Chapter 345

Nondetects-Data Regression

Introduction

This module fits the regression relationship between a positive-valued dependent variable (with, possibly, some nondetected responses) and one or more independent variables. The distribution of the residuals (errors) is assumed to follow the exponential, extreme value, logistic, log-logistic, lognormal, lognormal10, normal, or Weibull distribution. The Distribution Fitting module may be useful for determining a suitable distribution for use in Nondetects Regression.

Nondetects analysis is the analysis of data in which one or more of the values cannot be measured exactly because they fall below one or more detection limits. Detection limits often arise in environmental studies because of the inability of instruments to measure small concentrations. Some examples of sampling scenarios that lead to datasets with nondetects values are finding pesticide concentrations in water, determining chemical composition of soils, or establishing the number of particulates of a compound in the air.

A common practice for dealing with values which fall below the detection threshold is substitution. Often, each value which is below the detection limit is substituted with one half the detection limit. Evaluation of relationships among variables are then carried out using standard techniques (multiple regression) with the substituted data. Helsel (2005) warns of the potential data analysis biases that result if nondetects values are substituted. He particularly warns about the arbitrariness of substituting one half the detection limit (or zero, or the detection limit). Alternatively, if a proper distribution can be assumed for the variable with nondetects values, maximum likelihood distribution regression is a more appropriate analog to multiple regression with substituted values.

For a complete account of nondetects analysis, we suggest the book by Helsel (2005).

Technical Details

The linear regression equation is

$$Y = B_0 + B_1 X_1 + B_2 X_2 + \cdots + Se$$

Here, S represents the value of a constant standard deviation, Y is the response or a transformation of the response (*ln()* or *log()*), the *X's* are one or more independent variables, the B's are the regression coefficients, and e is the residual (error) that is assumed to follow a particular probability distribution. The problem reduces to estimating the B's and S. The density functions of the eight distributions that are fit by this module are given in the Distribution Fitting section and will not be repeated here.

As an example, we give detailed results for the lognormal distribution. The results for other distributions follow a similar pattern.

The lognormal probability density function may be written as

$$f(y|M,S)=\frac{1}{yS\sqrt{2\pi}}e^{-\frac{1}{2}\left(\frac{\ln(y)-M}{S}\right)^2}$$

If we replace the location parameter, *M*, with the regression model, the density now becomes

$$f(y|B_0\cdots B_p,S)=\frac{1}{yS\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\left(\frac{\ln(y)-\sum_{i=0}^{p}B_iX_i}{S}\right)^2\right\}$$

Maximum likelihood estimation consists of finding the values of the distribution parameters that maximize the log-likelihood of the data values. Loosely speaking, these are the values of the parameters, which maximize the probability that the current set of data values occur.

NCSS employs the Newton-Raphson algorithm with numerical differentiation to obtain the maximum likelihood estimates. These estimates have been shown to have optimality characteristics in large samples (number of responses greater than 20).

Data Structure

Nondetects responses are specified using up to three components: the response value (e.g., concentration or amount), an optional indicator of whether or not each observation was detected, and an optional frequency (count) specification. If no detection indicator is included, all response values represent detected responses. If the frequency (count) variable is omitted, all counts are assumed to be one.

Any number of independent variables may be specified as separate columns. In Nondetects Distribution Regression, all independent variables must be numeric. If categorical variables are to be used, corresponding zero-one variables must first be created.

Sample Dataset

The table below shows a dataset (fictitious) reporting 1,3-dichloropropene (1,3-DCP) concentrations (in µg/L) for 53 randomly chosen soil locations. Concentrations were determined following addition of one of two solutions to each sample: water or NaHSO4. Some of the soil samples resulted in concentrations below the laboratory minimum reporting limit of 0.13µg/L. The percent moisture in the soil sample is also reported. A value of zero in the DNondet column indicates 1,3-DCP was detected. A value of one in the DNondet column indicates 1,3-DCP was not detected. The Solution column is repeated with an appropriate zero-one variable column. These data are contained in the DCP dataset.

DCP dataset (subset)

DCP	DNondet	Moisture	Solution	Solution2
0.17	0	8.14	water	0
0.25	0	6.23	water	0
0.22	0	4.56	NaHSO4	1
0.28	0	7.39	water	0
0.13	1	11.91	water	0
0.18	0	6.43	NaHSO4	1
0.13	1	6.97	water	0
0.18	0	5.48	NaHSO4	1
0.26	0	6.12	NaHSO4	1
0.13	1	5.42	NaHSO4	1

Procedure Options

This section describes the options available in this procedure.

Variables Tab

This panel specifies the probability distribution that is fit and the variables used in the analysis.

Response Variable

Response Variable

The values of this variable represent either the magnitude of a detected observations or detection limits, depending on the corresponding values of the Nondetection (Censor) Variable.

The values in this variable must be greater than zero. If the value is missing or non-positive, it is not used during the estimation phase.

Frequency Variable

Frequency Variable

This variable gives the count, or frequency, of the response displayed on that row. When omitted, each row receives a frequency of one. Frequency values should be positive integers. A frequency variable is often used to indicate the number of Nondetects.

Nondetection Variable

Nondetection (Censor) Variable

The values in this variable indicate whether the value of the Response Variable represents a nondetected (censored) observation or a detected observation. When a particular value of this variable indicates a Nondetect, the corresponding value of the Response Variable represents a lower detection limit.

These values may be text or numeric. The interpretation of these codes is specified by the 'Detected' and 'Not Detected' (Censored) options to the right of this option.

Only two values are used, the Detected value and the Not Detected value. The Unknown Censor option specifies what is to be done with values that do not match either the Detected value or the Not Detected value.

Rows with missing values (blanks) in this variable are omitted from the estimation phase, but results are shown in any reports that output predicted values.

Detected

When this value is encountered under the Nondetection (Censor) Variable it indicates that the value under the Response Variable was observed or detected. The value may be a number or a letter.

We suggest the letter 'D' or the number '0' when you are in doubt as to what to use.

A detected observation is one in which the value was measured exactly; for example, the concentration was such that the instrument was able to measure it.

Not Detected

When this value is encountered under the Nondetection (Censor) Variable it indicates that the value under the Response Variable was not actually observed (i.e., a nondetect) but represents a lower detection limit. That is, the observation is left-censored, and the actual value of the response is something below the detection limit.

The value may be a number or a letter. We suggest the letter 'N' or the number '1' when you are in doubt as to what to use.

A nondetect is a response in which the value was not measured exactly; for example, the concentration was such that the instrument was not able to measure it.

Independent Variables

X's: Independent Variables

Specify the independent variables. At least one independent variable must be specified here.

These variables may be thought of as additional variables for which statistical adjustment is desired. Discrete and/or continuous variables may be specified here. If discrete variables are to be specified, you should create and specify the appropriate number of indicator (dummy) variables. For example, if three groups are to be compared, two indicator variables will be needed to distinguish these groups.

Probability Distribution

Distribution

This option specifies the probability distribution of the residuals (errors). All results are based on the probability distribution specified here.

Alpha Level

Alpha Level

This is the value of alpha used in the calculation of confidence limits. For example, if you specify 0.04 here, then 96% confidence limits will be calculated.

Estimation Tab

The following options control the searching algorithms used during parameter estimation.

Estimation Options

Maximum Iterations

Many of the parameter estimation algorithms are iterative. This option assigns a maximum to the number of iterations used in any one algorithm. We suggest a value of at least 100. This should be large enough to let the algorithm converge, but small enough to avoid a large delay if convergence cannot be obtained. If the number of iterations reaches this amount, you should re-run your analysis with a larger value.

Minimum Relative Change

This value is used to control the iterative algorithms used in maximum likelihood estimation. When the relative change in all of the parameters is less than this amount, the iterative procedure is terminated.

Parameter Adjustment

Newton's method calculates a change for each parameter value at each step. Instead of taking the whole parameter change, this option lets you take only a fraction of the indicated change. For datasets that diverge, taking only partial steps may allow the algorithm to converge. In essence, the algorithm tends to over correct the parameter values. This factor allows you to dampen this over correction. We suggest a value of about 0.2. This may increase the number of iterations (and you will have to increase the Maximum Iterations accordingly), but it provides a greater likelihood that the algorithm will converge.

Starting Sigma

Specify a starting value for *S*, the standard deviation of the residuals (errors). Select '0 - Data' to calculate an appropriate value from the data. If convergence fails, try a different value.

Derivatives

This value specifies the machine precision value used in calculating numerical derivatives. Slight adjustments to this value can change the accuracy of the numerical derivatives (which impacts the variance/covariance matrix estimation).

Remember from calculus that the derivative is the slope calculated at a point along the function. It is the limit found by calculating the slope between two points on the function curve that are very close together. Numerical differentiation mimics this limit by calculating the slope between two function points that are very close together and then computing the slope. This value controls how close together these two function points are.

Numerical analysis suggests that this distance should be proportional to the machine precision of the computer. We have found that our algorithm achieves four-place accuracy in the variance-covariance matrix no matter what value is selected here (within reason). However, increasing or decreasing this value by two orders of magnitude may achieve six or seven place accuracy in the variance-covariance matrix. We have found no way to find the optimal value except trial and error.

Note that the parameter estimates do not seem to be influenced a great deal, only their standard errors.

Reports Tab

The following options control which reports are displayed and the format of those reports.

Select Reports

Data Summary Report ... Residual Report

Each of these options specifies whether the indicated report is calculated and displayed.

Report Options

Precision

Specify the precision of numbers in the report. A single-precision number will show seven-place accuracy, while a double-precision number will show thirteen-place accuracy. Note that the reports are formatted for single precision. If you select double precision, some numbers may run into others. Also, note that all calculations are performed in double precision regardless of which option you select here. Single precision is for reporting purposes only.

Variable Names

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

Value Labels

This option lets you select whether to display only values, only value labels, or both for values of the group variable. Use this option if you want to automatically attach labels to the values of the group variable (such as 1=Male, 2=Female, etc.). See the section on specifying *Value Labels* elsewhere in this manual.

Report Options – Decimal Places

Response and Probability Decimals

This option specifies the number of decimal places shown on reported response and probability values.

Plots Tab

These options control the attributes of the plots.

Select Plots

X - Y Plots ... X - Residual Plots

Each of these options specifies whether the indicated plot is displayed. Click the plot format button to change the plot settings.

Number Predicted

This options sets resolution of the plot along the horizontal axis. A value near 50 is usually adequate.

Example 1 – Nondetects-Data Regression

This section presents an example of how to perform a nondetects normal distribution regression. The DCP dataset that will be used was described above. Suppose the researchers wish to establish the relationship between percent moisture in the soil sample and 1,3-DCP concentration. Further, they wish to determine if there are differences in the two solutions used for determining 1,3-DCP concentrations.

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 Nondetects-Data Regression window.

1 Open the DCP dataset.

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

2 Open the Nondetects-Data Regression window.

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

3 Specify the variables.

- On the Nondetects Regression window, select the **Variables tab**.
- Double-click in the **Response Variable** box. This will bring up the variable selection window.
- Select **DCP** from the list of variables and then click **Ok**.
- Double-click in the **Nondetection (Censor) Variable** box. This will bring up the variable selection window.
- Select **DNondet** from the list of variables and then click **Ok**.
- Enter the values **0** and **1** for the **Detected** and **Not Detected** fields, respectively.
- Double-click in the **X's: Independent Variables** box. This will bring up the variable selection window.
- Select **Moisture** and **Solution2** (you can use the control key) from the list of variables and then click **Ok**.
- Set the **Distribution** to **Normal**.

4 Specify the reports.

- On the Nondetects Regression window, select the **Reports tab**.
- Check the boxes of all the report options.

5 Run the procedure.

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

Data Summary Section

Data Summary Section

Type of Observation	Rows	Count	Hours Minimum	Hours Maximum
Missing or Prediction	0			
Detected	39	39	0.140	0.350
Not Detected	13	13	0.130	0.130
Total (Nonmissing)	52	52	0.130	0.350

Means

Variable	Mean
DCP	0.2071154
Moisture	7.520385
Solution2	0.5769231

This report displays a summary of the data that were analyzed. Scan this report to determine if there are any obvious data-entry errors by double-checking the counts and the minimum and maximum.

The means given for each variable are for detected and nondetected rows combined.

Parameter Estimation Section

Maximum Likelihood Parameter Estimation Section

Parameter Name	Parameter Estimate	Standard Error	Z Value	Prob Level	Lower 95.0% C.L.	Upper 95.0% C.L.
Intercept	0.1821516	0.03317943	5.4899	0.0000	0.1171212	0.2471821
Moisture	-0.002174384	0.003558569	-0.6110	0.5412	-0.009149052	0.004800284
Solution2	0.051851	0.02308581	2.2460	0.0247	0.006603649	0.09709834
Sigma	0.07899635	0.009540257	8.2803	0.0000	0.06234603	0.1000933

Approximate R-Squared	
Log Likelihood	30.68732
Iterations	32

This report displays parameter estimates along with standard errors, significance tests, and confidence limits. Note that the significance levels and confidence limits all use large sample formulas. We suggest that you only use these results when the number of detected items is greater than twenty.

Parameter Estimates

These are the maximum likelihood estimates (MLE) of the parameters. They are the estimates that maximize the likelihood function. Details are found in Nelson (1990) pages 287 - 295.

Standard Error

The standard errors are the square roots of the diagonal elements of the estimated Variance Covariance matrix.

Z Value

The z value is equal to the parameter estimate divided by the estimated standard error. This ratio, for large samples, follows the normal distribution. It is used to test the hypothesis that the parameter value is zero. This value corresponds to the t value that is used in multiple regression.

Prob Level

This is the two-tailed p-value for testing the significance of the corresponding parameter. You would deem independent variables with small p-values (less than 0.05) important in the regression equation.

Upper and Lower 100(1-Alpha)% Confidence Limits

These are the lower and upper confidence limits for the corresponding parameters. They are large sample limits. They should be ignored when the number of detected items is less than thirty. For the regression coefficients *B*, the formulas are

$$CL_i = \hat{B}_i \pm z_{1-\alpha/2}\hat{\sigma}_{\hat{B}_i} \quad i = 0, \cdots, p$$

where $\hat{B}_i$ is the estimated regression coefficient, $\hat{\sigma}_{\hat{B}_i}$ is its standard error, and *z* is found from tables of the standard normal distribution.

For the estimate of sigma, the formula is

$$CL = \hat{S}\exp\left\{\frac{\pm z_{1-\alpha/2}\hat{\sigma}_{\hat{S}}}{\hat{S}}\right\}$$

Approximate R-Squared

R-Squared reflects the percent of variation in response explained by the independent variables in the model. A value near zero indicates a complete lack of fit, while a value near one indicates nearly a perfect fit.

This value is an 'approximate' R-squared because it is computed using the failed observations with regression coefficients which were based on all observations. The formula used is

$$R^2 = 1 - \frac{\sum_{k=1}^{n}\delta_k\left(y_k - \sum_{i=0}^{p}X_{ik}\hat{B}_i\right)^2}{\sum_{k=1}^{n}\delta_k\left(y_k - \bar{y}\right)^2}, \quad \bar{y} = \frac{\sum_{k=1}^{n}\delta_k y_k}{\sum_{k=1}^{n}\delta_k}$$

where δ_i is one if the observation was a failure, and zero otherwise. Approximate R-Squared values greater than one or less than zero are not reported.

Log Likelihood

This is the value of the log likelihood function. This is the value being maximized. It is often used as a goodness-of-fit statistic. You can compare the log likelihood value from the fits of your data to several distributions and select as the best fitting the one with the largest value.

Iterations

This is the number of iterations that were required to solve the likelihood equations. If this is greater than the maximum you specified, you will receive a warning message. You should then increase the Maximum Iterations and rerun the analysis.

Variance Covariance Matrix

Variance Covariance Matrix

	Intercept	Moisture	Solution2	Sigma
Intercept	0.001100874	-9.995691E-05	-0.0003854267	-8.733427E-06
Moisture	-9.995691E-05	1.266342E-05	9.760901E-06	-1.532926E-06
Solution2	-0.0003854267	9.760901E-06	0.0005329545	3.791392E-06
Sigma	-8.733427E-06	-1.532926E-06	3.791392E-06	9.10165E-05

This table gives an estimate of the asymptotic variance covariance matrix which is the inverse of the Fisher information matrix. The elements of the Fisher information matrix are calculated using numerical differentiation.

Residual Section

Residual Section

Row	(T) DCP	T	Predicted T	Raw Residual	Standardized Residual	Cox-Snell Residual
1	0.170	0.17	0.1644522	0.005547851	0.0702292	0.7507644
2	0.250	0.25	0.1686052	0.08139478	1.030361	1.887696
3	0.220	0.22	0.2240874	-0.004087442	-0.05174216	0.65271
4	0.280	0.28	0.1660829	0.1139171	1.442055	2.595032
5L	0.130	0.13	0.1562547	-0.02625472	-0.3323536	0.4617357
6	0.180	0.18	0.2200213	-0.04002135	-0.5066227	0.3655857
7L	0.130	0.13	0.1669962	-0.03699618	-0.4683277	0.3853318
8	0.180	0.18	0.222087	-0.04208701	-0.5327716	0.3525347
9	0.260	0.26	0.2206954	0.0393046	0.4975495	1.173118
10L	0.130	0.13	0.2222175	-0.09221748	-1.167364	0.1295755
11	0.170	0.17	0.1652349	0.004765073	0.06032016	0.7424418
12	0.330	0.33	0.2200648	0.1099352	1.391649	2.50086
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

This report displays the predicted value and residual for each row. The report provides predicted values for all rows with values for the independent variables. Hence, you can add rows of data with missing time values to the bottom of your database and obtain the predicted values for them from this report. The report also allows you to obtain predicted values for nondetects observations.

You should ignore the residuals for nondetects observations, since the residual is calculated as if the response value were an actual response.

Row

This is the number of the observation being reported on. Nondetects observations have a letter (L for left-censored) appended to the row number.

(T) Response

This is the original value of the dependent variable.

Predicted T

This is the predicted transformed value of the dependent variable (usually time). Note that *y* depends on the distribution being fit. For the Weibull, exponential, lognormal, and log-logistic distributions, the *y* is *ln(t)*. For the lognormal10 distribution, *y* is *log(t)*. For the extreme value, normal, and logistic distributions, *y* is *t*. The formula for *y* is

$$\hat{y} = \sum_{i=0}^{p} x_i B_i$$

Raw Residual

This is the residual in the *y* scale. The formula is

$$r_k = y_k - \sum_{i=0}^{p} x_i B_i$$

Note that the residuals of censored observations are not directly interpretable, since there is no obvious value of y. The row is displayed so that you can see the predicted value for this censored observation.

Standardized Residual

This is the residual standardized by dividing by the standard deviation. The formula is

$$r'_k = \frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}$$

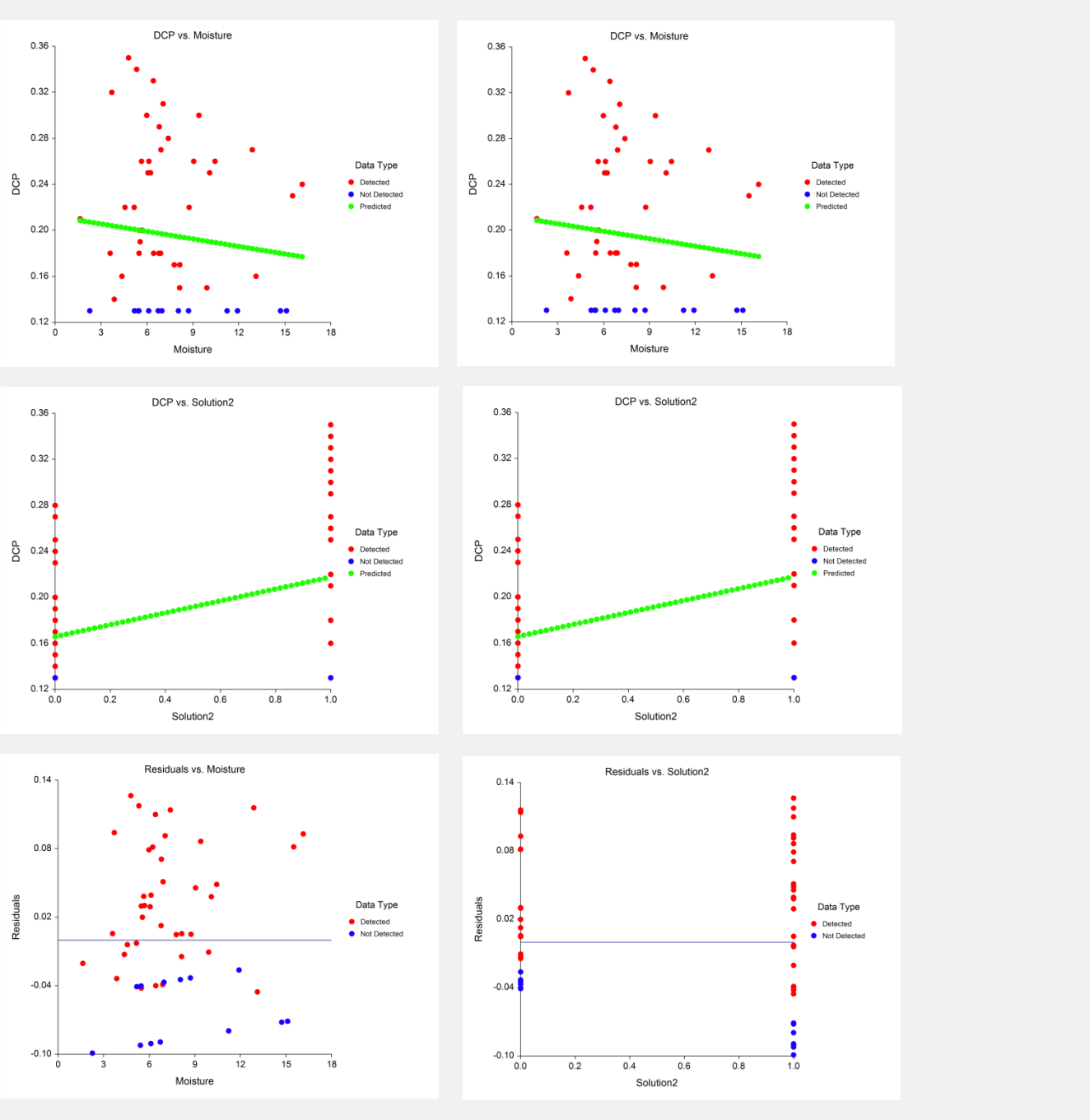
Cox-Snell Residual

The Cox-Snell residual is defined as

$$r_k'' = -\log\left\{1 - F\left(\frac{y_k - \sum_{i=0}^{p} x_i B_i}{\hat{S}}\right)\right\}$$

Here again, the residual does not have a direct interpretation for censored values.

X-Y, X-Trans(Y), and X-Resid Plots



The first two pairs of plots show the data values from which the analysis was run. The plots on the left show the response versus the independent variable in the original scale. The plots on the right show the response versus the independent variable in the transformed metric (for the normal distribution there is no transformation, so that the plots on the left and right are the same). The third pair of plots shows the residuals in the transformed scale (again, here, there is no transformation because the normal distribution is used).

Example 2 – Validation using Helsel (2005)

On pages 134-138, Helsel (2005) presents an example of using nondetects lognormal distribution regression to compare zinc concentrations among two zones. The estimate of the zone effect is given as -0.257408. The corresponding Z value and probability level are -1.60 and 0.110, respectively. The Log-likelihood is -407.296. The data are contained in the Zinc dataset.

These data can be run in this procedure to see that ***NCSS*** gets the same results. 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 Nondetects-Data Regression window.

1 Open the Zinc dataset.

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

2 Open the Nondetects-Data Regression window.

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

3 Specify the variables.

- On the Nondetects Regression window, select the **Variables tab**.
- Double-click in the **Response Variable** box. This will bring up the variable selection window.
- Select **Zinc** from the list of variables and then click **Ok**.
- Double-click in the **Nondetection (Censor) Variable** box. This will bring up the variable selection window.
- Select **ZNondet** from the list of variables and then click **Ok**.
- Enter the values **0** and **1** for the **Detected** and **Not Detected** fields, respectively.
- Double-click in the **X's: Independent Variable** box. This will bring up the variable selection window.
- Select **Zone** from the list of variables and then click **Ok**. Note that the values of Zone are appropriate for this problem.
- Set the **Distribution** to **Lognormal**.

4 Specify the estimation parameters.

- On the Nondetects Regression window, select the **Estimation tab**.
- Set **Derivatives** to **0.0005**.

5 Specify the reports.

- On the Nondetects Regression window, select the **Reports tab**.
- Uncheck all boxes except **Parameter Report**.

6 Run the procedure.

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

Parameter Estimation Section

Maximum Likelihood Parameter Estimation Section

Parameter Name	Parameter Estimate	Standard Error	Z Value	Prob Level	Lower 95.0% C.L.	Upper 95.0% C.L.
Intercept	2.723747	0.1203683	22.6284	0.0000	2.48783	2.959665
Zone	**-0.2574348**	**0.1612933**	**-1.5961**	**0.1105**	-0.5735639	0.0586942
Sigma	0.8428832	6.194304E-02	13.6074	0.0000	0.7298154	0.9734681

Approximate R-Squared	
Log Likelihood	**-407.2973**
Iterations	39

The results of ***NCSS*** match those of Helsel (2005) to several decimal places.