Analysis procedure options

- Find and open the **Mediation Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Y (Dependent Variable)	**Water**
X (Independent Variable)	**Temp**
M (Mediator Variable)	**Thirst**
Terms	**1-Way**

Robust, Bootstrap Tab	
Estimation Method	**Robust Regression using Huber's Method**
Type of Weights Assumed	**Random**
Bootstrap Calculations	**Checked**
Samples	**300**

Reports Tab	
Run Summary	**Checked**
Mediation Effects	**Checked**
Individual Regressions	**Checked**
Robust Iterations - Coefficients	**Checked**
Y, X, M, and Weights	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Y (Dependent Variable)	Water	Number Processed	100
X (Independent Variable)	Temp	Number Used in Estimation	100
M (Mediator Variable)	Thirst	Number Filtered Out	0
Number of Covariates	0	Number with X's Missing	0
Weight Variable	None	Number with Weight Missing	0
Robust Method	Huber's Method	Number with Y Missing	0
Tuning Constant	1.345	Sum of Robust Weights	89.561
MAD Scale Factor	0.675		
Bootstrap Samples	300		
Bootstrap Random Seed	4835559		
Bootstrap C.L. Type	Reflection		

Run Summary Details

Regression Model	R^2	Robust Iterations	Max % Change in Any Robust Coefficient	S from MAD	S from MSE	Completion Status
Y = X M	0.9225	8	0.000	2.225856	3.55961	Normal Completion
Y = X	0.9220	7	0.000	2.247608	3.556176	Normal Completion
M = X	0.9176	7	0.000	0.4067543	0.5114287	Normal Completion

These reports summarize the mediation analysis results. They present the estimation method used, the variables used, the number of rows used, and the R^2 of each of the three models.

To allow us to compare the two analyses, the Run Summary Details report from Example 1 is repeated here.

Run Summary Details

Regression Model	R^2	S from MSE	Completion Status
Y = X M	0.5482	10.78697	Normal Completion
Y = X	0.5475	10.74047	Normal Completion
M = X	0.4311	1.515404	Normal Completion

By comparing these two reports, we notice what the robust regression option has done. The R^2 values have increased from about 0.55 to 0.92. A large change. Also, S from MSE has been reduced from 10.8 to 3.6. Again, a large change.

Note that the Sum of Robust Weights has decreased from 100 to 89.6. This gives us a view of what robust regression has done. It has more or less omitted the 10 rows that didn't fit well. It is as if these rows were deleted from the dataset and then the analysis using ordinary least squares is rerun on the remaining 90 rows.

Direct, Indirect, and Total Effects

Direct, Indirect, and Total Effects

Estimation:	Huber robust regression
Sb(i):	Random row weights, adjust Sb(i)
Sb(Indirect):	First-order
Bootstrap:	Number of samples = 300, Random seed = 4835559, Confidence limit type = Reflection
Y:	Water
X:	Temp
M:	Thirst
Covariates:	0

			Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Test Statistic	P-Value	Lower	Upper
Total	0.8516012	0.01545002	55.120	0.0000	0.8209411	0.8822612
Direct (X → Y)	0.8532954	0.02053553	41.552	0.0000	0.8125381	0.8940527
Indirect (X → M → Y)						
Normal Theory	-0.002583202	0.0186791	-0.138	0.8900	-0.03919356	0.03402716
Bootstrap	-0.002583202	0.01974407	-0.131	0.8959	-0.04461664	0.0190477

This report shows the main results of the mediation analysis, this time using the three robust regressions.

So that we can compare the items, we are repeating this report from Example 1.

Direct, Indirect, and Total Effects

Estimation:	Ordinary least squares
Sb(Indirect):	First-order
Bootstrap:	Number of samples = 3000, Random seed = 4655484, Confidence limit type = Reflection
Y:	Water
X:	Temp
M:	Thirst
Covariates:	0

			Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Test Statistic	P-Value	Lower	Upper
Total	0.7923434	0.07276677	10.889	0.0000	0.6479401	0.9367467
Direct (X → Y)	0.8175452	0.0968889	8.438	0.0000	0.6252475	1.009843
Indirect (X → M → Y)						
Normal Theory	-0.02520181	0.06367939	-0.396	0.6923	-0.1500111	0.0996075
Bootstrap	-0.02520181	0.07635163	-0.330	0.7413	-0.09416629	0.08528873

Now we can see that the total and direct regression coefficients have changed only a little. However, the indirect has increased from -0.025 to -0.0026, quite a change.

Example 3 – Mediation Analysis (Adding Covariates)

This section presents an example of how to run a mediation analysis using robust regression. The residual plots in Example 1 showed the presence of outliers in the data. This suggests that the data in Example 1 should be reanalyzed using robust regression.

The data are in the *Mediation* dataset. In this example, it is supposed that the amount of water consumption (Y) is directly related to the temperature (X). The mediator is an index of how thirsty each subject was. This mediator is contained in the column named Thirst. Thus, for this example set X = Temp, M = Thirst, and Y = Water.

Setup

To run this example, complete the following steps:

1 Open the Mediation example dataset

- From the File menu of the NCSS Data window, select **Open Example Data**.
- Select **Mediation** and click **OK**.

2 Specify the Mediation Analysis procedure options

- Find and open the **Mediation Analysis** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab	
Y (Dependent Variable)	**Water**
X (Independent Variable)	**Temp**
M (Mediator Variable)	**Thirst**
C (Numeric Covariates)	**Age**
C (Categorical Covariates)	**Adults**
Default Recoding Scheme	**Compare Each with Next**
Terms	**1-Way**

Robust, Bootstrap Tab	
Estimation Method	**Ordinary Least Squares**
Bootstrap Calculations	**Checked**
Samples	**300**

Reports Tab	
Run Summary	**Checked**
Mediation Effects	**Checked**
Individual Regressions	**Checked**
Robust Iterations - Coefficients	**Checked**
Y, X, M, and Weights	**Checked**

3 Run the procedure

- Click the **Run** button to perform the calculations and generate the output.

Run Summary

Run Summary

Item	Value	Rows	Value
Y (Dependent Variable)	Water	Number Processed	100
X (Independent Variable)	Temp	Number Used in Estimation	100
M (Mediator Variable)	Thirst	Number Filtered Out	0
Number of Covariates	2	Number with X's Missing	0
Weight Variable	None	Number with Weight Missing	0
Robust Method	None (OLS)	Number with Y Missing	0
Bootstrap Samples	300	Sum of Robust Weights	100
Bootstrap Random Seed	4902628		
Bootstrap C.L. Type	Reflection		

Run Summary Details

Regression Model	R^2	S from MSE	Completion Status
Y = X M C	0.5836	10.52012	Normal Completion
Y = X C	0.5828	10.47447	Normal Completion
M = X C	0.4378	1.529981	Normal Completion

These reports summarize the mediation analysis results. They present the estimation method used, the variables used, the number of rows used, and the R^2 of each of the three models. Note that the R^2 values have not changed much, indicating that the covariates where not useful in this case.

Direct, Indirect, and Total Effects

Direct, Indirect, and Total Effects

Estimation:	Ordinary least squares
Sb(Indirect):	First-order
Bootstrap:	Number of samples = 300, Random seed = 4902628, Confidence limit type = Reflection
Y:	Water
X:	Temp
M:	Thirst
Covariates:	2

			Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Test Statistic	P-Value	Lower	Upper
Total	0.7944282	0.07240131	10.973	0.0000	0.6506934	0.9381629
Direct (X → Y)	0.8200662	0.09484039	8.647	0.0000	0.6317584	1.008374
Indirect (X → M → Y)						
Normal Theory	-0.02563802	0.0609658	-0.421	0.6741	-0.1451288	0.09385276
Bootstrap	-0.02563802	0.09566278	-0.268	0.7887	-0.1028662	0.09393113

This report shows the main results of the mediation analysis, this time using the three robust regressions.

So that we can compare the items, we are repeating this report from Example 1.

Direct, Indirect, and Total Effects

Estimation:	Ordinary least squares
Sb(Indirect):	First-order
Bootstrap:	Number of samples = 3000, Random seed = 4655484, Confidence limit type = Reflection
Y:	Water
X:	Temp
M:	Thirst
Covariates:	0

			Test of H0: β(i) = 0		95% Confidence Interval Limits for β(i)	
Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Test Statistic	P-Value	Lower	Upper
Total	0.7923434	0.07276677	10.889	0.0000	0.6479401	0.9367467
Direct (X → Y)	0.8175452	0.0968889	8.438	0.0000	0.6252475	1.009843
Indirect (X → M → Y)						
Normal Theory	-0.02520181	0.06367939	-0.396	0.6923	-0.1500111	0.0996075
Bootstrap	-0.02520181	0.07635163	-0.330	0.7413	-0.09416629	0.08528873

Now we can see that, in this case, adding the covariates has not changed the regression coefficients a great deal.