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

$$Y = B_0 + B_1X_1 + B_2X_2 + \cdots + Se$$

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

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

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

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.

Setup

To run this example, complete the following steps:

1 Open the DCP example dataset

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

2 Specify the Nondetects-Data Regression procedure options

- Find and open the **Nondetects-Data 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 file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Variables Tab

Response Variable	**DCP**
Nondetection (Censor) Variable	**DNondet**
Detected	**0**
Not Detected	**1**
X's Independent Variables	**Moisture,Solution2**
Distribution	**Normal**

Reports Tab

Data Summary Report	**Checked**
Parameter Report	**Checked**
Information Matrix	**Checked**
Residual Report	**Checked**

3 Run the procedure

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

Data Summary Section

```
Data Summary Section
────────────────────────────────────────────────────────────────────────────────
Type of                                              DCP         DCP
Observation               Rows      Count        Minimum     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   Parameter     Standard                     Prob    Lower          Upper
Name        Estimate      Error            Z Value    Level    95.0% C.L.     95.0% C.L.
─────────────────────────────────────────────────────────────────────────────────────────
Intercept   0.182152      0.03229247        5.6407   0.0000    0.1188599      0.2454441
Moisture    -0.002174445  0.003432803      -0.6334   0.5265    -0.008902616   0.004553725
Solution2   0.05185108    0.02305408        2.2491   0.0245    0.006665923    0.09703624
Sigma       0.07899634    0.009541152       8.2795   0.0000    0.06234464     0.1000956
─────────────────────────────────────────────────────────────────────────────────────────

Model Estimation Information
────────────────────────────────────
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

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

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

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

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	0.001042804	-9.281725E-05	-0.0003765752	-9.828295E-06
Moisture	-9.281725E-05	1.178414E-05	8.688428E-06	-1.399243E-06
Solution2	-0.0003765752	8.688428E-06	0.0005314904	3.995561E-06
Sigma	-9.828295E-06	-1.399243E-06	3.995561E-06	9.103358E-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

	(T)		Predicted	Raw	Standardized	Cox-Snell
Row	DCP	T	T	Residual	Residual	Residual
1	0.170	0.17	0.164452	0.005547956	0.07023054	0.7507655
2	0.250	0.25	0.1686052	0.08139476	1.030361	1.887696
3	0.220	0.22	0.2240876	-0.00408764	-0.05174468	0.6527081
4	0.280	0.28	0.1660829	0.1139171	1.442056	2.595034
5L	0.130	0.13	0.1562544	-0.02625439	-0.3323494	0.4617382
6	0.180	0.18	0.2200214	-0.04002143	-0.5066239	0.3655851
7L	0.130	0.13	0.1669962	-0.03699614	-0.4683274	0.385332
8	0.180	0.18	0.2220871	-0.04208715	-0.5327735	0.3525338
9	0.260	0.26	0.2206955	0.03930449	0.4975483	1.173116
10L	0.130	0.13	0.2222176	-0.09221762	-1.167366	0.129575
11	0.170	0.17	0.1652348	0.004765155	0.06032122	0.7424427
12	0.330	0.33	0.2200649	0.1099351	1.391648	2.500859
13	0.200	0.2	0.1697794	0.03022056	0.3825566	1.0469
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.

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

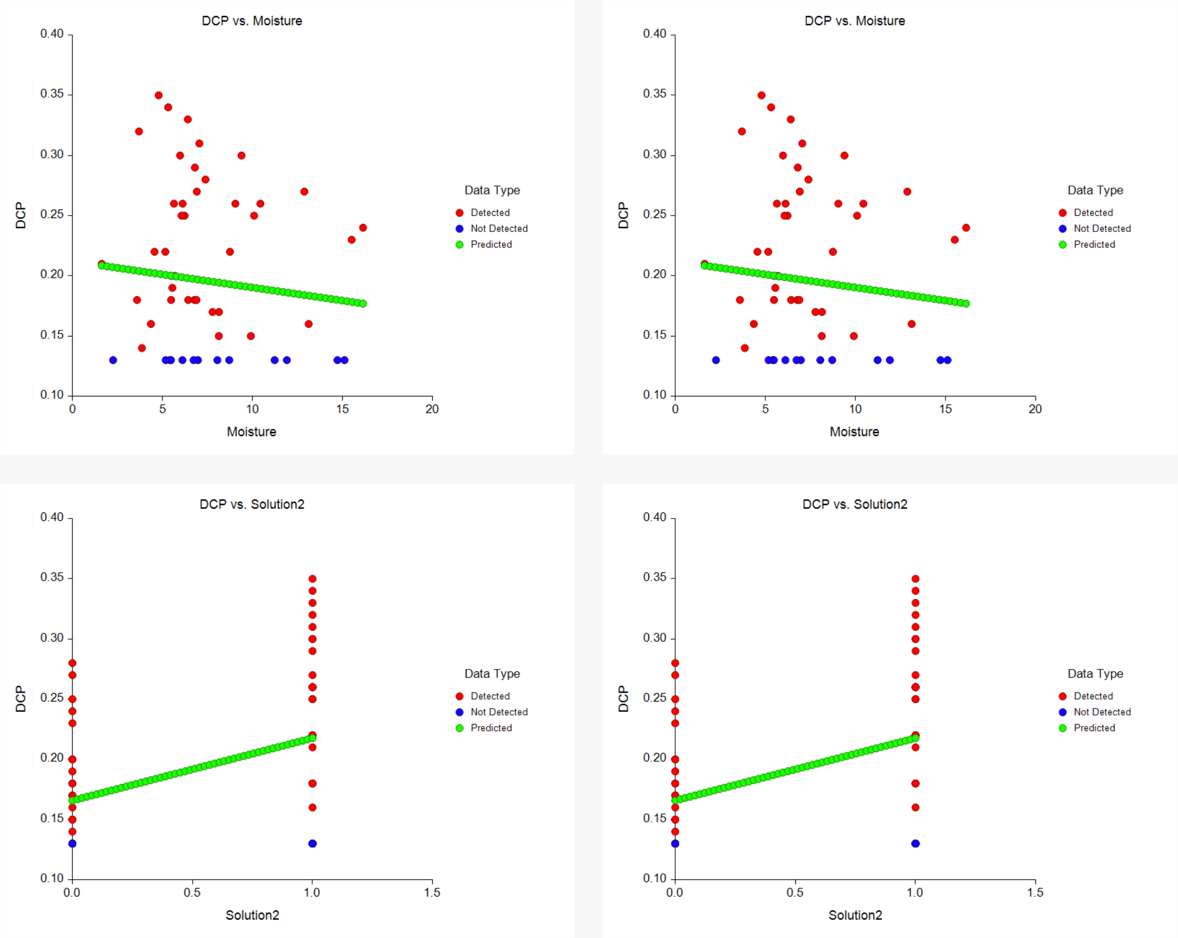
The Cox-Snell residual is defined as

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

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

X vs Y and X vsTrans(Y) Plots

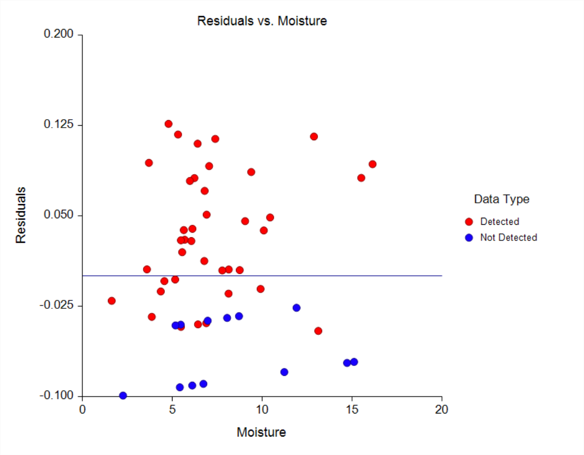
X's vs Y and X's vs Transformed(Y) Plots

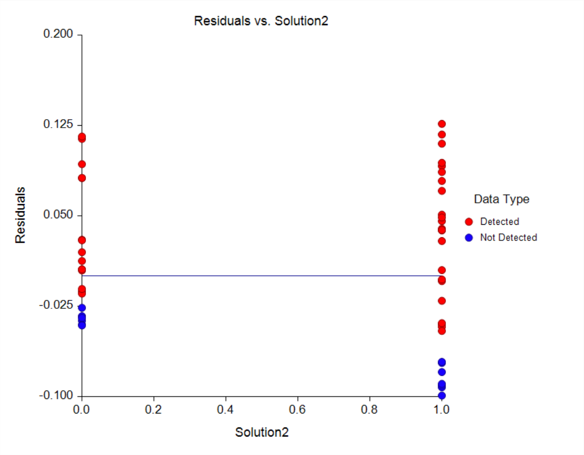


The two pairs of plots show the data values from which the analysis was run. The plots on the left show the response versus the independent variable in the original scale. The plots on the right show the response versus the independent variable in the transformed metric (for the normal distribution there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

X’s vs Residuals Plots

X's vs Residuals Plots





These plots show the residuals in the transformed scale.

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.

Setup

To run this example, complete the following steps:

1 Open the Zinc example dataset

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

2 Specify the Nondetects-Data Regression procedure options

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

Variables Tab

Response Variable	**Zinc**
Nondetection (Censor) Variable	**ZNondet**
Detected	**0**
Not Detected	**1**
X's Independent Variables	**Zone**
Distribution	**Lognormal**

Estimation Tab

Derivatives	**0.0005**

Reports Tab

Data Summary Report	**Unchecked**
Parameter Report	**Checked**
Information Matrix	**Unchecked**
Residual Report	**Unchecked**

3 Run the procedure

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

Parameter Estimation Section

```
Maximum Likelihood Parameter Estimation Section
──────────────────────────────────────────────────────────────────────────────────────────────────
Parameter     Parameter      Standard                     Prob     Lower         Upper
Name          Estimate       Error           Z Value      Level    95.0% C.L.    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    0.05869421
Sigma         0.8428832      0.06194304      13.6074      0.0000   0.7298154     0.9734681

Model Estimation Information
Approximate R-Squared
Log-Likelihood            -407.2973
Iterations                39
──────────────────────────────────────────────────────────────────────────────────────────────────
```

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