Chapter 308

Robust Regression

Introduction

Multiple regression analysis is documented in *Chapter 305 – Multiple Regression*, so that information will not be repeated here. Refer to that chapter for in depth coverage of multiple regression analysis. This chapter will deal solely with the topic of 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 %. 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**.

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.

M-Estimators

Several families of robust estimators have been developed. The robust methods found in ***NCSS*** fall into the family of *M-estimators*. This estimator minimizes the sum of a function $\rho(\cdot)$ of the residuals. That is, these estimators are defined as the β*'s* that minimize

$$\min_{\beta} \sum_{j=1}^{N} \rho\left(y_j - x_j'\beta\right) = \min_{\beta} \sum_{j=1}^{N} \rho\left(e_j\right)$$

M in *M*-estimators stands for maximum likelihood since the function $\rho(\cdot)$ is related to the likelihood function for a suitable choice of the distribution of the residuals. In fact, when the residuals follow the normal distribution, setting $\rho(u) = \frac{1}{2}u^2$ results in the usual method of least squares.

Unfortunately, *M*-estimators are not necessarily *scale invariant*. That is, these estimators may be influenced by the scale of the residuals. A scale-invariant estimator is found by solving

$$\min_{\beta}\sum_{j=1}^{N}\rho\left(\frac{y_j - x_j'\beta}{s}\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(\frac{e_j}{s}\right) = \min_{\beta}\sum_{j=1}^{N}\rho\left(u_j\right)$$

where *s* is a robust estimate of scale. The estimate of *s* is used in *NCSS* is

$$s = \frac{median\left|e_j - median\left(e_j\right)\right|}{0.6745}$$

This estimate of *s* yields an approximately unbiased estimator of the standard deviation of the residuals when *N* is large and the error distribution is normal.

The function

$$\sum_{j=1}^{N}\rho\left(\frac{y_j - x_j'\beta}{s}\right)$$

is minimized by setting the first partial derivatives of $\rho(\cdot)$ with respect to each β_i to zero which forms a set of $p + 1$ nonlinear equations

$$\sum_{j=1}^{N}x_{ij}\psi\left(\frac{y_j - x_j'\beta}{s}\right) = 0, \quad i = 0,1,\cdots,p$$

where $\psi\left(u\right) = \rho'\left(u\right)$ is the *influence function*.

These equations are solved iteratively using an approximate technique called iteratively reweighted least squares (IRLS). At each step, new estimates of the regression coefficients are found using the matrix equation

$$\beta_{t+1} = \left(\mathbf{X}'\mathbf{W}_t\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{W}_t\mathbf{Y}$$

where W_t is an *N-by-N* diagonal matrix of weights $w_{1t}, w_{2t}, \cdots, w_{Nt}$ defined as

$$w_{jt} = \begin{cases} \dfrac{\psi\left[\left(y_j - x'\beta_{jt}\right)/s_t\right]}{\left(y_j - x'\beta_{jt}\right)/s_t} & \text{if } y_j \neq x'\beta_{jt} \\ 1 & \text{if } y_j = x'\beta_{jt} \end{cases}$$

The ordinary least squares regression coefficients are used at the first iteration to begin the iteration process. Iterations are continued until there is little or no change in the regression coefficients from one iteration to the next. Because of the masking nature of outliers, it is a good idea to run through at least five iterations to allow the outliers to be found.

Two functions are available in ***NCSS.*** These are Huber's method and Tukey's biweight. Huber's method is the most frequently recommended in the regression texts that we have seen. The specifics for each of these functions are as follows.

Huber's Method

$$\rho(u) = \begin{cases} u^2 & \text{if } |u| < c \\ |2u|c - c^2 & \text{if } |u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u & \text{if } |u| < c \\ c\,\text{sign}(u) & \text{if } |u| \geq c \end{cases}$$

$$w(u) = \begin{cases} 1 & \text{if } |u| < c \\ c / |u| & \text{if } |u| \geq c \end{cases}$$

$$c = 1.345$$

Tukey's Biweight

$$\rho(u) = \begin{cases} \frac{c^2}{3}\left\{1 - \left[1 - \left(\frac{u}{c}\right)^2\right]^3\right\} & \text{if } |u| < c \\ 2c & \text{if } |u| \geq c \end{cases}$$

$$\psi(u) = \begin{cases} u\left[1 - \left(\frac{u}{c}\right)^2\right]^2 & \text{if } |u| < c \\ 0 & \text{if } |u| \geq c \end{cases}$$

$$w(u) = \begin{cases} \left[1 - \left(\frac{u}{c}\right)^2\right]^2 & \text{if } |u| < c \\ 0 & \text{if } |u| \geq c \end{cases}$$

$$c = 4.685$$

This gives you a sketch of what robust regression is about. If you find yourself using the technique often, we suggest that you study one of the modern texts on regression analysis. All of these 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 and Tests for M-Estimates

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 will cause large changes in the resulting tests and confidence intervals. For this reason, both methods are available.

Data Structure

The data are entered in two or more columns. An example of data appropriate for this procedure is shown below. These data are from a study of the relationship of several variables with a person's I.Q. Fifteen people were studied. Each person's IQ was recorded along with scores on five different personality tests. The data are contained in the IQ dataset. We suggest that you open this database now so that you can follow along with the example.

IQ dataset

Test1	Test2	Test3	Test4	Test5	IQ
83	34	65	63	64	106
73	19	73	48	82	92
54	81	82	65	73	102
96	72	91	88	94	121
84	53	72	68	82	102
86	72	63	79	57	105
76	62	64	69	64	97
54	49	43	52	84	92
37	43	92	39	72	94
42	54	96	48	83	112
71	63	52	69	42	130
63	74	74	71	91	115
69	81	82	75	54	98
81	89	64	85	62	96
50	75	72	64	45	103

Missing Values

Rows with missing values in the variables being analyzed are ignored. If data are present on a row for all but the dependent variable, a predicted value and confidence limits are generated for that row.

Example 1 – Robust Regression (Common Reports)

This section presents an example of how to run a robust regression analysis of the data presented earlier in this chapter. The data are in the IQ dataset. This example will run a robust regression of *IQ* on *Test1* through *Test5*. This program outputs over thirty different reports and plots, many of which contain duplicate information. If you want to obtain complete documentation for all reports, refer to the Multiple Regression chapter. Only those reports that are specifically needed for a robust regression will be presented here.

Setup

To run this example, complete the following steps:

1 Open the IQ example dataset

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

2 Specify the Robust Regression procedure options

- Find and open the **Robust Regression** procedure using the menus or the Procedure Navigator.
- The settings for this example are listed below and are stored in the **Example 1** settings template. To load this template, click **Open Example Template** in the Help Center or File menu.

Option	Value
Variables, Model Tab	
Y	**IQ**
Numeric X's	**Test1-Test5**
Terms	**1-Way**

3 Run the procedure

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

Run Summary

Run Summary Report

Item	Value
Dependent Variable	IQ
Number Ind. Variables	5
Weight Variable	None
Robust Method	Huber's Method
Tuning Constant	1.345
MAD Scale Factor	0.6745

Rows	Value
Number Processed	17
Number Used in Estimation	15
Number Filtered Out	0
Number with X's Missing	0
Number with Weight Missing	0
Number with Y Missing	2
Sum of Robust Weights	13.065

Run Information	Value
Iterations	15
Max % Change in any Coef	0.001
R^2 after Robust Weighting	0.6521
S using MAD	3.88
S using MSE	6.41
Completion Status	Normal Completion

This report summarizes the robust regression results. It presents the variables used, the number of rows used, and the basic results. Of particular interest is the number of iterations performed (15 here) and the Max % Change in any Coefficient since they establish whether the algorithm converged before iterations were stopped. The S using MAD should be compared to the S using MSE to determine the impact of the outliers.

Descriptive Statistics

Descriptive Statistics

Variable	Count	Mean	Standard Deviation	Minimum	Maximum
Test1	15	67.991	16.708	37	96
Test2	15	60.117	18.462	19	89
Test3	15	73.025	13.171	43	96
Test4	15	64.922	13.108	39	88
Test5	15	71.981	14.104	42	94
IQ	15	102.921	8.714	92	130

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

Robust Iterations - Coefficients

Robust Iterations - Coefficients

Robust Iteration	Max Percent Change in Coefficients	b(0)	b(1)	b(2)	b(3)
0		85.24038	-1.93357	-1.65988	0.10495
1	244.726	71.67678	-1.67990	-1.42832	0.16483
2	61.163	66.77071	-1.58814	-1.34456	0.18649
3	23.552	62.35068	-1.47177	-1.23677	0.19507
4	3.886	60.89347	-1.41804	-1.18871	0.19520
5	1.493	60.86422	-1.41349	-1.18315	0.19330
6	0.795	60.79848	-1.40846	-1.17854	0.19262
7	0.431	60.78759	-1.40668	-1.17690	0.19226
8	0.199	60.78211	-1.40586	-1.17614	0.19210
9	0.091	60.77952	-1.40548	-1.17579	0.19203
10	0.042	60.77829	-1.40531	-1.17563	0.19200
11	0.019	60.77772	-1.40523	-1.17556	0.19199
12	0.009	60.77745	-1.40519	-1.17552	0.19198
13	0.004	60.77733	-1.40517	-1.17551	0.19198
14	0.002	60.77727	-1.40516	-1.17550	0.19198
15	0.001	60.77725	-1.40516	-1.17550	0.19198

This report shows the largest percent change in any of the coefficients as well as the first four coefficients. The 0^{th} iteration shows the ordinary least squares estimates on the full dataset.

The report allows you to determine if enough iterations have been run for the coefficients to have stabilized. In this example, the coefficients have stabilized. If they had not, we would increase the number of robust iterations and rerun the analysis.

Robust Iterations - Residuals

Robust Iterations - Residuals

Robust Iteration	Max Percent Change in Coefficients	— Percentiles of Absolute Residuals — 25th	50th	75th	100th
0		2.767	5.073	9.167	22.154
1	244.726	1.726	4.446	7.637	27.573
2	61.163	1.573	3.093	7.084	29.533
3	23.552	1.511	2.599	7.083	30.626
4	3.886	1.564	2.285	7.296	30.714
5	1.493	1.569	2.271	7.387	30.604
6	0.795	1.581	2.252	7.440	30.553

7	0.431	1.586	2.246	7.464	30.525
8	0.199	1.589	2.243	7.475	30.513
9	0.091	1.590	2.242	7.480	30.507
10	0.042	1.590	2.241	7.483	30.504
11	0.019	1.590	2.241	7.484	30.503
12	0.009	1.590	2.241	7.484	30.502
13	0.004	1.591	2.241	7.484	30.502
14	0.002	1.591	2.241	7.485	30.502
15	0.001	1.591	2.240	7.485	30.502

The purpose of this report is to highlight the percentage changes among the coefficients and to show the convergence of the absolute value of the residuals after a selected number of iterations.

Robust Iteration

This is the robust iteration number.

Max Percent Change in Coefficients

This is the maximum percentage change in any of the regression coefficients from one iteration to the next. This quantity can be used to determine if enough iterations have been run. Once this value is less than 0.01%, little is gained by further iterations.

Percentiles of Absolute Residuals

The absolute values of the residuals for this iteration are sorted and the percentiles are calculated. We want to terminate the iteration process when there is little change in median of the absolute residuals.

Regression Coefficients T-Tests Assuming Fixed Weights

Regression Coefficients T-Tests Assuming Fixed Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Standard-ized Coefficient	T-Statistic to Test H0: β(i)=0	Prob Level	Reject H0 at 5%?
Intercept	60.77725	15.683629	0.0000	3.875	0.0038	Yes
Test1	-1.40516	0.633752	-2.6941	-2.217	0.0538	No
Test2	-1.17550	0.540272	-2.4904	-2.176	0.0576	No
Test3	0.19198	0.139920	0.2902	1.372	0.2033	No
Test4	2.86553	1.128164	4.3102	2.540	0.0317	Yes
Test5	0.11523	0.132366	0.1865	0.871	0.4066	No

This report gives the coefficients, standard errors, and significance tests assuming that the robust weights are fixed, known quantities.

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 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. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error Sb(i)

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests or confidence limits.

Standardized Coefficient

Standardized regression coefficients are the coefficients that would be obtained if you standardized the independent variables and the dependent variable. Here *standardizing* is defined as subtracting the mean and dividing by the standard deviation of a variable. A regression analysis on these standardized variables would yield these standardized coefficients.

When the independent variables have vastly different scales of measurement, this value provides a way of making comparisons among variables. The formula for the standardized regression coefficient is:

$$b_{j,\,std} = b_j\left(\frac{s_{X_j}}{s_Y}\right)$$

where s_Y and s_{X_j} are the standard deviations for the dependent variable and the j^{th} independent variable.

T-Statistic to test H0: β(i)=0

This is the t-test value for testing the hypothesis that $\beta_j = 0$ versus the alternative that $\beta_j \neq 0$ 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.

Reject H0 at 5%?

This is the conclusion reached about the null hypothesis. It will be either reject *H0* at the 5% level of significance or not.

Note that the level of significance is specified in the Test Alpha box on the Format tab panel.

Regression Coefficients T-Tests Assuming Random Weights

Regression Coefficients T-Tests Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Standard-ized Coefficient	T-Statistic to Test H0: β(i)=0	Prob Level	Reject H0 at 5%?
Intercept	60.77725	15.927008	0.0000	3.816	0.0041	Yes
Test1	-1.40516	0.691721	-2.6941	-2.031	0.0728	No
Test2	-1.17550	0.586729	-2.4904	-2.003	0.0761	No
Test3	0.19198	0.147810	0.2902	1.299	0.2263	No
Test4	2.86553	1.233082	4.3102	2.324	0.0452	Yes
Test5	0.11523	0.135253	0.1865	0.852	0.4164	No

This report gives the coefficients, standard errors, and significance tests assuming that the robust weights are random, unknown quantities found from the data. This is a much more reasonable assumption than that the weights are fixed.

Regression Coefficients Confidence Intervals Assuming Fixed Weights

Regression Coefficients Confidence Intervals Assuming Fixed Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	60.77725	15.683629	25.29841	96.25608
Test1	-1.40516	0.633752	-2.83881	0.02849
Test2	-1.17550	0.540272	-2.39768	0.04668
Test3	0.19198	0.139920	-0.12454	0.50850
Test4	2.86553	1.128164	0.31345	5.41761
Test5	0.11523	0.132366	-0.18421	0.41466

Note: The T-Value used to calculate these confidence limits was 2.262.

This report gives the coefficients, standard errors, and confidence interval assuming fixed weights.

Independent Variable

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

Regression Coefficient

The regression coefficients are the least squares estimates of the parameters. The value indicates how much change in *Y* occurs for a one-unit change in *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. These coefficients are the values of $b_0, b_1, \cdots, b_p$.

Standard Error Sb(i)

The standard error of the regression coefficient, s_{b_j}, is the standard deviation of the estimate. It is used in hypothesis tests and confidence limits.

Lower - Upper 95% Conf. Limit of β(i)

These are the lower and upper values of a $100(1-\alpha)\%$ interval estimate for β_j 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:

$$b_j \pm t_{1-\alpha/2,n-p-1}\ s_{b_j}$$

Note: The T-Value …

This is the value of $t_{1-\alpha/2,n-p-1}$ used to construct the confidence limits.

Regression Coefficients Confidence Intervals Assuming Random Weights

Regression Coefficients Confidence Intervals Assuming Random Weights

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Lower 95% Conf. Limit of β(i)	Upper 95% Conf. Limit of β(i)
Intercept	60.77725	15.927008	24.74785	96.80664
Test1	-1.40516	0.691721	-2.96994	0.15962
Test2	-1.17550	0.586729	-2.50277	0.15177
Test3	0.19198	0.147810	-0.14239	0.52635
Test4	2.86553	1.233082	0.07610	5.65495
Test5	0.11523	0.135253	-0.19074	0.42119

Note: The T-Value used to calculate these confidence limits was 2.262.

This report gives the coefficients, standard errors, and confidence interval assuming random weights.

Estimated Equation

Estimated Equation

IQ =
60.7772455334515 - 1.40516102388176 * Test1 - 1.17549846722092 * Test2 + 0.191975633021348 * Test3 + 2.86552848363198 * Test4 + 0.115225465460253 * Test5

This is the estimated robust regression line presented in double precision. Besides showing the regression model in long form, it may be used as a transformation by copying and pasting it into the Transformation portion of the spreadsheet.

Robust Residuals and Weights

Robust Residuals and Weights

Row	Actual IQ	Predicted IQ	Residual	Absolute Percent Error	Robust Weight
1	106.000	104.563	1.437	1.356	1.0000
2	92.000	96.874	-4.874	5.298	1.0000
3	102.000	100.096	1.904	1.867	1.0000
4	121.000	121.713	-0.713	0.590	1.0000
5	102.000	98.569	3.431	3.364	1.0000
6	105.000	100.337	4.663	4.441	1.0000
7	97.000	98.486	-1.486	1.532	1.0000
8	92.000	94.240	-2.240	2.435	1.0000
9	94.000	95.953	-1.953	2.078	1.0000
10	112.000	103.822	8.178	7.302	0.6382
11	130.000	99.498	30.502	23.463	0.1711
12	115.000	113.409	1.591	1.383	1.0000
13	98.000	105.485	-7.485	7.637	0.6973
14	96.000	105.340	-9.340	9.729	0.5588
15	103.000	104.758	-1.758	1.707	1.0000
16		90.381			0.0000
17		96.301			0.0000

The predicted values, the residuals, and the robust weights are reported for the last iteration. These robust weights can be saved for use in a weighted regression analysis, or they can be used as a filter to delete observations with a weight less than some number, say 0.20, in an ordinary least squares regression analysis.

Note that in this analysis, row 11 appears to be an outlier.

Row
This is the number of the row.

Actual
This is the actual value of the dependent variable.

Predicted
This is the predicted value of *Y* based on the robust regression equation from the final iteration.

Residual
The residual is the difference between the Actual and Predicted values of *Y*.

Absolute Percent Error
This is the Residual divided by the Actual times 100.

Robust Weight
These are the final robust weights for each observation. These weights will range from zero to one. Observations with a low weight make a minimal contribution to the determination of the regression coefficients. In fact, observations with a weight of zero have been deleted from the analysis. These weights can be saved and used again in a weighted least squares regression.

Residuals vs X's Plots

These are the scatter plots of the residuals versus each independent variable. Again, the preferred pattern is a rectangular shape or point cloud. Any other nonrandom pattern may require a redefining of the regression model.

Residuals vs X's Plots