CSLOGISTIC

Logistic regression is a commonly used analytical tool for categorical responses. SPSS has procedures LOGISTIC REGRESSION (for binary response) and NOMREG (for multi-category response) under the standard sampling setting. This document considers multinomial logistic regression model under the complex sampling setting extending the model in NOMREG to complex sampling.

There are different approaches for analytic inference in complex sampling (Chambers and Skinner 2003). We will take the two-phase sampling and pseudo-likelihood estimation approaches.

Notation

y_i Categorical dependent/response variable for case *i*. Its category values are denoted as 1, 2, etc.

K The total number of categories for dependent variable.

$y_i(k)$ Indicator variable for category *k*, i.e. $y_i(k) = \begin{cases} 1 & \text{if } y_i = k \\ 0 & \text{otherwise} \end{cases}$.

$\mathbf{X}$ Design matrix $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)'$, where $\mathbf{x}_i = (x_{i1}, \cdots, x_{ip})'$ is for case *i*.

Note $x_{i1} = 1$ if model is with intercept.

π_i Inclusion probability for case *i*.

w_i Sampling weight for case *i*, $w_i = 1/\pi_i$.

$p_{\mathbf{x}}(k)$ The probability for response category *k* at $\mathbf{x}$: $p_{\mathbf{x}}(k) = \Pr(y = k \mid \mathbf{x})$, and denote $p_i(k) = p_{\mathbf{x}_i}(k)$ for case *i*.

N The number of cases in the whole population.

N The number of cases in the sample.

$\mathbf{B}$ The parameter of interest, the population or census parameter.

Input

- The sampling plan.

  This plan provides information about the sampling method, sampling weight, strata and cluster information.

- Observed sample data $\{y_i, \mathbf{x}_i, w_i\}_{i=1}^{n}$.

  Predictors can be either categorical or continuous. The intercept, main effect, interaction effects and nested effects can be in the model.

Superpopulation model

Two phases of sampling are assumed. The first phase generates a finite population by a model or super population. The second phase selects a sample according to a sampling plan from the finite population generated in the first phase.

Model generating the population

Assume that response variable *y* at a given $\mathbf{x}$ follows a multinomial distribution with probability $p_{\mathbf{x}}(k)$ for $y = k$. Without loss of generality, let the last category *K* be the reference category. Then for *k* = 1, …, *K*-1,

$$\log \frac{p_{\mathbf{x}}(k)}{p_{\mathbf{x}}(K)} = \mathbf{x}'\boldsymbol{\beta}_k .$$

Or

$$p_{\mathbf{x}}(k) = \begin{cases} \dfrac{\exp(\mathbf{x}'\boldsymbol{\beta}_k)}{1+\sum_{k=1}^{K-1}\exp(\mathbf{x}'\boldsymbol{\beta}_k)} & k = 1,\cdots,K-1 \\ \dfrac{1}{1+\sum_{k=1}^{K-1}\exp(\mathbf{x}'\boldsymbol{\beta}_k)} & k = K \end{cases},$$

where $\boldsymbol{\beta}_k = (\beta_{k1},\cdots,\beta_{kp})'$ is the regression parameter vector for response category *k*. There are *p*(*K*-1) regression parameters in total $\boldsymbol{\beta} = (\boldsymbol{\beta}_1',\ldots,\boldsymbol{\beta}_{K-1}')'$. This model is described in many books, for example Agresti (2002).

The parameter of interest

Let **B** denote the MLE of the model parameter $\boldsymbol{\beta}$ based on the whole population. This **B** is also called the census parameter. The parameter of interest is the census parameter **B**, rather than the model parameter $\boldsymbol{\beta}$. The exact definition and formulation of **B** is described below in the estimating equation.

Estimating parameters from a complex sample

For a sample $S = \{y_i, \mathbf{x}_i\}_{i=1}^{n}$ drawn from the finite population according to a sample plan, we take the pseudo-likelihood approach. In this approach, the pseudo-likelihood is a sample estimate of the population log-likelihood, and parameter estimates are derived by maximizing the pseudo-likelihood.

From the sample, an unbiased estimate of population log-likelihood l_U is

$$l_S(\boldsymbol{\beta}) = \sum_{i\in S}\sum_{k=1}^{K} w_i y_i(k)\log\bigl(p_i(k)\bigr).$$

We will maximize $l_S(\boldsymbol{\beta})$ to get the estimates for census parameter **B**. The pseudo-score function is, for k = 1, …, K-1,

$$S_S(\boldsymbol{\beta}) = \sum_{i\in S} w_i\left(\mathbf{y}_i^* - \mathbf{p}_i^*\right)\otimes \mathbf{x}_i$$

Estimating equation

For k = 1, …, K-1,

$$\sum_{i\in S} w_i\left(\mathbf{y}_i^* - \mathbf{p}_i^*\right)\otimes \mathbf{x}_i = 0\,.$$

The estimator obtained by solving this equation is an estimator of the census parameter **B**.

Redundant parameters

In this procedure, the over-parameterization approach is similar to that in the NOMREG procedure. If a parameter is found to be redundant, it is set to zero and will not affect the estimation procedure.

Parameter estimates

To obtain the maximum pseudo-likelihood estimate of **B**, the *Newton-Raphson iterative estimation method* is used to solve the estimating equation. Let $\mathbf{B}^{(v)}$ be the parameter estimate at iteration step v, the parameter estimate $\mathbf{B}^{(v+1)}$ at iteration step v + 1 is updated as

$$\mathbf{B}^{(v+1)} = \mathbf{B}^{(v)} - \xi\cdot J^{-}\left(\mathbf{B}^{(v)}\right)S_S\left(\mathbf{B}^{(v)}\right)$$

where

$$J(\boldsymbol{\beta}) = \frac{\partial S_S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\sum_{i\in S} w_i\left(diag(\mathbf{p}_i^*) - \mathbf{p}_i^*\left(\mathbf{p}_i^*\right)'\right)\otimes \mathbf{x}_i\mathbf{x}_i'\,,$$

the (k, j)th block element of $J(\boldsymbol{\beta})$, for k, j = 1, …, K-1, is

$$J_{kj}(\boldsymbol{\beta}) = \frac{\partial S_S(\boldsymbol{\beta}_k)}{\partial \boldsymbol{\beta}_j} = \begin{cases} \sum_{i\in S} w_i p_i(k) p_i(j)\mathbf{x}_i\mathbf{x}_i' & k \neq j \\ -\sum_{i\in S} w_i p_i(j)\left(1 - p_i(j)\right)\mathbf{x}_i\mathbf{x}_i' & k = j \end{cases}.$$

$J^{-}(\boldsymbol{\beta})$ is a generalized inverse of $J(\boldsymbol{\beta})$. The stepping scalar $\xi > 0$ is used to make $l_S(\mathbf{B}^{(v+1)}) \geq l_S(\mathbf{B}^{(v)})$. Use step-halving method if $l_S(\mathbf{B}^{(v+1)}) < l_S(\mathbf{B}^{(v)})$. Let *t* be the maximum number of steps in step-halving; the set of values of ξ is {$1/2^r$: *r* = 0, …, *t*-1}.

Starting with initial values $\mathbf{B}^{(0)}$, iteratively update $\mathbf{B}^{(v+1)}$ until one of the stopping criteria is satisfied. The final estimate is denoted as $\hat{\mathbf{B}}$.

Note:

- Sometimes, infinite parameters may be present in the model because of complete or quasi-complete separation of the data (Albert and Anderson, 1984) (Santner and Duffy, 1986). In CSLOGISTIC, a check for separation of the data can be performed. If either complete or quasi-complete separation is suggested by the test, a warning is issued and results based on the last iteration are given.

Initial values $\mathbf{B}^{(0)}$

For all non-intercept regression parameters, set their initial values to be zero. For intercepts, if there are any, set for *k*=1, …, *K*-1,

$$B_{k1}^{(0)} = \log\left(\frac{\hat{N}_k}{\hat{N}_K}\right).$$

Where $\hat{N}_k = \sum_{i \in S} w_i y_i(k)$ is the estimated population number of responses in category *k*.

Stopping criteria

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

1. $$\begin{cases} \dfrac{\left|l_S(\mathbf{B}^{(v+1)}) - l_S(\mathbf{B}^{(v)})\right|}{\left|l_S(\mathbf{B}^{(v)})\right| + 10^{-6}} < \varepsilon_l & \text{if relative change} \\ \left|l_S(\mathbf{B}^{(v+1)}) - l_S(\mathbf{B}^{(v)})\right| < \varepsilon_l & \text{if absolute change} \end{cases}.,$$

2. $$\begin{cases} \max_{k,j}\left(\dfrac{\left|B_{kj}^{(v+1)} - B_{kj}^{(v)}\right|}{\left|B_{kj}^{(v)}\right| + 10^{-6}}\right) < \varepsilon_p & \text{if relative change} \\ \max_{k,j}\left(\left|B_{kj}^{(v+1)} - B_{kj}^{(v)}\right|\right) < \varepsilon_p & \text{if absolute change} \end{cases}.$$

3. The maximum number of iterations is reached.

Properties of the estimates

Variance matrix

The design-based variance of $\hat{\mathbf{B}}$ (Binder 1983) has estimate

$$\hat{V}(\hat{\mathbf{B}}) \approx J^{-}(\hat{\mathbf{B}})\hat{I}(\hat{\mathbf{B}})J^{-}(\hat{\mathbf{B}}),$$

where $\hat{I}(\boldsymbol{\beta})$ is the estimate of design based variance of $S_S(\boldsymbol{\beta})$. Let $\mathbf{d}_i = \left(\mathbf{y}_i^* - \mathbf{p}_i^*\right) \otimes \mathbf{x}_i$, then $S_S(\boldsymbol{\beta}) = \sum_{i \in S} w_i \left(\mathbf{y}_i^* - \mathbf{p}_i^*\right) \otimes \mathbf{x}_i = \sum_{i \in S} w_i \mathbf{d}_i$ is an estimate for population total of $\mathbf{d}_i$ vectors. How to calculate designed based variance matrix for total is given in "Complex Samples: Covariance Matrix of Total" (cs_covariance.pdf).

Confidence intervals

The confidence interval for a single regression parameter B_{kj} is approximately

$$\left[\hat{B}_{kj} - t_{df,1-\frac{\alpha}{2}} se(\hat{B}_{kj}), \hat{B}_{kj} + t_{df,1-\frac{\alpha}{2}} se(\hat{B}_{kj})\right].$$

Where $se(\hat{B}_{kj}) = \hat{V}(\hat{B}_{kj})$ is the estimated standard error of $\hat{B}_{kj}$, and $t_{df,1-\frac{\alpha}{2}}$ is the $100(1-\alpha/2)$ percentile of t distribution with *df* degrees of freedom. The degrees of freedom *df* can be user specified, and default as the difference between the number of primary sampling units and the number of strata in the first stage of sampling.

Design effect

For each parameter B_{kj}, its design effect is the ratio of its variance under the design to its variance under the SRS design,

$$Deff(\hat{B}_{kj}) = \frac{\hat{V}(\hat{B}_{kj})}{\hat{V}_{srs}(\hat{B}_{kj})}.$$

For SRS design, the variance matrix is

$$V_{SRS}(\hat{\mathbf{B}}) \approx J^{-}(\hat{\mathbf{B}})I_{SRS}(\hat{\mathbf{B}})J^{-}(\hat{\mathbf{B}}),$$

where

$$\hat{I}_{srs}(\hat{\mathbf{B}}) = \hat{\mathbf{V}}_{srs}\left(S_S(\hat{\mathbf{B}})\right) = \left(1 - \frac{n}{\hat{N}}\right)\frac{\hat{N}}{n-1}\sum_{i\in S} w_i \mathbf{d}_i \mathbf{d}_i' ,$$

$$\hat{N} = \sum_{i\in S} w_i .$$

Pseudo -2 Log Likelihood

For the model under consideration, the pseudo –2 Log Likelihood is

$$-2l_S(\hat{\mathbf{B}}).$$

Let the initial model be the intercept-only model if the intercept is in the considered model, or the empty model otherwise. For the initial model, the pseudo –2 Log Likelihood is

$$-2l_S(\mathbf{B}^{(0)}),$$

where $\mathbf{B}^{(0)}$ happens to be the initial parameter values used in the iterative estimating procedure.

Pseudo R Squares

Let $L_U(\mathbf{B})$ be the likelihood function for the whole population; that is, $L_U(\mathbf{B}) = \exp\left(l_U(\mathbf{B})\right)$. A sample estimate is $\hat{L}_U(\mathbf{B}) = \exp\left(l_S(\mathbf{B})\right)$

Cox and Snell’s R Square

$$R_{CS}^2 = 1 - \left(\frac{\hat{L}_U(\mathbf{B}^{(0)})}{\hat{L}_U(\hat{\mathbf{B}})}\right)^{\frac{2}{\hat{N}}} = 1 - \exp\left\{-\frac{-2l_S(\mathbf{B}^{(0)}) - (-2l_S(\hat{\mathbf{B}}))}{\hat{N}}\right\}.$$

Nagelkerke’s R Square

$$R_N^2 = \frac{R_{\mathrm{CS}}^2}{1 - \{\hat{L}_U(\mathbf{B}^{(0)})\}^{2/\hat{N}}} .$$

McFadden’s R Square

$$R_{\mathrm{M}}^2 = 1 - \frac{l_S(\hat{\mathbf{B}})}{l_S(\mathbf{B}^{(0)})}$$

Hypothesis Testing

Contrasts defined as linear combination of regression parameters can be tested. Given matrix $\mathbf{L}$ with r rows and $p(K-1)$ columns, and vector $\mathbf{K}$ with r elements, CSLogistic tests the linear hypothesis $H_0 : \mathbf{LB} = \mathbf{K}$. See “Complex Samples: Model Testing” (cs_modeltesting.pdf) for details.

Custom tests

For a user specified $\mathbf{L}$ and $\mathbf{K}$, $H_0 : \mathbf{LB} = \mathbf{K}$ is tested only when it is testable, i.e. when $\mathbf{LB}$ is estimable. Let $\mathbf{L} = (\mathbf{L}_1, \cdots, \mathbf{L}_{K-1})$, where each $\mathbf{L}_k$ is a *r* by *p* matrix. The $\mathbf{LB}$ is estimable if for every $k = 1, \cdots, K-1$,

$$\mathbf{L}_k = \mathbf{L}_k \mathbf{H},$$

where $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-} \mathbf{X}'\mathbf{X}$ is a $p \times p$ matrix.

Note: In NOMREG, only block diagonal matrices such as $\mathbf{L} = diag(\mathbf{L}^*, \cdots, \mathbf{L}^*)$ are considered, where $\mathbf{L}^*$ is a $q \times p$ matrix. Also in NOMREG, testability is not checked.

Default tests of Model effects

For each effect specified in the model, matrix $\mathbf{L} = diag(\mathbf{L}^*, \cdots, \mathbf{L}^*)$ is constructed and $H_0 : \mathbf{LB} = \mathbf{0}$ is tested. The matrix $\mathbf{L}^*$ is chosen to be the type III test matrix constructed based on matrix $\mathbf{H} = (\mathbf{X}'\mathbf{X})^{-} \mathbf{X}'\mathbf{X}$. This construction procedure makes sure that $\mathbf{LB}$ is estimable. It involves parameters only for the given effect and the effects containing the given effect. It does not depend on the order of effects specified in the model. If such a matrix cannot be constructed, the effect is not testable.

Predicted values

For a predictor pattern $\mathbf{x}$, the predicted probability of each response category is

$$\hat{p}_{\mathbf{x}}(k) = \begin{cases} \dfrac{\exp(\mathbf{x}'\hat{\mathbf{B}}_k)}{1+\sum_{k=1}^{K-1}\exp(\mathbf{x}'\hat{\mathbf{B}}_k)} & k = 1,\cdots,K-1 \\ \dfrac{1}{1+\sum_{k=1}^{K-1}\exp(\mathbf{x}'\hat{\mathbf{B}}_k)} & k = K \end{cases}.$$

The predicted category $c(\mathbf{x})$ is the one with the highest predicted probability, i.e.

$$c(\mathbf{x}) = \arg\max_k \hat{p}_{\mathbf{x}}(k).$$

Equivalently,

$$c(\mathbf{x}) = \arg\max_k \left(\mathbf{x}'\hat{\mathbf{B}}_k\right)$$

where $\hat{\mathbf{B}}_K = 0$ is set for the last (reference) response category. This latter formula is less likely to have numerical problems and should be used.

Classification table

A two-way table with (*i*,*j*)-th element being the counts or the sum of weights for the observations whose actual response category is *i* (as row) and predicted response category is *j* (as column) respectively.

Odds ratio

The ratio of odds at $\mathbf{x}_1$ to odds at $\mathbf{x}_2$ for response category k_1 versus k_2 is

$$or(\mathbf{x}_1,\mathbf{x}_2;k_1,k_2) = \frac{p_{\mathbf{x}_1}(k_1)/p_{\mathbf{x}_1}(k_2)}{p_{\mathbf{x}_2}(k_1)/p_{\mathbf{x}_2}(k_2)} = \exp\left((\mathbf{x}_1-\mathbf{x}_2)'(\mathbf{B}_{k_1}-\mathbf{B}_{k_2})\right)$$

For $k_1 = k$ and $k_2 = K$ (the reference response category), odds ratio is simplified as

$$or(\mathbf{x}_1,\mathbf{x}_2;k,K) = \exp\left((\mathbf{x}_1-\mathbf{x}_2)'\mathbf{B}_k\right).$$

Equation for $or(\mathbf{x}_1,\mathbf{x}_2;k,K)$ will be the one we use to calculate odds ratios. The estimate and confidence interval for $or(\mathbf{x}_1,\mathbf{x}_2;k,K)$ are respectively

$$\exp\left((\mathbf{x}_1-\mathbf{x}_2)'\hat{\mathbf{B}}_k\right),$$

and

$$\left[\exp\left(\hat{C}\ -t_{df,1-\frac{\alpha}{2}}se(\hat{C})\right),\exp\left(\hat{C}+t_{df,1-\frac{\alpha}{2}}se(\hat{C})\right)\right]$$

where

$$\hat{C}=(\mathbf{x}_1-\mathbf{x}_2)'\hat{\mathbf{B}}_k\ ,$$

$$se(\hat{C})=\sqrt{(\mathbf{x}_1-\mathbf{x}_2)'Var\left(\hat{\mathbf{B}}_k\right)(\mathbf{x}_1-\mathbf{x}_2)}\ .$$

$\exp\left(B_{kj}\right)$ and its cofidence interval

$\exp\left(B_{kj}\right)$ can be interpreted as an odds ratio for main effects model. SUDAAN calls $\exp\left(B_{kj}\right)$ the odds ratio for parameter B_{kj} whether or not there is an interaction effect in the model. Even though they may not be odds ratios for models with interaction effects, they are still of interest. For each $\exp\left(B_{kj}\right)$, its $1-\alpha$ confidence interval is

$$\left[\exp\left(L(\hat{B}_{kj})\right),\exp\left(U(\hat{B}_{kj})\right)\right],$$

where $L(\hat{B}_{kj}),U(\hat{B}_{kj})$ are the lower and upper confidence limits for census parameter B_{kj} .

Subpopulation estimates

When analyses are requested for a given subpopulation *D*, we perform calculations on the following redefined $\mathbf{x}_i$ and $y_i(k)$:

$$\mathbf{x}_i\ =\mathbf{x}_i\delta_i(D)$$
$$y_i(k)=y_i(k)\delta_i(D)$$

where

$$\delta_i(D)=\begin{cases}1 & \text{if the sample unit } i \text{ is in the subpopulation D}\\ 0 & \text{otherwise}\end{cases}\ .$$

When computing point estimates, this substitution is equivalent to including only the subpopulation elements in the calculations. This is in contrast to computing the variance estimates where all elements in the sample need to be included.

Missing values

Missing values are handled using list-wise deletion; that is, any case without valid data on any design, dependent, or independent variable is excluded from the analysis.

References

Agresti, A. (2002), *Categorical Data Analysis, second edition*. Wiley

Albert, A. and Anderson, J.A. (1984), "On the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 71, 1 -10.

Binder, D. A. (1983), “On the variances of asymptotically normal estimators from complex surveys”, *International Statistical Review*, 51, 279-292.

Chambers R. L. and Skinner, C. J. (eds.) (2003), *Analysis of Survey Data*. Chichester:Wiley.

Santner, T.J. and Duffy, E.D. (1986), "A Note on A. Albert and J.A. Anderson's Conditions for the Existence of Maximum Likelihood Estimates in Logistic Regression Models," *Biometrika*, 73, 755 -758