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.

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

© NCSS, LLC. All Rights Reserved.

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.

© NCSS, LLC. All Rights Reserved.

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

   $$\phi(u) = 1 \quad |u| \le c$$
   $$\phi(u) = 0 \quad |u| > c$$

   for Tukey's biweight estimation

   $$\phi(u) = \left[1 - \frac{u^2}{c^2}\right]\left[1 - 5\frac{u^2}{c^2}\right] \quad |u| \le c$$
   $$\phi(u) = 0 \quad |u| > c$$

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 $\tilde{\mathbf{Y}}$ on $\mathbf{X}$. 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.

© NCSS, LLC. All Rights Reserved.

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 regression. This is often call *row-wise deletion.*

Procedure Options

This section describes the options available in this procedure.

Variables, Model Tab

This panel specifies the variables and model used in the analysis.

Variables

Y (Dependent Variable)

Specify the column containing the values of the dependent (Y) variable.

Y, also known as the outcome, response, or predicted variable must contain only numeric values.

In the mediation model, the independent variable X is thought to impact the dependent variable Y through a mediator variable M.

X (Independent Variable)

Specify the column containing the values of the independent (X) variable.

In the mediation model, the independent variable X is thought to impact the dependent variable Y through a mediator variable M.

X must contain only numeric values. If X is binary, it must be coded with numeric value such as 0 and 1 or 1 and 2.

© NCSS, LLC. All Rights Reserved.

M (Mediator Variable)

Specify the column containing the values of the mediator (M) variable.

In the mediation model, the independent variable X is thought to impact the dependent variable Y through a mediator variable M.

M must contain only numeric values. If M is binary, it must be coded with numeric value such as 0 and 1 or 1 and 2.

Weight Variable

The weight variable contains the (non-negative) weight given to each observation in regression calculations. By default, each observation receives an equal weight of 1/n (where n is the sample size). This variable allows you to specify different weights for different observations.

The weight variable is commonly created from the three weights generated and saved during the robust regression estimation of the three equations. For example, you might use the product or the minimum of the three stored weights as the weight.

The weight variable is used for all three regression models.

NCSS automatically scales the weights so they sum to one. Hence, you can enter integer numbers and NCSS will scale them to appropriate fractions.

Covariates

C (Numeric Covariates)

Specify any optional quantitative covariate columns. These covariates will be used in each of the three regression equations that are fit.

Numeric (Quantitative) Variables

We consider a variable as 'numeric' if its values are numbers that are at least ordinal. Nominal variables are classified as categorical, even if their values are numbers.

Although you can specify binary-indicator (0-1) variables here, it is often better to specify them as categorical variables.

Powers and Cross-Products

You can automatically generate additional covariates as powers and/or cross-products of existing covariates as internal variables that only exist at run-time. This is done using the 'Custom Model' box shown below when the Terms option is set to 'Custom Model.' Of course, you can also add these power and cross-product variables to the database using the transformation feature.

C (Categorical Covariates)

Specify any optional categorical covariates here. Additional covariates may be specified as interactions in the Custom Model below.

Regression analysis is only defined for numeric variables that are at least ordinal. Since categorical variables are nominal, they cannot be used directly in regression. Instead, an internal set of numeric variables must be substituted for each categorical covariate.

A categorical variable only takes on a few unique values which identify categories. For example, state of birth, hair color, and type of disease are categorical variables.

© NCSS, LLC. All Rights Reserved.

Recoding Categories to Numeric Values

NCSS automatically generates internal numeric variables from categorical covariates since only numeric values can be processed by multiple regression. One of the strengths of NCSS is the ease with which these new variables are generated.

The complete syntax for specifying a categorical variable is *VarName*(*CType; RefValue*) where *VarName* is the name of the variable, CType is the recoding scheme, and RefValue is the reference value, if needed.

CType

The recoding scheme is entered as a letter. Possible choices are B, P, R, N, S, L, F, A, 1, 2, 3, 4, 5, or E. The meaning of each of these letters is as follows.

- **B** for **binary** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	B1	B2	B3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

- **P** for **Polynomial** of up to 5th order (you cannot use this option with category variables with more than 6 categories.
  Example: Categorical variable Z with 4 categories.

Z	P1	P2	P3
1	-3	1	-1
3	-1	-1	3
5	1	-1	-3
7	3	1	1

- **R** to compare each with the **reference value** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **N** to compare each with the **next** category.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	0	0
3	-1	1	0
5	0	-1	1
7	0	0	-1

- **S** to compare each with the **average of all subsequent** values.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-3	0	0
3	1	-2	0
5	1	1	-1
7	1	1	1

© NCSS, LLC. All Rights Reserved.

- **L** to compare each with the **prior** category.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-1	0	0
3	1	-1	0
5	0	1	-1
7	0	0	1

- **F** to compare each with the **average of all prior** categories.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	1	1
3	1	1	-1
5	1	-2	0
7	-3	0	0

- **A** to compare each with the **average of all** categories (the Reference Value is skipped).
  Example: Categorical variable Z with 4 categories. Suppose the reference value is 3.

Z	S1	S2	S3
1	-3	1	1
3	1	1	1
5	1	-3	1
7	1	1	-3

- **1** to compare each with the **first** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	-1	-1	-1
B	1	0	0
C	0	1	0
D	0	0	1

- **2** to compare each with the **second** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	-1	-1	-1
C	0	1	0
D	0	0	1

- **3** to compare each with the **third** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	-1	-1	-1
D	0	0	1

© NCSS, LLC. All Rights Reserved.

- **4** to compare each with the **fourth** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **5** to compare each with the **fifth** category after sorting.
  Example: Categorical variable Z with 5 categories.

Z	C1	C2	C3	C4
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1
E	-1	-1	-1	-1

- **E** to compare each with the **last** category after sorting.
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

RefValue
A second, optional argument is the reference value. The reference value is one of the categories. The other categories are compared to it, so it is usually a baseline or control value. If neither a baseline or control value is evident, the reference value is the most frequent value.

For example, suppose you want to include a categorical independent variable, State, which has four values: Texas, California, Florida, and NewYork. Suppose the recoding scheme is specified as *Compare Each with Reference Value* with the reference value of *California*. You would enter

State(R;California)

Default Recoding Scheme

Select the default type of numeric variable that will be generated when processing categorical independent variables. The values in a categorical variable are not used directly in regression analysis. Instead, a set of numeric variables is automatically created and substituted for them. This option allows you to specify what type of numeric variable will be created. The options are outlined in the sections below.

The contrast type may also be designated within parentheses after the name of each categorical independent variable, in which case the default contrast type is ignored.

If your model includes interactions of categorical variables, this option should be set to ‘Contrast with Reference’or Compare with All Subsequent' in order to match GLM results for factor effects.

- **Binary** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	B1	B2	B3
A	1	0	0
B	0	1	0
C	0	0	1
D	0	0	0

© NCSS, LLC. All Rights Reserved.

- **Polynomial** of up to 5th order (you cannot use this option with category variables with more than 6 categories.
  Example: Categorical variable Z with 4 categories.

Z	P1	P2	P3
1	-3	1	-1
3	-1	-1	3
5	1	-1	-3
7	3	1	1

- **Compare Each with Reference Value** (the group with the reference value is skipped).
  Example: Categorical variable Z with 4 categories. Category D is the reference value.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **Compare Each with Next**.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	0	0
3	-1	1	0
5	0	-1	1
7	0	0	-1

- **Compare Each with All Subsequent**.
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-3	0	0
3	1	-2	0
5	1	1	-1
7	1	1	1

- **Compare Each with Prior**
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	-1	0	0
3	1	-1	0
5	0	1	-1
7	0	0	1

- **Compare Each with All Prior**
  Example: Categorical variable Z with 4 categories.

Z	S1	S2	S3
1	1	1	1
3	1	1	-1
5	1	-2	0
7	-3	0	0

© NCSS, LLC. All Rights Reserved.

- **Compare Each with Average**
  Example: Categorical variable Z with 4 categories. Suppose the reference value is 3.

Z	S1	S2	S3
1	-3	1	1
3	1	1	1
5	1	-3	1
7	1	1	-3

- **Compare Each with First**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	-1	-1	-1
B	1	0	0
C	0	1	0
D	0	0	1

- **Compare Each with Second**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	-1	-1	-1
C	0	1	0
D	0	0	1

- **Compare Each with Third**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	-1	-1	-1
D	0	0	1

- **Compare Each with Fourth**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

- **Compare Each with Fifth**
  Example: Categorical variable Z with 5 categories.

Z	C1	C2	C3	C4
A	1	0	0	0
B	0	1	0	0
C	0	0	1	0
D	0	0	0	1
E	-1	-1	-1	-1

© NCSS, LLC. All Rights Reserved.

- **Compare Each with Last**
  Example: Categorical variable Z with 4 categories.

Z	C1	C2	C3
A	1	0	0
B	0	1	0
C	0	0	1
D	-1	-1	-1

Default Reference Value

This option specifies the default reference value to be used when automatically generating indicator variables during the processing of selected categorical independent variables. The reference value is often the baseline, and the other values are compared to it. The choices are

- **First Value after Sorting – Fifth Value after Sorting**
  Use the first (through fifth) value in alpha-numeric sorted order as the reference value.

- **Last Value after Sorting**
  Use the last value in alpha-numeric sorted order as the reference value.

Regression Model of Covariates

These options control which terms are included in the covariate portion of the three regression models. Note that the resulting model does NOT include the X or M variables. They are added automatically and you do not specify them here.

Terms

This option specifies which terms (terms, powers, cross-products, and interactions) are included in the regression model. For a straight-forward regression model, select *1-Way*.

The options are

- **1-Way**
  All numeric and categorical covariates are included in the model. No interaction or power terms are included. Use this option when you just want to use the covariates you have specified.

  This is the option to select when you want to analyze the covariates specified without adding any other terms.

  For example, if you have three covariates A, B, and C, this would generate the model:

  A + B + C

- **Up to 2-Way**
  This option specifies that all individual covariates, two-way interactions, and squares of numeric covariates are included in the model. For example, if you have three numeric covariates A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*A + B*B + C*C

  On the other hand, if you have three categorical covariates A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C

© NCSS, LLC. All Rights Reserved.

- **Up to 3-Way**
  All individual covariates, two-way interactions, three-way interactions, squares of numeric covariates, and cubes of numeric covariates are included in the model. For example, if you have three numeric, covariates A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C + A*A + B*B + C*C + A*A*B + A*A*C + B*B*C +A*C*C + B*C*C

  On the other hand, if you have three categorical covariates A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

- **Up to 4-Way**
  All individual covariates, two-way interactions, three-way interactions, and four-way interactions are included in the model. Also included would be squares, cubes, and quartics of numeric covariates and their cross-products.

  For example, if you have four categorical covariates A, B, C, and D, this would generate the model:

  A + B + C + D + A*B + A*C + A*D + B*C + B*D + C*D + A*B*C + A*B*D + A*C*D + B*C*D + A*B*C*D

- **Interaction**
  Mainly used for categorical covariates. A saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-covariate combinations (you can have unequal numbers of observations for each combination of the categorical covariates). No squares, cubes, etc. are generated.

  For example, if you have three covariates A, B, and C, this would generate the model:

  A + B + C + A*B + A*C + B*C + A*B*C

  Note that the discussion of the Custom option discusses the interpretation of this model.

- **Custom**
  The model specified in the *Custom* box is used.

Replace Custom with Preview Model (button)

When this button is pressed, the Custom Model is cleared and a copy of the Preview model is stored in the Custom Model. You can then edit this Custom Model as desired.

Maximum Order of Custom Terms

This option specifies that maximum number of variables that can occur in an interaction (or cross-product) term in a custom model. For example, A*B*C is a third order interaction term and if this option were set to 2, the A*B*C term would not be included in the model.

This option is particularly useful when used with the bar notation of a custom model to allow a simple way to remove unwanted high-order interactions.

Custom

This option specifies a custom model. It is only used when the *Terms* option is set to *Custom*. This specifies that terms (single variables, cross-products, and interactions) that are to be kept in the model.

Interactions
An interaction expresses the combined relationship between two or more covariates and the dependent variable by creating a new covariate that is the product of the covariates. The interaction (cross-product) between two numeric covariates is generated by multiplying them. The interaction between to categorical covariates is generated by multiplying each pair of internal variables. The interaction between a numeric covariates and a

© NCSS, LLC. All Rights Reserved.

categorical covariates is created by generating all products between the numeric covariates and the generated, numeric variables.

Syntax

A model is written by listing one or more terms. The terms are separated by a blank or plus sign. Terms include variables and interactions. Specify regular variables (main effects) by entering the variable names. Specify interactions by listing each variable in the interaction separated by an asterisk (*), such as Fruit*Nuts or A*B*C.

You can use the bar (|) symbol as a shorthand technique for specifying many interactions quickly. When several variables are separated by bars, all of their interactions are generated. For example, A|B|C is interpreted as A + B + C + A*B + A*C + B*C + A*B*C.

You can use parentheses. For example, A*(B+C) is interpreted as A*B + A*C.

Some examples will help to indicate how the model syntax works:

A|B = A + B + A*B

A|B A*A B*B = A + B + A*B + A*A + B*B

Note that you should only repeat numeric variables. That is, A*A is valid for a numeric variable, but not for a categorical variable.

A|A|B|B (Max Term Order=2) = A + B + A*A + A*B + B*B

A|B|C = A + B + C + A*B + A*C + B*C + A*B*C

(A + B)*(C + D) = A*C + A*D + B*C + B*D

(A + B)|C = (A + B) + C + (A + B)*C = A + B + C + A*C + B*C

Robust, Bootstrap Tab

The options on this panel control the robust regression and bootstrap options.

Estimation Algorithm

Estimation Method

This option specifies the method used to estimate the three regression equations. Ordinary least squares is usually used to estimate the regression equations. However, the estimates are easily distorted by outliers.

The robust regression algorithms seek to reduce the influence of observations that are apparent outliers. Both robust methods are M-Estimators and use iteratively reweighted least squares.

- **Ordinary Least Squares**
  This option specifies that ordinary least squares regression should be used. No robust regression is attempted.

- **Robust Regression using Huber's Method**
  This option specifies that robust regression is to be run using Huber's method. This M-estimator gradually reduces the weight of observations with large residuals. However, no observations have a weight of zero. Also, it works with poor starting values, and it has better convergence properties.

- **Robust Regression using Tukey's Biweight**
  This option specifies that robust regression is to be run using Tukey's method. This M-estimator completely down weights observation with large outliers until their weight is set to zero. It provides the most protection against heavy-tailed error distributions.

© NCSS, LLC. All Rights Reserved.

Recommended
Huber's Method

Note
A common strategy is to use a robust algorithm to find which rows of data are associated with outliers. Then remove these rows and rerun the analysis using Ordinary Least Squares.

Huber's Tuning Constant
This option specifies the robust truncation constant for Huber's method. This is a cutoff point on the influence function designating when an observation's weight should be reduced.

The recommended value is 1.345.

Tukey's Tuning Constant
This option specifies the robust truncation constant for Tukey's Biweight method. This is a cutoff point on the influence function designating when an observation's weight should be set to zero.

The recommended value is 4.685.

Scale Factor

MAD Scale Factor
Specify the constant used to scale MAD. The default value of 0.6745 is suggested in several regression texts because it is appropriate for the Huber method when normal errors are assumed.

Stop Iterating When

Maximum Percent Change in Beta Estimates is Less Than or Equal To
This option specifies an early stopping value for the iteration procedure. Normally, the number of iterations is specified in the next option. However, if the percentage change in each of the estimated regression coefficients is less than this amount, the iteration procedure is terminated. If you want this option to be ignored, set it to zero.

We recommend setting this value to 0.001 and the *Number of Iterations* to 30.

Number of Iterations is
Specifies the maximum number of iterations allowed while finding a solution. If this number is reached, the procedure is terminated.

We recommend setting this value to 30 and the *Maximum Percent Change in Beta Estimates* to 0.001.

Adjustment of Standard Errors of Beta Estimates for Random Weights

Type of Weights Assumed
Specify whether the standard errors of the regression coefficients are to be adjusted for the random weights that are produced by robust regression.

- **Fixed (No Adjustment)**
  The standard error formulas assume that the robust regression weights are fixed constants known before running the analysis. No adjustment is made to account for the randomness in the regression weights.

© NCSS, LLC. All Rights Reserved.

- **Random**

  The standard error formulas assume that the robust regression weights are random values not known before running the analysis. Appropriate adjustment of the regression coefficient standard errors is made to account for this randomness.

Bootstrap Options

Bootstrap Calculations

Specify whether to calculate the bootstrap confidence interval of the Indirect Effect.

Note that this option uses Monte Carlo simulation and may require a long time to complete, especially for robust regression.

Samples

This is the number of bootstrap samples used. A general rule of thumb is that you use at least 100 when standard errors are your focus or at least 1000 when confidence intervals are your focus. If computing time is available, it does not hurt to do 4000 or 5000.

We recommend setting this value to at least 3000.

Retries

If the results from a bootstrap sample cannot be calculated, the sample is discarded and a new sample is drawn in its place. This parameter is the number of times that a new sample is drawn before the algorithm is terminated. We recommend setting the parameter to at least 50.

Percentile Type

The method used to create the percentiles when forming bootstrap confidence limits. You can read more about the various types of percentiles in the Descriptive Statistics chapter. We suggest you use the Ave X(p[n+1]) option.

C.I. Method

This option specifies the method used to calculate the bootstrap confidence intervals. The reflection method is recommended.

- **Percentile**

  The confidence limits are the corresponding percentiles of the bootstrap values

- **Reflection**

  The confidence limits are formed by reflecting the percentile limits. If X0 is the original value of the parameter estimate and XL and XU are the percentile confidence limits, the reflection interval is (2 X0 - XU, 2 X0 - XL).

Reports Tab

These options control which reports are displayed. Note that many of these reports are only needed is special situations. You will only need a few reports for a typical robust regression analysis.

Alphas and Confidence Levels

Test Alpha

Alpha is the significance level used in conducting the hypothesis tests. The value of 0.05 is usually used. This corresponds to a chance of 1 out of 20. However, you should not be afraid to use other values since 0.05 became popular in pre-computer days when it was the only value available. Typical values range from 0.01 to 0.20.

© NCSS, LLC. All Rights Reserved.

Assumptions Alpha

This value specifies the significance level that must be achieved to reject a preliminary test of an assumption. In regular hypothesis tests, common values of alpha are 0.05 and 0.01. However, most statisticians recommend that preliminary tests of assumptions use a larger alpha such as 0.10, 0.15, or 0.20.

We recommend 0.20.

Confidence Level

Enter the confidence level (or confidence coefficient) as a percentage for the confidence intervals reported. The interpretation of confidence level is that if confidence intervals are constructed across many experiments at the same confidence level, the percentage of such intervals that surround the true value of the parameter is equal to the confidence level.

Typical values range from 80 to 99.99. Usually, 95 is used.

Select Reports

Check those reports that you want to see.

Show

This option makes it possible to display fewer observations in the row-by-row lists. This is especially useful when you have a lot of observations.

Weight Cutoff

On the Residuals and Weights report, only rows with weights less than this amount in at least one regression are displayed. This report allows you to quickly focus on those rows that have been down-weighted.

The possible range is 0.000 to 1.00. We recommend 0.20.

Report Options Tab

These options specify the number of decimal places shown when the indicated value is displayed in a report. Note that the number of decimal places shown in plots is controlled by the Tick Label Settings buttons on the Axes tabs.

Variable Labels

Variable Names

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

Stagger label and output if label length is ≥

When writing a row of information to a report, some variable names/labels may be too long to fit in the space allocated. If the name (or label) contains more characters than specified here, the rest of the output for that line is moved down to the next line. Most reports are designed to hold a label of up to 15 characters.

Enter ***1*** when you always want each row's output to be printed on two lines.

Enter ***100*** when you want each row printed on only one line. Note that this may cause some columns to be miss-aligned.

© NCSS, LLC. All Rights Reserved.

Decimal Places

Precision

This option is used when the number of decimal places is set to ***All***. It specifies whether numbers are displayed as single (7-digit) or double (13-digit) precision numbers in the output. All calculations are performed in double precision regardless of the Precision selected here.

- **Single**
  Unformatted numbers are displayed with 7-digits.

- **Double**
  Unformatted numbers are displayed with 13-digits. This option is most often used when the extremely accurate results are needed for further calculation.

Double Precision Format Misalignment

Double precision numbers may require more space than is available in the output columns, causing column alignment problems. The double precision option is for those instances when accuracy is more important than format alignment.

Reg Coefficients ... Diagonal of X'X Inverse Decimals

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

- **All**
  Select ***All*** to display all digits available. The number of digits displayed by this option is controlled by whether the *Precision* option is *Single* (7) or *Double* (13).

Plots Tab

These options control the inclusion and the settings of each of the plots.

Select Plots

Histogram ... Resids vs X Plot

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

Edit During Run

This is the small check-box in the upper right-hand corner of the format button. If checked, the graphics format window for this plot will be displayed while the procedure is running so that you can format it with the actual data.

Storage Tab

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

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

© NCSS, LLC. All Rights Reserved.

Data Storage Options

Storage Option

This option controls whether the values indicated below are stored on the dataset when the procedure is run.

- **Do not store data**
  No data are stored even if they are checked.

- **Store in empty columns only**
  The values are stored in empty columns only. Columns containing data are not used for data storage, so no data can be lost.

Select Items to Store with the Dataset

Predicted Y ... VC(Betas) Matrix

Indicate whether to store these row-by-row values, beginning at the column indicated by the *Store First Variable In* option.

Note that the results for each of the three regressions are stored in separate columns.

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.

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 Mediation Analysis window.

1 Open the Mediation dataset.

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

2 Open the Mediation Analysis window.

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

3 Specify the variables and model.

- On the Mediation Analysis window, select the **Variables, Model tab**.
- Set the **Y** box to **Water**.
- Set the **X** box to **Temp**.
- Set the **M** box to **Thirst**.
- Set the **Terms** box to **1-Way**.

4 Specify the Robust Regression and Bootstrap options.

- On the Mediation Analysis window, select the **Robust, Bootstrap tab**.
- Set the **Estimation Method** box to **Ordinary Least Squares**.
- Check the **Bootstrap Calculations** box.
- Set the **Samples** to **3000**.

© NCSS, LLC. All Rights Reserved.

5 Specify the Reports to be displayed.

- On the Mediation Analysis window, select the **Reports tab**.
- Check all of the reports.

6 Specify the Plots to be displayed.

- On the Mediation Analysis window, select the **Plots tab**.
- Check the following plots: Histogram, Probability Plot, Y vs X, and Resids vs X.

7 Run the procedure.

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

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
		Sum of Robust Weights	100.000

Run Summary Detail Report

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

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

	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.

© NCSS, LLC. All Rights Reserved.

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

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

Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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	0.08543313	-0.295	0.7680	-0.1040223	0.08726753

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)

--- Estimation Results ---		--- Bootstrap Confidence Limits ---		
Parameter	Estimate	Conf. Level	Lower	Upper
Indirect Effect (AB)				
Original Value	-0.02520181	95.00	-0.1040223	0.08726753
Bootstrap Mean	-0.02918646			
Bias (BM – OV)	-0.003984644			
Bias Corrected	-0.02121717			
Standard Error	0.08543313			

Sampling Method = Observation, Confidence Limit Type = Reflection, Number of Samples = 3000.

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

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.

© NCSS, LLC. All Rights Reserved.

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.

Conf. 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 Section

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-Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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

$R^2 = 0.5482$

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-Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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

$R^2 = 0.5475$

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-Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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

$R^2 = 0.5482$

© NCSS, LLC. All Rights Reserved.

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 to test 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.

Prob Level

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 - Upper 95% Conf. Limit of 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

© NCSS, LLC. All Rights Reserved.

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 Report for Model: Y = X M

Source	DF	R^2	Sum of Squares	Mean Square	F-Ratio	Prob Level
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 Report for Model: Y = X

Source	DF	R^2	Sum of Squares	Mean Square	F-Ratio	Prob Level
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 Report for Model: M = X

Source	DF	R^2	Sum of Squares	Mean Square	F-Ratio	Prob Level
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*.

© NCSS, LLC. All Rights Reserved.

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.

Prob Level

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 Report for Model: Y = X M

Independent Variable (IV)	Total R² for this IV and IV's Above	Increase in R² if this IV 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 Report for Model: Y = X

Independent Variable (IV)	Total R² for this IV and IV's Above	Increase in R² if this IV 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 Report for Model: M = X

Independent Variable (IV)	Total R² for this IV and IV's Above	Increase in R² if this IV 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

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

© NCSS, LLC. All Rights Reserved.

Total R^2 for This I.V. 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.

R^2 Increase When This IV Added to Those 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.

R^2 Decrease When This IV is 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 When This IV Is 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 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 Reports

Multicollinearity Report for Model: Y = X M

Independent Variable	Variance Inflation Factor	R^2 Versus Other I.V.'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 Report for Model: Y = X

Independent Variable	Variance Inflation Factor	R^2 Versus Other I.V.'s	Tolerance	Diagonal of X'X Inverse
Temp	1.0000	0.0000	1.0000	4.59007E-05

Multicollinearity Report for Model: M = X

Independent Variable	Variance Inflation Factor	R^2 Versus Other I.V.'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.

© NCSS, LLC. All Rights Reserved.

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

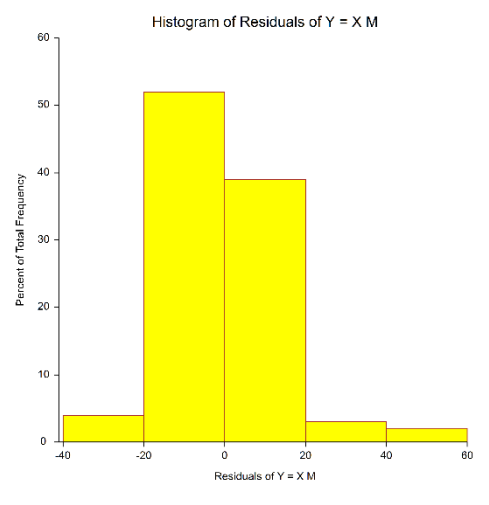
Row	Water Y	Temp X	Thirst M	Residual Y = X M	Residual Y = X	Residual M = X
1	84	68	9.0000	23.5495	23.2497	1.0524
2	33	36	5.0000	-2.4286	-2.3953	-0.1167
3	31	34	4.0000	-3.0784	-2.8106	-0.9398
4	21	20	3.0000	-1.9176	-1.7178	-0.7012
5	51	53	7.0000	2.2429	2.1348	0.3794
6	41	45	6.0000	-1.5016	-1.5264	0.0871
7	34	33	4.0000	0.7392	0.9817	-0.8513
8	57	62	8.0000	1.1699	1.0037	0.5832
9	58	66	8.0000	-1.1003	-1.1656	0.2293
10	23	69	9.0000	-38.2681	-38.5427	0.9639
11	47	53	7.0000	-1.7571	-1.8652	0.3794
12	37	36	5.0000	1.5714	1.6047	-0.1167
13	55	61	8.0000	-0.0126	-0.2039	0.6716
14	52	56	7.0000	0.7903	0.7578	0.1140
15	63	70	9.0000	0.9144	0.6650	0.8754
16	45	47	6.0000	0.8633	0.8889	-0.0898
17	61	70	2.0000	-3.0797	-1.3350	-6.1246
18	56	66	8.0000	-3.1003	-3.1656	0.2293
19	49	56	7.0000	-2.2097	-2.2422	0.1140
20	21	51	7.0000	-26.1220	-26.2805	0.5563
(continues for 100 rows)	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

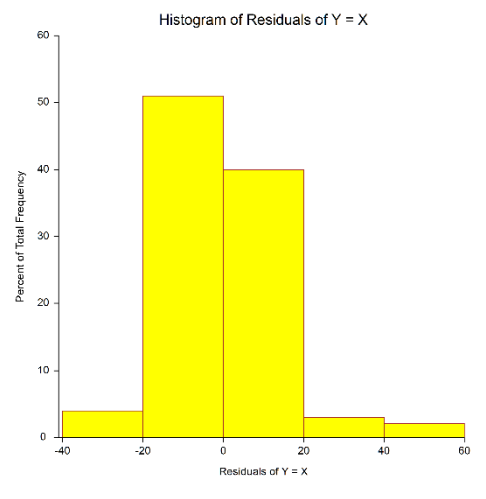
This report gives the values of Y, X, and M for each row followed by the residuals from the three regression models. It allows you to quickly see any rows with large residuals in at least one of the regressions.

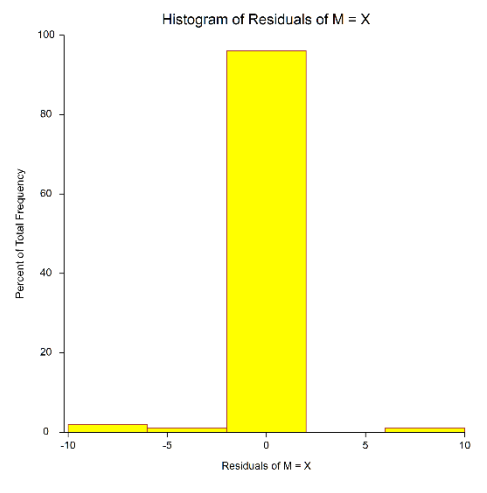
© NCSS, LLC. All Rights Reserved.

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.





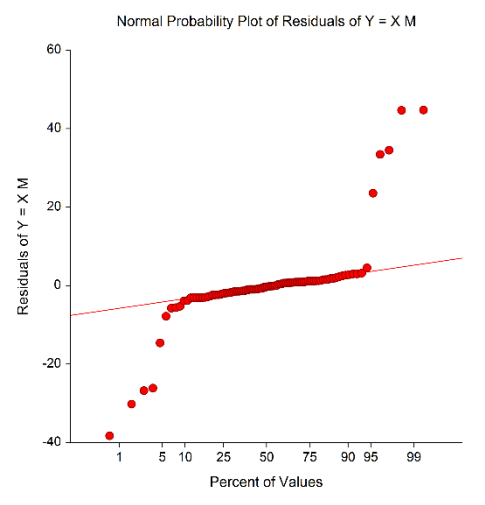


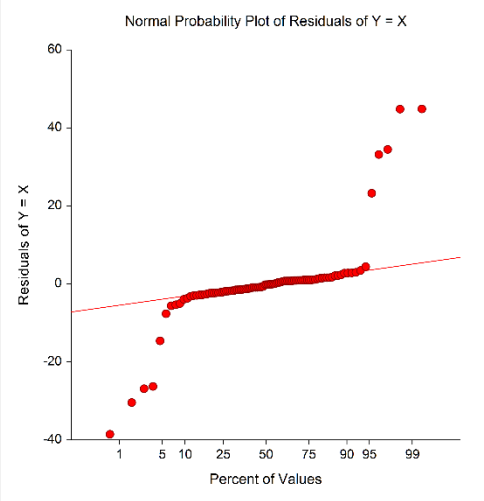
© NCSS, LLC. All Rights Reserved.

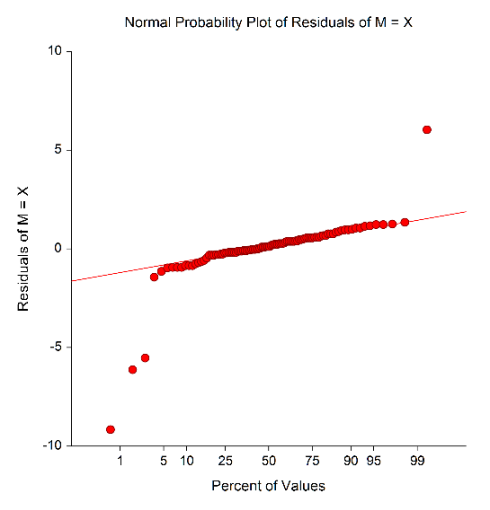
Probability Plots of Residuals

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.



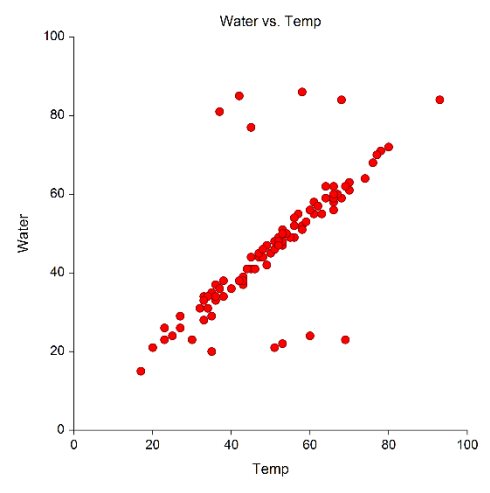


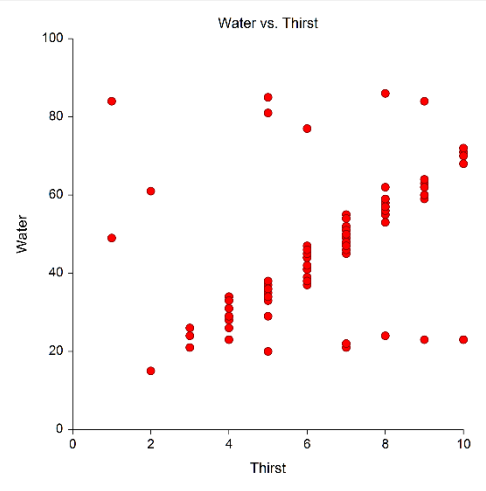


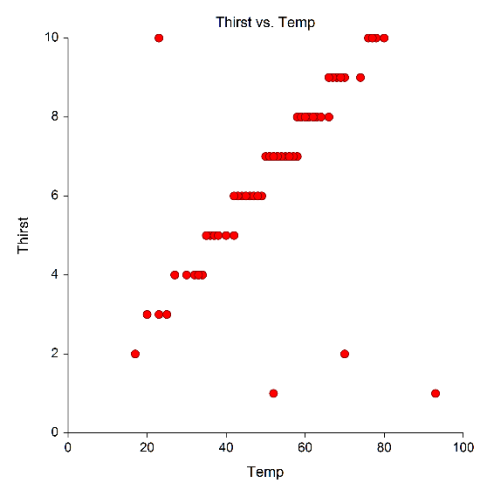
© NCSS, LLC. All Rights Reserved.

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.



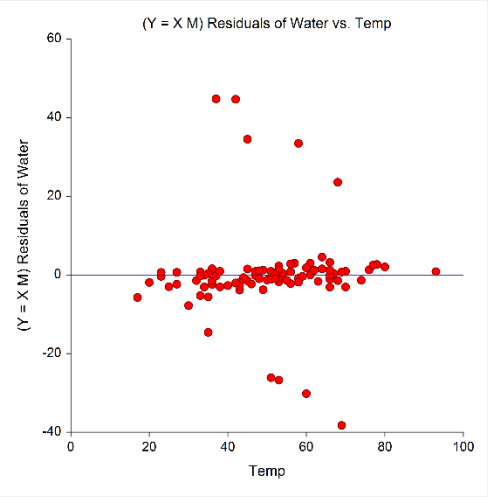


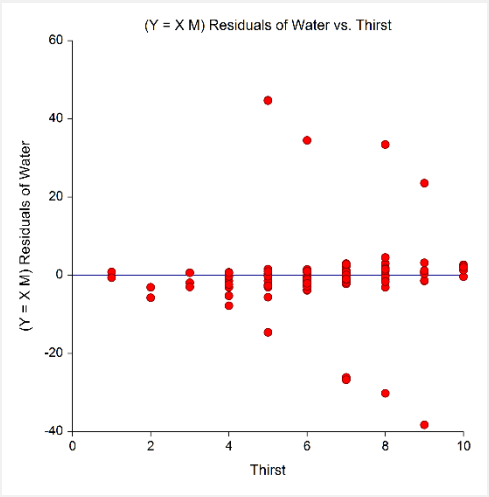


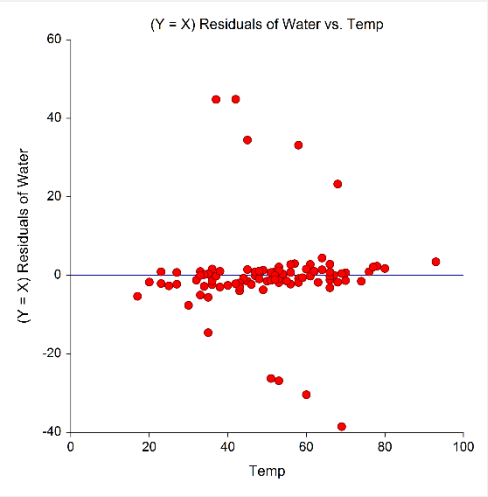
© NCSS, LLC. All Rights Reserved.

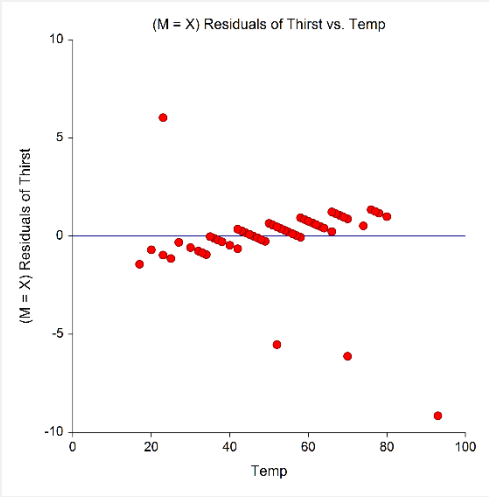
Scatter Plots of Residuals versus each Independent Variable

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









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

© NCSS, LLC. All Rights Reserved.

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.

You may follow along here by making the appropriate entries or load the completed template **Example 2** by clicking on Open Example Template from the File menu of the Mediation Analysis window.

1 Open the Mediation dataset.

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

2 Open the Mediation Analysis window.

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

3 Specify the variables and model.

- On the Mediation Analysis window, select the **Variables, Model tab**.
- Set the **Y** box to **Water**.
- Set the **X** box to **Temp**.
- Set the **M** box to **Thirst**.
- Set the **Terms** box to **1-Way**.

4 Specify the Robust Regression and Bootstrap options.

- On the Mediation Analysis window, select the **Robust, Bootstrap tab**.
- Set the **Estimation Method** box to **Robust Regression using Huber's Method**.
- Set the **Type of Weights Assumed** box to **Random**.
- Check the **Bootstrap Calculations** box.
- Set the **Samples** to **300**.

5 Specify the Reports to be displayed.

- On the Mediation Analysis window, select the **Reports tab**.
- Check the following reports: Run Summary, Mediation Effects, Individual Regressions, Robust Iterations – Coefficients, Y, X, M, and Weights.
- Set the **Weight Cutoff** to **0.3**.

6 Specify the Plots to be displayed.

- On the Mediation Analysis window, select the **Plots tab**.
- Check the following plots: Resids vs X.

7 Run the procedure.

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

© NCSS, LLC. All Rights Reserved.

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 C.L.Type	Reflection		

Run Summary Detail Report

Regression Model	R^2	Robust Iterations	Max % Chng In Any Robust Coef	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 Detail Report from Example 1 is repeated here.

Run Summary Detail Report (From Example 1)

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

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

Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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	0.01960082	-0.132	0.8951	-0.05787071	0.02083182

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

© NCSS, LLC. All Rights Reserved.

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

From Example 1

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

Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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	0.08543313	-0.295	0.7680	-0.1040223	0.08726753

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 changed.

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.

You may follow along here by making the appropriate entries or load the completed template **Example 3** by clicking on Open Example Template from the File menu of the Mediation Analysis window.

1 Open the Mediation dataset.

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

2 Open the Mediation Analysis window.

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

3 Specify the variables and model.

- On the Mediation Analysis window, select the **Variables, Model tab**.
- Set the **Y** box to **Water**.
- Set the **X** box to **Temp**.
- Set the **M** box to **Thirst**.
- Set **C (Numeric Covariates)** to **Age**.
- Set **C (Categorical Covariates)** to **Adults**.
- Set **Default Recoding Scheme** to **Compare Each with Next**.
- Set the **Terms** box to **1-Way**.

4 Specify the Robust Regression and Bootstrap options.

- On the Mediation Analysis window, select the **Robust, Bootstrap tab**.
- Set the **Estimation Method** box to **Ordinary Least Squares**.

© NCSS, LLC. All Rights Reserved.

- Check the **Bootstrap Calculations** box.
- Set the **Samples** to **300**.

5 Specify the Reports to be displayed.

- On the Mediation Analysis window, select the **Reports tab**.
- Check the following reports: Run Summary, Mediation Effects, Individual Regressions, Robust Iterations – Coefficients, Y, X, M, and Weights.
- Set the **Weight Cutoff** to **0.3**.

6 Specify the Plots to be displayed.

- On the Mediation Analysis window, select the **Plots tab**.
- Check the following plots: Resids vs X.

7 Run the procedure.

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

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
		Sum of Robust Weights	100.000

Run Summary Detail Report

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

From Example 3

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

Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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	0.0875705	-0.293	0.7697	-0.09372456	0.0815872

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

© NCSS, LLC. All Rights Reserved.

Mediation Analysis

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

From Example 1
Estimation: Ordinary least squares
Sb(Indirect): First-order
Bootstrap: Number of samples = 3000, Confidence limit type = Reflection
Y: Water
X: Temp
M: Thirst
Covariates: 0

Type of Effect	Regression Coefficient b(i)	Standard Error Sb(i)	Statistic to Test H0: β(i)=0	Prob Level	Lower 95% C.L. of β(i)	Upper 95% C.L. of β(i)
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	0.08543313	-0.295	0.7680	-0.1040223	0.08726753

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

© NCSS, LLC. All Rights Reserved.