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 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 generally 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

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

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

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

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, $V(\hat{\beta}) = I^{-1}$.

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

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

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

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

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 (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
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.

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 to study the relationship between myocardial infarction and the covariates smoking, blood pressure, and electrocardiogram status.

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.

Setup

To run this example, complete the following steps:

1 Open the Kleinbaum MI example dataset

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

2 Specify the Conditional Logistic Regression procedure options

- Find and open the **Conditional Logistic 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, Model Tab

Match Group	**Match**
Ties	**Breslow (1:n matching)**
Event	**MI**
Numeric X's	**SMK,SBP,ECG**

3 Run the procedure

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

Run Summary

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	7	Convergence Criterion	1E-06
Maximum Iterations Allowed	20	Achieved Convergence	9.326338E-15
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 Wald Z-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-Test of H0: β(i) = 0 vs. H1: β(i) ≠ 0 Z-Value	P-Value	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, $V(\hat{\beta}) = I^{-1}$. It is also used as the denominator of the Wald test.

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.

P-Value

This is the two-sided p-value. 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 Intervals for Regression Coefficients and Odds Ratios

Confidence Intervals for Regression Coefficients and Odds Ratios

Event Variable: MI
Match Group Variable: Match

Independent Variable	Regression Coefficient b(i)	95% Confidence Interval Limits for β(i)		Odds Ratio Exp(b(i))	95% Confidence Interval Limits for Exp(β(i))	
		Lower	Upper		Lower	Upper
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 Interval Limits for β(i)

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 Interval Limits for Exp(β(i))

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 Chi^2 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)	P-Value*	Amount R² Increased by This Term
All Terms	3	-185.7248	371.4496	22.201	0.0001	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			

* The P-Value is for testing the significance of that term after adjusting for all other terms.

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 (Chi^2)

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.

P-Value

This is the p-value 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 the 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.

Setup

To run this example, complete the following steps:

1 Open the Kleinbaum MI example dataset

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

2 Specify the Conditional Logistic Regression procedure options

- Find and open the **Conditional Logistic 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, Model Tab

Match Group	**Match**
Ties	**Breslow (1:n matching)**
Event	**MI**
Numeric X's	**SMK,SBP,ECG**
Search Method	**Hierarchical Forward with Switching**
Stop search when number of terms reaches	**2**

Reports Tab

Run Summary	**Checked**
Subset Summary	**Checked**
Subset Detail	**Checked**
Regression Coefficients	**Checked**
C.L. of Regression Coefficients	**Checked**
Log-Likelihood and Chi² Tests	**Checked**

3 Run the procedure

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

Subset Selection Summary

Subset Selection Summary

Event Variable: MI
Match Group Variable: Match

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

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

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

Subset Selection Detail

Event Variable: MI
Match Group Variable: Match

Step	Action	Number of Terms	Number of X's	Log-Likelihood	R^2	Terms Entered	Terms 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.

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 regression 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

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.