Chapter 564

Conditional Logistic Regression

Introduction

Logistic regression analysis studies the association between a binary dependent variable and a set of independent (explanatory) variables using a logit model (see Logistic Regression). *Conditional logistic regression* (CLR) is a specialized type of logistic regression usually employed when case subjects with a particular condition or attribute are each matched with *n* control subjects without the condition. In general, there may be 1 to *m* cases matched with 1 to *n* controls. However, the most common design is 1:1 matching, followed by 1:*n* matching in which *n* varies from 1 to 5.

The details of CLR are beyond the scope of this introduction. However, we will mention several facts:

1. CLR provides estimates of regression coefficients associated with independent variables (often called covariates) that vary within at least one strata. Likewise, CLR does not provide estimates for estimates for any regression coefficients associated with independent variables the do not vary within strata.
2. As the study sample size increases, the number of strata (clusters) increases at the same rate.
3. The stratum indicator variable is in the model, but no stratum by stratum output is shown.
4. CLR can be used when the matched sets have differing numbers of cases and controls.

Further Reading

Several books provide in some coverage of CLR. Hosmer and Lemeshow (2000) devote two chapters to this subject. Kleinbaum and Klein (2010) provide a somewhat more elementary discussion of the topic.

The Conditional Logistic Regression Model

If there are *S* strata (matched sets) and *p* independent variables (*x*'s), the CLR model is

$$logit(p) = \alpha_1 + \alpha_2 z_2 + \cdots + \alpha_S z_S + \beta_1 x_1 + \cdots + \beta_p x_p$$

where the *z*'s are binary indicator variables for each strata (note that there are only $S - 1$ *z* variables needed), the α's are the regression coefficients associated with the stratum indicator variables, the *x*'s are the covariates, and the β's are the population regression coefficients to be estimated.

The CLR algorithm estimates the β's, but not the α's. These can be used to analyze the odds ratios of each covariate adjusted for the others.

Maximum Likelihood Estimation

The estimation procedure used in NCSS makes use of the relationship between CLR and Cox Regression. This relationship allows us to estimate and test the significance of the β's using the Cox Regression calculation engine. However, it does not allow the calculation of predicted values and residuals.

As discussed in the Cox Regression chapter, there are two methods available for approximating the likelihood equation when there are ties present: Breslow and Efron. The Breslow method is often used as the default in other statistical packages. It is recommended for 1:1 and 1:*n* matching. Efron's method is general taken to be more accurate, but a little slower to compute. It is recommended for *m:n* matching where *m* is greater than one.

Statistical Tests and Confidence Intervals

Inferences about the regression coefficients are of interest. The inference procedures in Cox regression continue to be valid as long as the sample sizes are adequate. Two tests are available for testing the significance of one or more independent variables in a regression: the likelihood ratio test and the Wald test. Simulation studies usually show that the likelihood ratio test performs better than the Wald test. However, the Wald test is still used to test the significance of individual regression coefficients because of its ease of calculation.

These two testing procedures will be described next.

Likelihood Ratio and Deviance

The *Likelihood Ratio* test statistic is -2 times the difference between the log likelihoods of two models, one of which is a subset of the other. The distribution of the LR statistic is closely approximated by the chi-square distribution for large sample sizes. The degrees of freedom (DF) of the approximating chi-square distribution is equal to the difference in the number of regression coefficients in the two models. The test is named as a ratio rather than a difference since the difference between two log likelihoods is equal to the log of the ratio of the two likelihoods.

The likelihood ratio test is the test of choice in Cox regression. Various simulation studies have shown that it is more accurate than the Wald test in situations with small to moderate sample sizes. In large samples, it performs about the same. Unfortunately, the likelihood ratio test requires more calculations than the Wald test, since it requires the fitting of two maximum-likelihood models.

Deviance

When the full model in the likelihood ratio test statistic is the saturated model, *LR* is referred to as the *deviance*. A saturated model is one which includes all possible terms (including interactions) so that the predicted values from the model equal the original data. The formula for the deviance is

$$D = -2\left[L_{\text{Reduced}} - L_{\text{Saturated}}\right]$$

The deviance in Cox regression is analogous to the residual sum of squares in multiple regression. In fact, when the deviance is calculated in multiple regression, it is equal to the sum of the squared residuals.

The change in deviance, ΔD, due to excluding (or including) one or more variables is used in Cox regression just as the partial *F* test is used in multiple regression. Many texts use the letter *G* to represent ΔD. Instead of using the *F* distribution, the distribution of the change in deviance is approximated by the chi-square distribution. Note that since the log likelihood for the saturated model is common to both deviance values, ΔD can be calculated without actually fitting the saturated model. This fact becomes very important during subset selection.

The formula for ΔD for testing the significance of the regression coefficient(s) associated with the independent variable $X1$ is

$$\begin{aligned}\Delta D_{X1} &= D_{\text{without } X1} - D_{\text{with } X1} \\ &= -2\left[L_{\text{without } X1} - L_{\text{Saturated}}\right] + 2\left[L_{\text{with } X1} - L_{\text{Saturated}}\right] \\ &= -2\left[L_{\text{without } X1} - L_{\text{with } X1}\right]\end{aligned}$$

Note that this formula looks identical to the likelihood ratio statistic. Because of the similarity between the change in deviance test and the likelihood ratio test, their names are often used interchangeably.

Wald Test

The Wald test will be familiar to those who use multiple regression. In multiple regression, the common *t*-test for testing the significance of a particular regression coefficient is a Wald test. In Cox regression, the Wald test is calculated in the same manner. The formula for the Wald statistic is

$$z_j = \frac{b_j}{s_{b_j}}$$

where s_{b_j} is an estimate of the standard error of b_j provided by the square root of the corresponding diagonal element of the covariance matrix, $V\left(\hat{\beta}\right) = I^{-1}$.

With large sample sizes, the distribution of z_j is closely approximated by the normal distribution. With small and moderate sample sizes, the normal approximation is described as 'adequate' at best.

The Wald test is used in ***NCSS*** to test the statistical significance of individual regression coefficients.

Confidence Intervals

Confidence intervals for the regression coefficients are based on the Wald statistics. The formula for the limits of a $100(1-\alpha)\%$ two-sided confidence interval is

$$b_j \pm \left|z_{\alpha/2}\right| s_{b_j}$$

R^2

Hosmer and Lemeshow (1999) indicate that at the time of the writing of their book, there is no single, easy to interpret measure in Cox regression that is analogous to R^2 in multiple regression. They indicate that if such a measure "must be calculated" they would use

$$R_p^2 = 1 - \exp\left[\frac{2}{n}\left(L_0 - L_p\right)\right]$$

where L_0 is the log likelihood of the model with no covariates, n is the number of observations (censored or not), and L_p is the log likelihood of the model that includes the covariates.

Subset Selection

Subset selection refers to the task of finding a small subset of the available regressor variables that does a good job of predicting the dependent variable. Because Cox regression must be solved iteratively, the task of finding the best subset can be time consuming. Hence, techniques which look at all possible combinations of the regressor variables are not feasible. Instead, algorithms that add or remove a variable at each step must be used. Two such searching algorithms are available in this module: forward selection and forward selection with switching.

Before discussing the details of these two algorithms, it is important to comment on a couple of issues that can come up. The first issue is what to do about the binary variables that are generated for a categorical independent variable. If such a variable has six categories, five binary variables are generated. You can see that with two or three categorical variables, a large number of binary variables may result, which greatly increases the total number of variables that must be searched. To avoid this problem, the algorithms used here search on model terms rather than on the individual variables. Thus, the whole set of binary variables associated with a given term are considered together for inclusion in, or deletion from, the model. It's all or none. Because of the time consuming nature of the algorithm, this is the only feasible way to deal with categorical variables. If you want the subset algorithm to deal with them individually, you can generate the set of binary variables manually and designate them as Numeric Variables.

Hierarchical Models

A second issue is what to do with interactions. Usually, an interaction is not entered in the model unless the individual terms that make up that interaction are also in the model. For example, the interaction term A*B*C is not included unless the terms A, B, C, A*B, A*C, and B*C are already in the model. Such models are said to be *hierarchical*. You have the option during the search to force the algorithm to only consider hierarchical models during its search. Thus, if C is not in the model, interactions involving C are not even considered. Even though the option for non-hierarchical models is available, we recommend that you only consider hierarchical models.

Forward Selection

The method of forward selection proceeds as follows.

1. Begin with no terms in the model.
2. Find the term that, when added to the model, achieves the largest value of *R*-squared. Enter this term into the model.
3. Continue adding terms until a preset limit on the maximum number of terms in the model is reached.

This method is comparatively fast, but it does not guarantee that the best model is found except for the first step when it finds the best single term. You might use it when you have a large number of observations so that other, more time consuming methods, are not feasible, or when you have far too many possible regressor variables and you want to reduce the number of terms in the selection pool.

Forward Selection with Switching

This method is similar to the method of Forward Selection discussed above. However, at each step when a term is added, all terms in the model are switched one at a time with all candidate terms not in the model to determine if they increase the value of *R*-squared. If a switch can be found, it is made and the candidate terms are again searched to determine if another switch can be made.

When the search for possible switches does not yield a candidate, the subset size is increased by one and a new search is begun. The algorithm is terminated when a target subset size is reached or all terms are included in the model.

Discussion

These algorithms usually require two runs. In the first run, you set the maximum subset size to a large value such as 10. By studying the Subset Selection reports from this run, you can quickly determine the optimum number of terms. You reset the maximum subset size to this number and make the second run.

Data Structure

CLR data sets require at least three columns: one to hold the match group, one to hold the event of interest (case or control identifier), and one to hold an independent variable. The table below shows part of the Kleinbaum MI dataset. These data are discussed in Kleinbaum and Klein (2010). The variables in the dataset are

Match Match identification number
Person Person identification number (not used)
MI Myocardial infarction status (case or yes = 1; control or no = 0)
SMK Smoker (yes = 1; no = 0)
SBP Systolic blood pressure
ECG Electrocardiogram abnormality status (yes = 1; no = 0)

Kleinbaum MI dataset (a subset)

Match	Person	MI	SMK	SBP	ECG
1	1	1	0	160	1
1	2	0	0	140	0
1	3	0	0	120	0
2	4	1	0	160	1
2	5	0	0	140	0
2	6	0	0	120	0
3	7	1	0	160	0
3	8	0	0	140	0
3	9	0	0	120	0
4	10	1	0	160	0
4	11	0	0	140	0
4	12	0	0	120	0
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

Procedure Options

This section describes the options available in this procedure.

Variables, Model Tab

This panel lets you designate which variables and model are used in the analysis.

Variables

Match Group

This column contains the Match Groups, Match Sets, or Strata identifier values. Each match set consists m+n rows. Each set must include 1 or more (m) cases and 1 or more (n) controls. The number of cases and/or controls does not have to be the same in each set.

The group identifiers may be text or numeric values.

Ties Method

Choose one of two methods to approximate the exact likelihood in the presence of ties.

In the discussion next, the phrase ‘m:n’ refers to a match set that contains *m* cases and *n* controls.

- **Breslow**
  This method is suggested for designs with ‘1:1’ or ‘1:n’ matching.

- **Efron**
  This method is suggested for designs with ‘m:n’ matching, where *m* is two or more.

Event

The values in this column indicate whether the row is a case or a control.

The values may be text or numeric, but usually a ‘1’ is used for a case and a ‘0’ is used for a control. The actual interpretation of values is defined in the next three boxes.

Rows with missing values (blanks) in this variable are ignored.

Case

This value identifies those values of the Event variable that represent cases. The value may be a number or a letter. We suggest the number ‘1’ when you are in doubt as to what to use.

A case observation is one in which the event of interest occurred.

Control

This value identifies those values of the Event variable that represent controls. The value may be a number or a letter. We suggest the number ‘0’ when you are in doubt as to what to use.

A control observation is one in which the event of interest did not occur.

Other

This option specifies what the program is to assume about observations whose event value is not equal to either the Case value or the Control value. Note that rows with missing event values are always treated as missing and omitted.

- **Control**
  Rows with other event values are assumed to have been censored.

- **Case**
  Rows with other event values are assumed to have failed.

- **Missing**
  Rows with other event values are assumed to be missing and that row is omitted from the analysis.

Numeric X's

Specify one or more quantitative independent variables. It is not necessary to specify any of these variables if at least one Categorical X variable is specified.

- **Numeric (Quantitative) Variables**
  We consider a variable as 'numeric' if its values are numbers that are at least ordinal. This includes binary (indicator) variables. However, when creating powers of variables, binary indicator variables should be selected in the Categorical X's box.

- **Independent or Explanatory Variables**
  The X's are called the independent, explanatory, regressor, or predictor variables. We say that the dependent variable Y depends on these independent variables, the X's.

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

Categorical X's

Specify categorical independent variables here. It is not necessary to specify any Categorical X variables if at least one Numeric X variable is specified.

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 variable.

- **Categorical**
  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.

- **Independent or Explanatory Variables**
  The X's are called the independent, explanatory, regressor, or predictor variables. We say that the dependent variable Y depends on these independent variables, the X's.

- **Recoding Categories to Numeric Values**
  NCSS automatically generates internal numeric variables from categorical variables 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.

- **Specifying how categorical variables are recoded to numeric variables**

  The complete syntax for specifying a categorical variable is *VarName(Ctype;RefValue)* where *VarName* is the variable name from the database, *CType* is the method used to generate the variable, and *RefValue* (if needed) 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 often most frequently occurring value.

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

- **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

- **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

- **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

- **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

- **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.

Frequencies

This is an optional variable containing the frequency (observation count) for each row. Usually, you would leave this option blank and let each row receive the default frequency of one.

If your data have already been summarized, this option lets you specify how many actual rows each physical row represents.

Regression Model

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**
  This option generates a model in which each variable is represented by a single model term. No cross-products, interactions, or powers are added. Use this option when you want to use the variables you have specified, but you do not want to generate other terms.

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

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

  A + B + C

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

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

- **Up to 3-Way**
  All individual variables, two-way interactions, three-way interactions, squares of numeric variables, and cubes of numeric variables are included in the model. For example, if you have three numeric, independent variables 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 variables 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 variables, 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 variables and their cross-products.

  For example, if you have four categorical variables 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 variables. A saturated model (all terms and their interactions) is generated. This requires a dataset with no missing categorical-variable combinations (you can have unequal numbers of observations for each combination of the categorical variables). No squares, cubes, etc. are generated.

  For example, if you have three independent variables 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 Model option discusses the interpretation of this model.

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

Center X's

Indicate whether to center (subtract the mean) the independent variables. This usually improves the stability of the algorithm and is recommended.

Centering does not change the values of the regression coefficients, except that the algorithm might provide slightly different results because of better numerical stability.

The options are available:

- **Unchecked**
  The data are not centered.

- **Checked**
  All variables, both numeric and binary, are centered.

Replace Custom Model 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 Model

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

Interactions

An interaction expresses the combined relationship between two or more variables and the dependent variable by creating a new variable that is the product of the variables. The interaction between two numeric variables is generated by multiplying them. The interaction between to categorical variables is generated by multiplying each pair of indicator variables. The interaction between a numeric variable and a categorical variable is created by generating all products between the numeric variable and the indicator variables generated from the categorical variable.

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 variable. 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

Subset Selection

Search Method

This option specifies the subset selection algorithm used to reduce the number of independent variables that used in the regression model. Note that since the solution algorithm is iterative, the selection process can be very time consuming. The Forward algorithm is much quicker than the Forward with Switching algorithm, but the Forward algorithm does not usually find as good of a model.

Also note that in the case of categorical independent variables, the algorithm searches among the original categorical variables, not among the generated individual binary variables. That is, either all binary variables associated with a particular categorical variable are included or not—they are not considered individually.

Hierarchical models are such that if an interaction is in the model, so are the terms that can be derived from it. For example, if A*B*C is in the model, so are A, B, C, A*B, A*C, and B*C. Statisticians usually adopt hierarchical models rather than non-hierarchical models. The subset selection procedure can be made to consider only hierarchical models during its search.

The subset selection options are:

- **None – No Search is Conducted**

  No subset selection is attempted. All specified independent variables are used in the regression equation.

- **(Hierarchical) Forward**

  With this algorithm, the term with the largest log likelihood is entered into the model. Next, the term that increases the log likelihood the most is added. This selection is continued until all the terms have been entered or until the maximum subset size has been reach.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for selection. For example, the interaction term A*B will not be considered unless both A and B are already in the model.

  When using this algorithm, you must make one run that allows a large number of terms to find the appropriate number of terms. Next, a second run is made in which you decrease the maximum terms in the subset to the number after which the log likelihood does not change significantly.

- **(Hierarchical) Forward with Switching**

  This algorithm is similar to the Forward algorithm described above. The term with the largest log likelihood is entered into the regression model. The term which increases the log likelihood the most when combined with the first term is entered next. Now, each term in the current model is removed and the rest of the terms are checked to determine if, when they are used instead, the likelihood function is increased. If a term can be found by this switching process, the switch is made and the whole switching operation is begun again. The algorithm continues until no term can be found that improves the likelihood. This model then becomes the best two-term model.

  Next, the subset size is increased by one, the best third term is entered into the model, and the switching process is repeated. This process is repeated until the maximum subset size is reached. Hence, this model finds the optimum subset for each subset size. You must make one run to find an appropriate subset size by looking at the change in the log likelihood. You then reset the maximum subset size to this value and rerun the analysis.

  If hierarchical models are selected, only those terms that will keep the model hierarchical are candidates for addition or deletion. For example, the interaction term A*B will not be considered unless both A and B are already in the model. Likewise, the term A cannot be removed from a model that contains A*B.

Stop search when number of terms reaches

Once this number of terms has been entered into the model, the subset selection algorithm is terminated. Often you will have to run the procedure twice to find an appropriate value. You would set this value high for the first run and then reset it appropriately for the second run, depending upon the values of the log likelihood.

Note that the intercept is counted in this number.

Iteration Tab

This panel lets you control the maximum likelihood estimation algorithm.

Iteration Options

These options control the number of iterations used while the algorithm is searching for the maximum likelihood solution.

Maximum Iterations

This option specifies the maximum number of iterations used while finding a solution. If this number is reached, the procedure is terminated prematurely. This is used to prevent an infinite loop and to reduce the running time of lengthy variable selection runs.

Usually, only 20 iterations are needed. In fact, many runs converge in about 7 or 8 iterations.

During a variable selection run, it may be advisable reset this value to 4 or 5 to speed up the variable selection. Usually, the last few iterations make little difference in the estimated values of the regression coefficients.

Convergence Zero

This option specifies the convergence target for the maximum likelihood estimation procedure. The algorithm finds the maximum relative change of the regression coefficients. If this amount is less than the value set here, the maximum likelihood procedure is terminated.

For large datasets, you might want to increase this value to about 0.0001 so that fewer iterations are used, thus decreasing the running time of the procedure.

Regression Coefficient Starting Values

These options control the starting regression coefficient values (the B’s).

Start B’s at

Select a starting value (or enter a list of individual starting values) for the regression coefficients (the B’s).

Although the B’s can be any numeric value, the typical range is between -1 and 1. So starting values in this range usually allow the algorithm to converge to a useful solution.

The Cox regression algorithm solves for the maximum likelihood estimates of the regression coefficients by the iterative Newton-Raphson algorithm. This algorithm begins at a set of starting values for the regression coefficients and, at each iteration, modifies the B’s in a way that leads to a local maximum of the likelihood function. Sometimes the algorithm does not converge to the global maximum, so a different set of starting values must be tried.

Typically, a ‘good’ solution is one with all B’s less than 50 in absolute value. If your solution has one or more B’s that are large (over 10,000), you should rerun the algorithm with a different set of starting values.

List of Starting B’s (*If Start B’s at = “List”*)

Enter a list of starting values. If not enough values are entered, the last value will be used over and over.

Although the B's can be any numeric value, the typical range is between -1 and 1. So starting values in this range usually allow the algorithm to converge to a useful solution.

Reports Tab

The following options control which reports are displayed.

Alpha

Alpha Level

Alpha is the significance level used in the hypothesis tests. One minus alpha is the confidence level of the confidence intervals. A value of 0.05 is most commonly used. This corresponds to a chance of error of 1 in 20. 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 0.001 to 0.20.

Select Reports – Summaries

Run Summary

Indicate whether to display this summary report.

Select Reports – Subset Selection

Subset Selection - Summary and Subset Selection - Detail

Indicate whether to display these subset selection reports.

Select Reports – Estimation

Regression Coefficients ... C.L. of Regression Coefficients

Indicate whether to display these estimation reports.

Select Reports – Goodness-of-Fit

Log-Likelihood and Deviance

Indicate whether to display this report.

Report Options Tab

Variable Names

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

Stagger label and output if label length is ≥

The names of the indicator variables can be too long to fit in the space provided. If the name contains more characters than the number specified here, only the name is shown on the first line of the report and the rest of the output is placed on the next line.

Enter *1* when you want the each variable's results printed on two lines.

Enter *100* when you want each variable's results printed on a single line.

Report Options – Decimal Places

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. This is for reporting purposes only.

b(i) ... S.E. of b(i)

These options specify the number of decimal places shown on the reports for the indicated values.

Example 1 – Conditional Logistic Regression Analysis and Validation

This section presents an example of how to run a CLR. The data used are found in the Kleinbaum MI dataset. These data are from a matched case control study reported in Kleinbaum and Klein (2010). The purpose of this analysis is study the relationship between myocardial infarction and the covariates smoking, blood pressure, and electrocardiogram status. 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 Conditional Logistic Regression window.

Kleinbaum and Klein (2010) present the results of fitting this model. They obtained the following parameter estimates for SMK, SBP, and ECG: 0.7291, 0.0456, and 1.5993.

1 Open the Kleinbaum MI dataset.

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

2 Open the Conditional Logistic Regression window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Conditional Logistic Regression** procedure
- On the menus, select **File**, then **New Template**. This will load the default template.

3 Specify the variables.

- On the Conditional Logistic Regression window, select the **Variables, Model tab**.
- Enter **Match** in the **Match Group** variable box.
- Set the **Ties Method** to **Breslow**.
- Enter **MI** in the **Event** variable box.
- Enter **SMK,SBP,ECG** in the **Numeric X's** variables box.

4 Run the procedure.

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

Run Summary

Parameter	Value	Parameter	Value
Rows Read	117	Match Group Variable	Match
Rows Filtered Out	0	Number of Match Groups	39
Rows Missing X's	0	Event Variable	MI
Rows Processed	117	Case Value (Count)	1 (39)
Sum of Frequencies	117	Control Value (Count)	0 (78)
Independent Variables Available	3	Frequency Variable	None
Number of DF's in Model	3	Subset Method	None
Iterations Used	20	Convergence Criterion	1E-06
Maximum Iterations Allowed	20	Achieved Convergence	2.792446E-09
Log Likelihood	-174.6244	Completion Status	Normal completion
Deviance	349.2487	Starting B's	0

This report summarizes the characteristics of the dataset and provides useful information about the reports to follow. It should be studied to make sure that the data were read in properly and that the estimation algorithm terminated normally. We will only discuss those parameters that need special explanation.

Rows Read

This is the number of rows processed during the run. Check this count to make certain it agrees with what you anticipated.

Iterations

This is the number of iterations used by the maximum likelihood procedure. This value should be compared against the value of the Maximum Iterations option to see if the iterative procedure terminated early.

Achieved Convergence

This is the maximum of the relative changes in the regression coefficients on the last iteration. If this value is less than the Convergence Criterion, the procedure converged normally. Otherwise, the specified convergence precision was not achieved.

Note that coefficients near machine zero (see Options) are not included in the convergence test.

Log Likelihood

This is the log likelihood of the model.

Regression Coefficients and Significance Tests

Regression Coefficients and Wald Z Tests
Event Variable: MI
Match Group Variable: Match

Independent Variable	Regression Coefficient b(i)	Standard Error Sb(i)	Wald Z-Value H0: β=0	Wald Prob Level	Odds Ratio Exp(b(i))	Mean
SMK	0.7291	0.5613	1.299	0.1940	2.073	0.28
SBP	0.0456	0.0152	2.994	0.0028	1.047	136.41
ECG	1.5993	0.8534	1.874	0.0609	4.949	0.21

This report displays the results of the estimation. You can check that these regression coefficients match those of Kleinbaum and Klein (2010), page 614 and thus validate this procedure.

Following are the detailed definitions:

Independent Variable

This is the variable from the model that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the binary variable is given. For example, suppose that a discrete independent

GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by the *Stagger label and output* option of the Report Options tab.

Regression Coefficient, b(i)

This is the estimate of the regression coefficient, β_i. The quantity β_i is the amount that the log of the odds ratio changes when x_i is increased by one unit.

Standard Error, Sb(i)

This is s_{b_j}, the large-sample estimate of the standard error of the regression coefficient. This is an estimate of the precision of the regression coefficient. It is provided by the square root of the corresponding diagonal element of the covariance matrix, $V\left(\hat{\beta}\right) = I^{-1}$. It is also used as the denominator of the Wald test.

Wald Z-Value

This is the *z* value of the Wald test used for testing the hypothesis that $\beta_i = 0$ against the alternative $\beta_i \neq 0$. The Wald test is calculated using the formula

$$z_i = \frac{b_{ij}}{s_{b_i}}$$

The distribution of the Wald statistic is closely approximated by the normal distribution in large samples. However, in small samples, the normal approximation may be poor.

Wald Prob Level

This is the two-sided probability level. This is the probability of obtaining a z-value larger in absolute value than the one obtained. If this probability is less than the specified significance level (say 0.05), the regression coefficient is significantly different from zero.

Odds Ratio Exp(b(i))

This the value of e^{β_i}. This value is often called the *adjusted odds ratio*. However, you must keep in mind that this interpretation is only valid with the corresponding variable is a 0-1 binary variable. We refer you to Kleinbaum 1994 for a detailed discussion of the interpretation of logistic regression coefficients as odds ratios.

Mean

This is the average of this independent variable.

Confidence Limits

Confidence Limits of Regression Coefficients and Odds Ratios
Event Variable: MI
Match Group Variable: Match

Independent Variable	Regression Coefficient b(i)	Lower 95.0% C.L. of β	Upper 95.0% C.L. of β	Odds Ratio Exp(b(i))	Lower 95.0% C.L. of Exp(β)	Upper 95.0% C.L. of Exp(β)
SMK	0.7291	-0.3710	1.8291	2.073	0.690	6.228
SBP	0.0456	0.0158	0.0755	1.047	1.016	1.078
ECG	1.5993	-0.0734	3.2719	4.949	0.929	26.362

This report provides the confidence intervals for the regression coefficients and the odds ratios. The confidence coefficient, in this example 95%, was specified on the Reports tab by specifying the Alpha Level. You can check that these confidence intervals match those of Kleinbaum and Klein (2010), page 614 and thus validate this procedure.

Independent Variable

This is the independent variable that is displayed on this line. If the variable is continuous, it is displayed directly. If the variable is discrete, the definition of the binary variable that was generated is given. For example, suppose that a discrete independent GRADE variable has three values: A, B, and C. The name shown here would be something like *GRADE=B*. This refers to a binary variable that is one for those rows in which GRADE was B and zero otherwise.

Note that the placement of the name is controlled by *Stagger label and output* option of the Report Options tab.

Regression Coefficient, b(i)

This is the estimate of the regression coefficient, β_i. Thus the quantity β_i is the amount that the log of the odds ratio changes when x_i is increased by one unit.

Confidence Limits of β

A 95% confidence interval for β_i is given by an upper and lower limit. These limits are based on the Wald statistic using the formula

$$b_i \pm z_{1-\alpha/2} s_{b_i}$$

Since they are based on the Wald test, they are only valid for large samples.

Odds Ratio Exp(b(i))

This the value of e^{β_i}. This value is often called the *odds ratio*.

Confidence Limits of Exp(β)

A 95% confidence interval for e^{β_i} is given by an upper and lower limit. These limits are based on the Wald statistic using the formula

$$\exp\left(b_i \pm z_{1-\alpha/2} s_{b_i}\right)$$

Since they are based on the Wald test, they are only valid for large samples.

Log Likelihood and Chi² Tests

Log Likelihood and Chi² Tests
Event Variable: MI
Match Group Variable: Match

Term(s) Omitted	DF	Log Likelihood	-2 Log Likelihood	Increase Above Model Deviance (Chi²)	Prob Level	Amount R² Increased By This Term
All Terms	3	-185.7248	371.4496	22.201	0.9927	0.173
SMK	1	-175.4818	350.9636	1.715	0.1904	0.012
SBP	1	-179.9278	359.8556	10.607	0.0011	0.078
ECG	1	-176.7488	353.4977	4.249	0.0393	0.031
None(Model)	3	-174.6244	349.2487			

This report is the conditional logistic regression analog of the analysis of variance table. It displays the results of chi-square tests used to test whether each of the individual terms in the regression are statistically significant after adjusting for all other terms in the model.

Since this report requires that a separate regression be run for each term, it may require a long time to calculate.

This report is not produced during a subset selection run.

Term(s) Omitted

This is the model term that is being tested. The test is formed by comparing the deviance statistic when the term is removed with the deviance of the complete model. Thus, the deviance when the term is left out of the model is shown.

The "All" line refers to a no-covariates model. The "None(Model)" refers to the complete model with no terms removed.

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 *Stagger label and output* option of the Report Options tab. This should create a better looking report when the names are extra long.

DF

This is the degrees of freedom of the chi-square test displayed on this line.

Log Likelihood

This is the log likelihood achieved by the model being described on this line of the report.

-2 Log Likelihood

This is -2 times the log likelihood. It is analogous to the sums of squares in an ANOVA table. The final value (the value with no terms omitted) is known as the *Deviance*.

Increase above Model Deviance

This is a measure of the predictability added by this term. It is the increase over the final value. It is the Chi^2 value.

Prob Level

This is the significance level of a Chi^2 test. This is the probability that a Chi^2 value with degrees of freedom DF is equal to this value or greater than the test value. If this value is less than 0.05 (or other appropriate value), the term is said to be statistically significant.

Amount R^2 Increased By This Term

This is amount that R^2 is reduced when this term is omitted from the regression model. This reduction is calculated from the R^2 achieved by the full model.

This quantity is used to determine if removing a term causes a large reduction in R^2. If it does not, then the term can be safely removed from the model.

Example 2 – Subset Selection

This section presents an example of how to conduct a subset selection. We will again use the Kleinbaum MI dataset that was used in Example 1. In this run, we will be trying to find a subset of two covariates that should be kept in the regression model.

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 Conditional Logistic Regression window.

1 Open the Kleinbaum MI dataset.

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

2 Open the Conditional Logistic Regression window.

- Using the Analysis menu or the Procedure Navigator, find and select the **Conditional Logistic Regression** procedure
- On the menus, select **File**, then **New Template**. This will load the default template.

3 Specify the variables.

- On the Conditional Logistic Regression window, select the **Variables, Model** tab.
- Enter **Match** in the **Match Group** variable box.
- Set the Ties Method to Breslow.
- Enter **MI** in the **Event** variable box.
- Enter **SMK,SBP,ECG** in the **Numeric X's** variables box.

4 Specify the Subset Selection.

- Set the **Search Method** box to **Hierarchical Forward with Switching**.
- Set the **Stop search when number of terms reaches** box to **2**.

5 Specify the reports.

- On the Cox Regression window, select the **Reports tab**.
- Uncheck all of the reports except **Subset Selection – Summary** and **Subset Selection - Detail**.

6 Run the procedure.

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

Subset Selection Summary

Number of Terms	Number of X's	Log Likelihood	R^2 Value	R^2 Change
0	0	-185.7248	0.000	0.000
1	1	-177.6814	0.128	0.128
2	2	-175.4818	0.161	0.032

This report shows the best log-likelihood value for each subset size. In this example, it appears that a model with three terms provides the best model. Note that adding the fourth variable does not increase the R-squared value very much.

Number of Terms

The number of terms in the regression model.

Number of X's

The number of *X*'s that were included in the model. Note that in this case, the number of terms matches the number of *X*'s. This would not be the case if some of the terms were categorical variables.

Log Likelihood

This is the value of the log likelihood function evaluated at the maximum likelihood estimates. Our goal is to find a subset size above which little is gained by adding more variables.

R^2 Value

This is the value of R^2 calculated using the formula

$$R_k^2 = 1 - \exp\left[\frac{2}{n}\left(L_0 - L_k\right)\right]$$

as discussed in the introduction. We are looking for the subset size after which this value does not increase by a meaningful amount.

R^2 Change

This is the increase in R^2 that occurs when each new subset size is reached. Search for the subset size below which the R^2 value does not increase by more than 0.02 for small samples or 0.01 for large samples.

In this example, the optimum subset size appears to be three terms.

Subset Selection Detail

Step	Action	No. of Terms	No. of X's	Log Likelihood	R^2	Term Entered	Term Removed
0	Begin	0	0	-185.7248	0.000		
1	Add	1	1	-177.6814	0.128	SBP	
2	Add	2	2	-175.4818	0.161	ECG	

This report shows the highest log likelihood for each subset size.

Action

This item identifies the action that was taken at this step. A term was added, removed, or two were switched.

No. of Terms

The number of terms.

No. of X's

The number of *X*'s that were included in the model.

Log Likelihood

This is the value of the log likelihood function after the completion of this step. Our goal is to find a subset size above which little is gained by adding more variables.

R^2 Value

This is the value of R^2 calculated using the formula that was discussed in the introduction. We are looking for the subset size after which this value does not increase by a meaningful amount.

Terms Entered and Removed

These columns identify the terms added, removed, or switched.