NOMREG

The purpose of the Multinomial Logistic Regression procedure is to model the dependence of a nominal categorical response on a set of discrete and/or continuous predictor variables.

Notation

The following notation is used throughout this chapter unless otherwise stated:

Y	The response variable, which takes integer values from 1 to *J*.
J	The number of categories of the nominal response.
m	The number of subpopulations.
$\mathbf{X}^A$	$m \times p^A$ matrix with vector-element x_i^A, the observed values at the *i*th subpopulation, determined by the independent variables specified in the command.
$\mathbf{X}$	$m \times p$ matrix with vector-element x_i, the observed values of the location model's independent variables at the *i*th subpopulation.
f_{ijs}	The frequency weight for the *s*-th observation which belongs to the cell corresponding to $Y = j$ at subpopulation *i*.
n_{ij}	The sum of frequency weights of the observations that belong to the cell corresponding to $Y = j$ at subpopulation *i*.
N	The sum of all n_{ij}'s.
π_{ij}	The cell probability corresponding to $Y = j$ at subpopulation *i*.
$\log(\pi_{ij} / \pi_{ik})$	The logit of response category *j* to response category *k*.
$\beta_j = (\beta_{j1},...,\beta_{jp})'$	$p \times 1$ vector of unknown parameters in the *j*-th logit (i.e., logit of response category *j* to response category *J*).
p	Number of parameters in each logit. $p \geq 1$.
p_j^{nr}	Number of non-redundant parameters in logit *j* after maximum likelihood estimation. $p \geq p_j^{nr} \geq 0$.
p^{nr}	The total number of non-redundant parameters after maximum likelihood estimation. $p^{nr} = \sum_{j=1}^{k-1} p_j^{nr}$.
$\mathbf{B} = (\beta_1^{'},...,\beta_{J-1}^{'})^{'}$	$(k-1)p \times 1$ vector of unknown parameters in the model.
$\hat{\mathbf{B}} = (\hat{\beta}_1^{'},...,\hat{\beta}_{J-1}^{'})^{'}$	The maximum likelihood estimate of $\mathbf{B}$.
$\hat{\pi}_{ij}$	The maximum likelihood estimate of π_{ij}.

Data Aggregation

Observations with negative or missing frequency weights are discarded. Observations are aggregated by the definition of subpopulations. Subpopulations are defined by the cross-

classifications of either the set of independent variables specified in the command or the set of independent variables specified in the subpopulation command.

Let n_i be the marginal count of subpopulation *i*,

$$n_i = \sum_{j=1}^{k} n_{ij} .$$

If there is no observation for the cell of $Y = j$ at subpopulation *i*, it is assumed that $n_{ij} = 0$, provided that $n_i \neq 0$. A non-negative scalar $\delta \in [0, 1)$ may be added to any zero cell (i.e., cell with $n_{ij} = 0$) if its marginal count n_i is nonzero. The value of δ is zero by default.

Data Assumptions

Let $(n_{i1}, \ldots, n_{iJ})^{\mathrm{T}}$ be the $J \times 1$ vector of counts for the categories of *Y* at subpopulation. It is assumed that each $(n_{i1}, \ldots, n_{iJ})^{\mathrm{T}}$ is independently multinomial distributed with probability vector $(\pi_{i1}, \ldots, \pi_{iJ})^{\mathrm{T}}$ of dimension $J \times 1$ and fixed total n_i.

Model

Generalized Logit Model

In a Generalized Logit model, the probability π_{ij} of response category *j* at subpopulation *i* is

$$\pi_{ij} = \frac{\exp(\mathbf{x}_i' \beta_j)}{1 + \sum_{k=1}^{J-1} \exp(\mathbf{x}_i' \beta_k)},$$

where the last category *J* is assumed to be the reference category.

In terms of logits, the model can be expressed as

$$\log\left(\frac{\pi_{ij}}{\pi_{iJ}}\right) = \mathbf{x}_i' \beta_j$$

for *j* = 1, …, *J*-1.

When *J* = 2, this model is equivalent to the binary Logistic Regression model. Thus, the above model can be thought of as an extension of the binary Logistic Regression model from binary response to polytomous nominal response.

1-1 Matched Case Control Model by Conditional Likelihood Approach

The above model can also be used to estimate the parameters in the conditional likelihood of the 1-1 Matched Case Control Model. In this case, let m be the number of matching pairs, $\boldsymbol{x}_{i1}$ be the vector of independent variables for the case and $\boldsymbol{x}_{i2}$ that for the control. The conditional log-likelihood for the m matched pairs is given by

$$l = \frac{\exp\{(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\beta\}}{1 + \exp\{(\mathbf{x}_{i1} - \mathbf{x}_{i2})'\beta\}}$$

in which $\boldsymbol{\beta}$ is the vector of parameters for the difference between the values of independent variables of the case and those of the control. This conditional likelihood is identical to the unconditional log-likelihood of a binary (i.e., $k = 2$) logistic regression model when

- There is no intercept term in the model.
- The set of subpopulations is defined by the set of matching pairs.
- The independent variables in the model are set to equal to the differences between the values for the case and the control.
- The number of response categories is $J = 2$, and the value of the response is 1 (or a constant), i.e., $Y = 1$.

Log-likelihood

The log-likelihood of the model is given by

$$\begin{aligned} l(\mathbf{B}) &= \sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij} \log\left(\pi_{ij}\right) \\ &= \sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij} \log\left(\frac{\exp(\mathbf{x}_i' \beta_j)}{1 + \sum_{k=1}^{J-1}\exp(\mathbf{x}_i' \beta_k)}\right) \end{aligned}$$

A constant that is independent of parameters has been excluded here. The value of the constant is $c = \sum_{i=1}^{m} \log\{n_i!/(n_{i1}!\ldots n_{iJ}!)\}$.

Parameter Estimation

First and Second Derivatives of the Log-likelihood

For any $j = 1, \ldots, J\text{-}1$, $s = 1, \ldots, p$, the first derivative of l with respect to β_{js} is

$$\frac{\partial l}{\partial \beta_{js}} = \sum_{i=1}^{m} x_{is}(n_{ij} - n_i \pi_{ij}).$$

For any $j, j' = 1, \ldots, J\text{-}1$, $s, t = 1, \ldots, p$, the second derivative of l with respect to β_{js} and $\beta_{j't}$ is

$$\frac{\partial^2 l}{\partial \beta_{js} \partial \beta_{j't}} = -\sum_{i=1}^{m} n_i x_{is} x_{it} \pi_{ij} (\delta_{jj'} - \pi_{ij'})$$

where $\delta_{jj'} = 1$ if $j = j'$, 0 otherwise.

Maximum Likelihood Estimate

To obtain the maximum likelihood estimate of $\mathbf{B}$, a Newton-Raphson iterative estimation method is used. Notice that this method is the same as Fisher-Scoring iterative estimation method in this model, since the expectation of the second derivative of l with respect to $\mathbf{B}$ is the same as the observed one.

Let $\partial l / \partial \mathbf{B}$ be the $(J-1)p \times 1$ vector of the first derivative of l with respect to $\mathbf{B}$. Moreover, let $\left[\partial^2 l / \partial \mathbf{B} \partial \mathbf{B}\right]$ be the $(J-1)p \times (J-1)p$ matrix of the second derivative of l with respect to $\mathbf{B}$. Notice that $-\left[\partial^2 l / \partial \mathbf{B} \partial \mathbf{B}\right] = \sum_{i=1}^{m} \mathbf{X}_i^* \Delta_i \mathbf{X}_i^{*\prime}$ where Δ_i is a $(J-1) \times (J-1)$ matrix as

$$\Delta_i = n_i (\mathrm{Diag}(\pi_i^{(-J)}) - \pi_i^{(-J)} \pi_i^{(-J)\prime}),$$

in which $\pi_i^{(-J)} = (\pi_{i1}, \ldots, \pi_{iJ-1})'$ and $\mathrm{Diag}(\pi_i^{(-J)})$ is a $(J-1) \times (J-1)$ diagonal matrix of $\pi_i^{(-J)}$. Let $\mathbf{B}^{(v)}$ be the parameter estimate at iteration v, the parameter estimate $\mathbf{B}^{(v+1)}$ at iteration $v + 1$ is updated as

$$\mathbf{B}^{(v+1)} = \mathbf{B}^{(v)} + \xi \left(\sum_{i=1}^{m} \mathbf{X}_i^* \Delta_i^{(v)} \mathbf{X}_i^{*\prime} \right)^{-1} \frac{\partial l}{\partial \mathbf{B}^{(v)}}$$

and $\xi > 0$ is a stepping scalar such that $l\left(\mathbf{B}^{(v+1)}\right) - l\left(\mathbf{B}^{(v)}\right) \geq 0$, $\mathbf{X}^*$ is a $(J-1)p \times (J-1)$ matrix of independent vectors,

$$\mathbf{X}_\mathbf{i}^* = \begin{pmatrix} \mathbf{x}_i & 0 & \cdots & 0 \\ 0 & \mathbf{x}_i & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{x}_i \end{pmatrix},$$

and $\Delta_i^{(\nu)}$ is Δ_i and $\partial l / \partial \mathbf{B}^{(\nu)}$ is $\partial l / \partial \mathbf{B}$, both evaluated at $\mathbf{B} = \mathbf{B}^{(\nu)}$.

Stepping

Use step-halving method if $l\left(\mathbf{B}^{(\nu+1)}\right) - l\left(\mathbf{B}^{(\nu)}\right) < 0$. Let V be the maximum number of steps in step-halving, the set of values of ξ is $\{1/2^{\nu}: \nu = 0, \ldots, V\text{-}1\}$.

Starting Values of the Parameters

If intercepts are included in the model, set $\beta_j^{(0)} = (\beta_{j1}^{(0)}, 0, \ldots, 0)'$ where

$$\beta_{j1}^{(0)} = \log\left(\frac{\tilde{\pi}_{ij}}{\tilde{\pi}_{iJ}}\right) = \log\left(\frac{\sum_{i=1}^{m} n_{ij}}{\sum_{i=1}^{m} n_{iJ}}\right),$$

for $j = 1, \ldots, J\text{-}1$.

If intercepts are not included in the model, set

$\beta_j^{(0)} = (0, \ldots, 0)'$

for $j = 1, \ldots, J\text{-}1$.

Convergence Criteria

Given two convergence criteria $\varepsilon_k > 0$ and $\varepsilon_p > 0$, the iteration is considered to be converged if one of the following criteria are satisfied:

1. $\left| l(\mathbf{B}^{(\nu+1)}) - l(\mathbf{B}^{(\nu)}) \right| < \varepsilon_k$.

2. $\max_i \left| \mathbf{B}_i^{(\nu+1)} - \mathbf{B}_i^{\nu} \right| < \varepsilon_p$.

3. The maximum above element in $\partial l / \partial \mathbf{B}^{(\nu+1)}$ is less than $\min(\varepsilon_l, \varepsilon_p)$.

Statistics

Model Information

Initial Model, Intercept-Only

If intercepts are included in the model, the predicted probability for the initial model (that is, the model with intercepts only) is

$$\tilde{\pi}_{ij} = \frac{\sum_{i=1}^{m} n_{ij}}{N}$$

and the value of the -2log-likelihood of the initial model is

$$-2l(\tilde{\pi}) = -2\sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij}\log(\tilde{\pi}_{ij}).$$

Initial Model, Empty

If intercepts are not included in the model, the predicted probability for the initial model is

$$\tilde{\pi}_{ij} = \frac{1}{J}$$

and the value of the -2log-likelihood of the initial model is

$$-2l(\tilde{\pi}) = -2N\log(\frac{1}{J}).$$

Final Model

The value of -2log-likelihood of the final model is

$$-2l(\tilde{\pi}) = -2\sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij}\log(\hat{\pi}_{ij}).$$

Model Chi-Square

The Model Chi-square is given by

$-2l(\tilde{\pi}) - \{-2l(\hat{\pi})\}$.

Model with Intercepts versus Intercept-only Model

If the final model includes intercepts, then the initial model is an intercept-only model. Under the null hypothesis that $H_0: \beta^{\text{intercepts}} = \mathbf{0}$, the Model Chi-square is asymptotically chi-squared distributed with $p^{nr} - (J - 1)$ degrees of freedoms.

Model without Intercepts versus Empty Model

If the model does not include intercepts, then the initial model is an empty model. Under the null hypothesis that $H_0: \beta = \mathbf{0}$, the Model Chi-square is asymptotically chi-squared distributed with p^{nr} degrees of freedoms.

Pseudo R Square

Cox and Snell's R Square

The Cox and Snell's R^2 is

$$R^2_{\text{CS}} = 1 - \left(\frac{L(\tilde{\pi})}{L(\hat{\pi})} \right)^{\frac{2}{n}}.$$

Nagelkerke's R Square

The Nagelkerke's R^2 is

$$R^2_{\text{N}} = \frac{R^2_{\text{CS}}}{1 - L(\tilde{\pi})^{2/n}}.$$

McFadden's R Square

The McFadden's R^2 is

$$R^2_{\text{M}} = 1 - \left(\frac{l(\hat{\pi})}{l(\tilde{\pi})} \right).$$

Goodness of Fit Measures

Pearson Goodness of Fit Measure

The Pearson goodness of fit measure is

$$X^2 = \sum_{i=1}^{m}\sum_{j=1}^{J} \frac{(n_{ij} - n_i\hat{\pi}_{ij})^2}{n_i\hat{\pi}_{ij}}.$$

Under the null hypothesis, the Pearson goodness-of-fit statistic is asymptotically chi-squared distributed with $m(J-1) - p^{nr}$ degrees of freedom.

Deviance Goodness of Fit Measure

The Deviance goodness of fit measure is

$$D = 2\sum_{i=1}^{m}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{n_i\hat{\pi}_{ij}}\right).$$

Under the null hypothesis, the Deviance goodness-of-fit statistic is asymptotically chi-squared distributed with $m(J-1) - p^{nr}$ degrees of freedom.

Overdispersion Adjustments

Let $\hat{\kappa} > 0$ be an estimate of the overdispersion parameter. Possible estimates of this parameter are

- A positive value specified in the command. If no value is specified, 1 is assumed.
- The ratio of Pearson goodness-of-fit measure to its degrees of freedom:

$$\hat{\kappa} = \frac{X^2}{m(k-1) - p^{nr}}$$

- The ratio of Deviance goodness of fit measure to its degrees of freedoms:

$$\hat{\kappa} = \frac{D}{m(k-1) - p^{nr}}$$

Covariance and Correlation Matrices

The estimate of the covariance matrix of the parameters is the inverse of the negative of the second derivative of the log-likelihood evaluated at $\mathbf{B} = \mathbf{B}^{(v)}$, multiplied by the estimate of the overdispersion parameter.

$$\text{Cov}(\hat{\mathbf{B}}) = \hat{\kappa}\left[\sum_{i=1}^{m} \mathbf{X}_i^{*} \hat{\Delta}_i \mathbf{X}_i^{*\prime}\right]^{-1}.$$

Let $\hat{\sigma}$ be the $(J\text{-}1)p \times 1$ vector of the square roots of the diagonal elements in $\text{Cov}(\hat{\mathbf{B}})$. The estimate of the correlation matrix of $\hat{\mathbf{B}}$ is

$$\text{Cor}(\hat{\mathbf{B}}) = \text{Diag}(\hat{\sigma}^{-1})\text{Cov}(\hat{\mathbf{B}})\text{Diag}(\hat{\sigma}^{-1}).$$

Parameter Statistics

An estimate of the standard deviation of $\hat{B}_{js}$ is $\hat{\sigma}_{js}$. The Wald statistic for $\hat{B}_{js}$ is

$$\text{Wald}_{js} = \frac{\hat{B}_{js}}{\hat{\sigma}_{js}}$$

Under the null hypothesis that $H_0: B_{js} = 0$, Wald_{js} is asymptotically chi-squared distributed with 1 degree of freedom.

If B_{js} Based on the asymptotic normality of the parameter estimate, a $100(1\text{-}\alpha)$ % Wald confidence interval for $\hat{B}_{js}$ is

$$\hat{B}_{js} \pm z_{1-\alpha/2}\hat{\sigma}_{js}$$

where $z_{1-\alpha/2}$ is the upper $(1\text{-}\alpha/2)100^{\text{th}}$ percentile of the standard normal distribution.

Predicted Cell Counts

At each subpopulation *i*, the predicted count for response category $Y = j$ is

$$\hat{n}_{ij} = n_i \hat{\pi}_{ij}$$

The (raw) residual is $n_{ij} - \hat{n}_{ij}$ and the standardized residual is $(n_{ij} - \hat{n}_{ij}) / \sqrt{n_i \hat{\pi}_{ij}(1 - \hat{\pi}_{ij})}$.

Likelihood Based Partial Effects

A likelihood ratio test is performed for any effect (except intercept) in the model. The procedure to perform a likelihood ratio test for any effect e is as follows:

1. Form a submodel that has all the effects in the working model but the one (e) of interest.
2. Fit the submodel and calculate the value of its –2log-likelihood, denote it by $-2l(\hat{\pi}_{(e)})$. Moreover, let the number of non-redundant parameters in this submodel be $p^{nr}_{(e)}$.
3. Calculate the difference between the –2log-likelihood of the submodel and that of the working model, $\{-2l(\hat{\pi}_{(e)})\} - \{-2l(\hat{\pi})\}$.

Under the null hypothesis that the effect e of interest is zero, $\{-2l(\hat{\pi}_{(e)})\} - \{-2l(\hat{\pi})\}$ is asymptotically chi-squared distribution with $p^{nr} - p^{nr}_{(e)}$ degrees of freedoms.

Linear Hypothesis Testings

For each $q \times p$ matrix of linear combinations $\mathbf{L}$, J Wald's tests are performed. Each of the first $J - 1$ Wald's tests corresponds to a Wald's test on each of the $J - 1$ logits. The last Wald's tests corresponds to a Wald's test joint for all the $J - 1$ logits. In the followings, it is assumed that $q = \text{Rank}(\mathbf{L}) \le p$.

The Wald's test corresponding to the j-th logit is

$$\text{Wald}(\mathbf{L}, j) = (\mathbf{L}\hat{\beta}_j)'\{\mathbf{L}\text{Cov}(\hat{\beta}_j)\mathbf{L}'\}^{-1}(\mathbf{L}\hat{\beta}_j).$$

Under the null hypothesis that $H_0: \mathbf{L}\beta_j = \mathbf{0}$, $\text{Wald}(\mathbf{L}, j)$ is asymptotically chi-squared distributed with q degrees of freedoms.

Let $\boldsymbol{L}^*$ be a $(J-1)q \times (J-1)p$ matrix,

$$\mathbf{L}^* = \begin{pmatrix} \mathbf{L} & 0 & \cdots & 0 \\ 0 & \mathbf{L} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \mathbf{L} \end{pmatrix}.$$

The Wald's joint test for all logits is

$$\text{Wald}(\mathbf{L}, \circ) = (\mathbf{L}^*\hat{\mathbf{B}})'\{\mathbf{L}^*\text{Cov}(\hat{\mathbf{B}})\mathbf{L}^{*'}\}^{-1}(\mathbf{L}^*\hat{\mathbf{B}}).$$

Under the null hypothesis that $H_0: \mathbf{L}^*\mathbf{B} = \mathbf{0}$, $\text{Wald}(\mathbf{L}, \circ)$ is asymptotically chi-squared distributed with $(J\text{-}1)q$ degrees of freedoms.

Classification Table

Suppose that $c(j, j')$ is the (j, j')-th element of the classification table, $j, j' = 1, \ldots, J$. $c(j, j')$ is the sum of the frequencies for the observations whose actual response category is j (as row) and predicted response category is j' (as column) respectively.

The predicted response category for subpopulation i is

$$j^*: \hat{\pi}_{ij^*} = \max_j \left(\hat{\pi}_{ij} \right)$$

Should there be a tie, choose the category with the smallest category number.

For $j, j' = 1, \ldots, J$, $c(j, j')$ is given as

$$c(j, j') = \sum_{i=1}^{m} n_{ij} \delta_{j_i^* j'}.$$

The percentage of total correct predictions of the model is

$$p(\circ) = \left(\frac{\sum_{j=1}^{n} c(j, j)}{n} \right) 100\%.$$

The percentage of correct predictions of the model for response category j is

$$p(\circ) = \left(\frac{c(j, j)}{\sum_{i=1}^{m} n_{ij}} \right) 100\%.$$

Checking for Separation

The algorithm checks for separation in the data starting with iteration ν^{chksep} (20 by default). To check for separation:

1. For each subpopulation i, find $j^*: \hat{\pi}_{ij^*} = \max_j \left(\hat{\pi}_{ij} \right)$.
2. If $n_{ij^*} = n_i$, then there is a perfect prediction for subpopulation i.
3. If all subpopulations have perfect prediction, then there is complete separation. If some patterns have perfect prediction and the Hessian of $\hat{\mathbf{B}}$ is singular, then there is quasi-complete separation.

References

Agresti, A. (1990). *Categorical Data Analysis*. NY: John Wiley.

Cramer, J.S., and Ridder, G. (1988). The Logit Model in Economics. *Statistica Neerlandica*, 42, 291-314.

Cohen, A., and Rom, M. (1994). A Method for Hypothesis Tests in Polychotomous Logistic Regression. *Computational Statistics and Data Analysis*, 17, 277-288.

Cox, D.R., and Snell, E.J. (1989). *The Analysis of Binary Data*. Second Ed.. London: Chapman and Hall.

Haberman, S. J. (1974). *The Analysis of Frequency Data*. Chicago: University of Chicago, 351-373.

Hauck, W.W., and Donner, A. (1977). Wald's Test as Applied to Hypotheses in Logit Analysis. *J. American. Statist. Ass.*, 72, 851-853.

Hosmer, D. W., and Lemeshow, S. (1989). *Applied Logistic Regression*. NY: Wiley.

Long, J. S. (1997). *Regression Models for Categorical and Limited Dependent Variables*. NY: Advanced Quantitative Techniques in the Social Sciences Series.

Luce, R. D. (1959). *Individual Choice Behavior*. NY: John Wiley.

McCullagh, P., and Nelder, J.A. (1990). Generalized Linear Models, 2nd Edition. NY: Chapman and Hall.

McFadden, D. (1973). Conditional Logit Analysis of Qualitative Choices Behavior. *Frontiers in Econometrics*, edited by Zarembka, P., NY: John Wiley, 105-35.

Nagelkerke, N.J.D. (1991). A Note on a General Definition of the Coefficient of Determination. *Biometrika*, 78, 691-692.

Searle, R.S. (1987). *Linear Models for Unbalanced Data*. NY: Wiley.

Zhang, J., and Hoffman, S.D. (1993). Discrete-Choice Logit Models. Testing the IIA Property. *Sociological Methods and Research*, 22, 2, 193-213.