# LOGISTIC REGRESSION

Logistic regression regresses a dichotomous dependent variable on a set of independent variables. Several methods are implemented for selecting the independent variables.

## Notation

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

| | |
|---|---|
| $n$ | The number of observed cases |
| $p$ | The number of parameters |
| $\mathbf{y}$ | $n \times 1$ vector with element $y_i$, the observed value of the $i$th case of the dichotomous dependent variable |
| $\mathbf{X}$ | $n \times p$ matrix with element $x_{ij}$, the observed value of the $i$th case of the $j$th parameter |
| $\beta$ | $p \times 1$ vector with element $\beta_j$, the coefficient for the $j$th parameter |
| $\mathbf{w}$ | $n \times 1$ vector with element $w_i$, the weight for the $i$th case |
| $l$ | Likelihood function |
| $L$ | Log likelihood function |
| $I$ | Information matrix |

## Model

The linear logistic model assumes a dichotomous dependent variable $Y$ with probability $\pi$, where for the $i$th case,

$$\pi_i = \frac{\exp(\eta_i)}{1+\exp(\eta_i)},$$

or

$$\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \eta_i = \mathbf{X}_i'\beta$$

Hence, the likelihood function $l$ for $n$ observations $y_1, \ldots, y_n$, with probabilities $\pi_1, \ldots, \pi_n$ and case weights $w_1, \ldots, w_n$, can be written as

$$l = \prod_{i=1}^{n} \pi_i^{w_i y_i} \left(1-\pi_i\right)^{w_i\left(1-y_i\right)}$$

It follows that the logarithm of $l$ is

$$L = \ln(l) = \sum_{i=1}^{n} \left(w_i y_i \ln(\pi_i) + w_i\left(1-y_i\right)\ln\left(1-\pi_i\right)\right)$$

and the derivative of $L$ with respect to $\beta_j$ is

$$L_{X_j}^{*} = \frac{\partial L}{\partial \beta_j} = \sum_{i=1}^{n} w_i\left(y_i - \pi_i\right)x_{ij}$$

# Maximum Likelihood Estimates (MLE)

The maximum likelihood estimates for $\beta$ satisfy the following equations

$$\sum_{i=1}^{n} w_i\left(y_i - \hat{\pi}_i\right)x_{ij} = 0, \text{ for the } j\text{th parameter}$$

where $x_{i0} = 1$ for $i = 1, \ldots, n$.

Note the following:

(1) A Newton-Raphson type algorithm is used to obtain the MLEs. Convergence can be based on

   (a) Absolute difference for the parameter estimates between the iterations

   (b) Percent difference in the log-likelihood function between successive iterations

   (c) Maximum number of iterations specified

(2) During the iterations, if $\hat{\pi}_i(1-\hat{\pi}_i)$ is smaller than $10^{-8}$ for all cases, the log-likelihood function is very close to zero. In this situation, iteration stops and the message "All predicted values are either 1 or 0" is issued.

After the maximum likelihood estimates $\hat{\beta}$ are obtained, the asymptotic covariance matrix is estimated by $I^{-1}$, the inverse of the information matrix *I*, where

$$I = -\left[ E\left( \frac{\partial^2 L}{\partial \beta_i \partial \beta_j} \right) \right] = \mathbf{X}'\mathbf{W}\hat{\mathbf{V}}\mathbf{X},$$

$$\hat{\mathbf{V}} = \text{Diag}\left\{ \hat{\pi}_1(1-\hat{\pi}_1), \ldots, \hat{\pi}_n(1-\hat{\pi}_n) \right\},$$

$$\mathbf{W} = \text{Diag}\{ w_1, \ldots, w_n \},$$

$$\hat{\pi}_i = \frac{\exp(\hat{\eta}_i)}{1+\exp(\hat{\eta}_i)},$$

and

$$\hat{\eta}_i = \mathbf{X}_i'\hat{\beta}.$$

# Stepwise Methods of Selecting Variables

Several methods are available for selecting independent variables. With the forced entry method, any variable in the variable list is entered into the model. There are two stepwise methods: forward and backward. The stepwise methods can use either the Wald statistic, the likelihood ratio, or a conditional algorithm for variable removal. For both stepwise methods, the score statistic is used to select variables for entry into the model.

Three statistics used later are defined as follows:

## Score Statistic

The score statistic is calculated for each variable not in the model to determine whether the variable should enter the model. Assume that there are $r_1$ variables, namely, $\alpha_1, \ldots, \alpha_{r_1}$ in the model and $r_2$ variables, $\gamma_1, \ldots, \gamma_{r_2}$, not in the model. The score statistic for $\gamma_i$ is defined as

$$\mathbf{S}_i = \left(\mathbf{L}^*_{\gamma_i}\right)^2 \mathbf{B}_{22,i}$$

if $\gamma_i$ is not a categorical variable. If $\gamma_i$ is a categorical variable with *m* categories, it is converted to a $(m-1)$-dimension dummy vector. Denote these new $m-1$ variables as $\tilde{\gamma}_i, \ldots, \tilde{\gamma}_{i+m-2}$. The score statistic for $\gamma_i$ is then

$$\mathbf{S}_i = \left(\mathbf{L}^*_{\tilde{\gamma}}\right)' \mathbf{B}_{22,i} \mathbf{L}^*_{\tilde{\gamma}}$$

where $\mathbf{L}^*_{\tilde{\gamma}}{}' = \left(L^*_{\tilde{\gamma}_i}, \ldots, L^*_{\tilde{\gamma}_{i+m-2}}\right)$ and the $(m-1)\times(m-1)$ matrix $\mathbf{B}_{22,i}$ is

$$\mathbf{B}_{22,i} = \left(\mathbf{A}_{22,i} - \mathbf{A}_{21,i}\mathbf{A}_{11}^{-1}\mathbf{A}_{12,i}\right)^{-1}$$

with

$$\mathbf{A}_{11} = \underset{\sim}{\alpha}' \hat{\mathbf{V}} \underset{\sim}{\alpha},$$

$$\mathbf{A}_{12,i} = \underset{\sim}{\alpha}' \hat{\mathbf{V}} \underset{\sim}{\gamma}_i,$$

$$\mathbf{A}_{22,i} = \underset{\sim}{\gamma}_i' \hat{\mathbf{V}} \underset{\sim}{\gamma}_i$$

in which $\underset{\sim}{\alpha}$ is the design matrix for variables $\alpha_1, \ldots, \alpha_{r_1}$ and $\underset{\sim}{\gamma}_i$ is the design matrix for dummy variables $\tilde{\gamma}_i, \ldots, \tilde{\gamma}_{i+m-2}$. Note that $\underset{\sim}{\alpha}$ contains a column of ones unless the constant term is excluded from $\eta$. Based on the MLEs for the parameters in the model, $\mathbf{V}$ is estimated by $\hat{\mathbf{V}} = \text{Diag}\left\{\hat{\pi}_1(1-\hat{\pi}_1), \ldots, \hat{\pi}_n(1-\hat{\pi}_n)\right\}$. The asymptotic distribution of the score statistic is a chi-square with degrees of freedom equal to the number of variables involved.

Note the following:

(1) If the model is through the origin and there are no variables in the model, $\mathbf{B}_{22,i}$ is defined by $\mathbf{A}_{22,i}^{-1}$ and $\hat{\mathbf{V}}$ is equal to $\frac{1}{4}\mathbf{I}_n$.

(2) If $\mathbf{B}_{22,i}$ is not positive definite, the score statistic and residual chi-square statistic are set to be zero.

## Wald Statistic

The Wald statistic is calculated for the variables in the model to determine whether a variable should be removed. If the *i*th variable is not categorical, the Wald statistic is defined by

$$Wald_i = \frac{\hat{\beta}_i^2}{\hat{\sigma}^2_{\hat{\beta}_i}}$$

If it is a categorical variable, the Wald statistic is computed as follows:

Let $\hat{\beta}_i$ be the vector of maximum likelihood estimates associated with the $m-1$ dummy variables, and $\mathbf{C}$ the asymptotic covariance matrix for $\hat{\beta}_i$. The Wald statistic is

$$Wald_i = \hat{\beta}_i' \mathbf{C}^{-1} \hat{\beta}_i$$

The asymptotic distribution of the Wald statistic is chi-square with degrees of freedom equal to the number of parameters estimated.

## Likelihood Ratio (LR) Statistic

The LR statistic is defined as two times the log of the ratio of the likelihood functions of two models evaluated at their MLEs. The LR statistic is used to determine if a variable should be removed from the model. Assume that there are $r_1$ variables in the current model which is referred to as a full model. Based on the MLEs of the full model, *l*(*full*) is calculated. For each of the variables removed from the full model one at a time, MLEs are computed and the likelihood function *l*(*reduced*) is calculated. The LR statistic is then defined as

$$LR = -2\ln\left(\frac{l(reduced)}{l(full)}\right) = -2\big(L(reduced) - L(full)\big)$$

LR is asymptotically chi-square distributed with degrees of freedom equal to the difference between the numbers of parameters estimated in the two models.

## Conditional Statistic

The conditional statistic is also computed for every variable in the model. The formula for the conditional statistic is the same as the LR statistic except that the

parameter estimates for each reduced model are conditional estimates, not MLEs. The conditional estimates are defined as follows. Let $\hat{\beta} = \left(\hat{\beta}_1, \ldots, \hat{\beta}_{r_1}\right)$ be the MLE for the $r_1$ variables in the model and **C** be the asymptotic covariance matrix for $\hat{\beta}$. If variable $x_i$ is removed from the model, the conditional estimate for the parameters left in the model given $\hat{\beta}$ is

$$\tilde{\beta}_{(i)} = \hat{\beta}_{(i)} - \mathbf{c}_{12}^{(i)}\left(\mathbf{c}_{22}^{(i)}\right)^{-1}\hat{\beta}_i$$

where $\hat{\beta}_i$ is the MLE for the parameter(s) associated with $x_i$ and $\hat{\beta}_{(i)}$ is $\hat{\beta}$ with $\hat{\beta}_i$ removed, $\mathbf{c}_{12}^{(i)}$ is the covariance between $\hat{\beta}_{(i)}$ and $\hat{\beta}_i$, and $\mathbf{c}_{22}^{(i)}$ is the covariance of $\hat{\beta}_i$. Then the conditional statistic is computed by

$$-2\left(L\left(\tilde{\beta}_{(i)}\right) - L(full)\right)$$

where $L\left(\tilde{\beta}_{(i)}\right)$ is the log likelihood function evaluated at $\hat{\beta}_{(i)}$.

# Stepwise Algorithms

### Forward Stepwise (FSTEP)

(1) If FSTEP is the first method requested, estimate the parameter and likelihood function for the initial model. Otherwise, the final model from the previous method is the initial model for FSTEP. Obtain the necessary information: MLEs of the parameters for the current model, predicted probability $\hat{\pi}_i$, likelihood function for the current model, and so on.

(2) Based on the MLEs of the current model, calculate the score statistic for every variable eligible for inclusion and find its significance.

(3) Choose the variable with the smallest significance. If that significance is less than the probability for a variable to enter, then go to step 4; otherwise, stop FSTEP.

(4) Update the current model by adding a new variable. If this results in a model which has already been evaluated, stop FSTEP.

(5) Calculate LR or Wald statistic or conditional statistic for each variable in the current model. Then calculate its corresponding significance.

(6) Choose the variable with the largest significance. If that significance is less than the probability for variable removal, then go back to step 2; otherwise, if

the current model with the variable deleted is the same as a previous model, stop FSTEP; otherwise, go to the next step.

(7) Modify the current model by removing the variable with the largest significance from the previous model. Estimate the parameters for the modified model and go back to step 5.

**Backward Stepwise (BSTEP)**

(1) Estimate the parameters for the full model which includes the final model from previous method and all eligible variables. Only variables listed on the BSTEP variable list are eligible for entry and removal. Let the current model be the full model.

(2) Based on the MLEs of the current model, calculate the LR or Wald statistic or conditional statistic for every variable in the model and find its significance.

(3) Choose the variable with the largest significance. If that significance is less than the probability for a variable removal, then go to step 5; otherwise, if the current model without the variable with the largest significance is the same as the previous model, stop BSTEP; otherwise, go to the next step.

(4) Modify the current model by removing the variable with the largest significance from the model. Estimate the parameters for the modified model and go back to step 2.

(5) Check to see any eligible variable is not in the model. If there is none, stop BSTEP; otherwise, go to the next step.

(6) Based on the MLEs of the current model, calculate the score statistic for every variable not in the model and find its significance.

(7) Choose the variable with the smallest significance. If that significance is less than the probability for variable entry, then go to the next step; otherwise, stop BSTEP.

(8) Add the variable with the smallest significance to the current model. If the model is not the same as any previous models, estimate the parameters for the new model and go back to step 2; otherwise, stop BSTEP.

# Statistics

### Initial Model Information

If $\beta_0$ is not included in the model, the predicted probability is estimated to be 0.5 for all cases and the log likelihood function $L(0)$ is

$$L(0) = W\ln(0.5) = -0.6931472W$$

with $W = \sum_{i=1}^{n} w_i$. If $\beta_0$ is included in the model, the predicted probability is estimated as

$$\hat{\pi}_0 = \frac{\sum_{i=1}^{n} w_i y_i}{W}$$

and $\beta_0$ is estimated by

$$\hat{\beta}_0 = \ln\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right)$$

with asymptotic standard error estimated by

$$\hat{\sigma}_{\hat{\beta}_0} = \frac{1}{\sqrt{W\hat{\pi}_0(1-\hat{\pi}_0)}}$$

The log likelihood function is

$$L(0) = W\left[\hat{\pi}_0 \ln\left(\frac{\hat{\pi}_0}{1-\hat{\pi}_0}\right) + \ln(1-\hat{\pi}_0).\right]$$

### Model Information

The following statistics are computed if a stepwise method is specified.

**(a) -2 Log Likelihood**

$$-2\sum_{i=1}^{n}\left(w_i y_i \ln(\hat{\pi}_i) + w_i(1-y_i)\ln(1-\hat{\pi}_i)\right)$$

**(b) Model Chi-Square**

2(log likelihood function for current model - log likelihood function for initial model)

The initial model contains a constant if it is in the model; otherwise, the model has no terms. The degrees of freedom for the model chi-square statistic is equal to the difference between the numbers of parameters estimated in each of the two models. If the degrees of freedom is zero, the model chi-square is not computed.

**(c) Block Chi-Square**

2(log likelihood function for current model - log likelihood function for the final model from the previous method.)

The degrees of freedom for the block chi-square statistic is equal to the difference between the numbers of parameters estimated in each of the two models.

**(d) Improvement Chi-Square**

2(log likelihood function for current model - log likelihood function for the model from the last step )

The degrees of freedom for the improvement chi-square statistic is equal to the difference between the numbers of parameters estimated in each of the two models.

**(e) Goodness of Fit**

$$\sum_{i=1}^{n}\frac{w_i(y_i-\hat{\pi}_i)^2}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

**(f) Cox and Snell's $R^2$ (Cox and Snell, 1989; Nagelkerke, 1991)**

$$R^2_{CS} = 1 - \left( \frac{l(0)}{l(\hat{\beta})} \right)^{\frac{2}{W}}$$

where $l\left(\hat{\beta}\right)$ is the likelihood of the current model and $l(0)$ is the likelihood of the initial model; that is, $l(0) = W\log(0.5)$ if the constant is not included in the model; $l(0) = W\left[\hat{\pi}_o \log\left\{\hat{\pi}_o / (1-\hat{\pi}_o)\right\} + \log(1-\hat{\pi}_o)\right]$ if the constant is included in the model, where $\hat{\pi}_o = \sum_i^n w_i y_i / W$.

**(g) Nagelkerke's $R^2$ (Nagelkerke, 1981)**

$$R^2_N = R^2_{CS} / \max\left(R^2_{CS}\right)$$

where $\max\left(R^2_{CS}\right) = 1 - \left\{l(0)\right\}^{2/W}$.

**Hosmer-Lemeshow Goodness-of-Fit Statistic**

The test statistic is obtained by applying a chi-square test on a $2 \times g$ contingency table. The contingency table is constructed by cross-classifying the dichotomous dependent variable with a grouping variable (with $g$ groups) in which groups are formed by partitioning the predicted probabilities using the percentiles of the predicted event probability. In the calculation, approximately 10 groups are used ($g = 10$). The corresponding groups are often referred to as the "deciles of risk" (Hosmer and Lemeshow , 1989).

If the values of independent variables for observation $i$ and $i$' are the same, observation $i$ and $i$' are said to be in the same block. When one or more blocks occur within the same decile, the blocks are assigned to this same group. Moreover, observations in the same block are not divided when they are placed into groups. This strategy may result in fewer than 10 groups (that is, $g \le 10$) and consequently, fewer degrees of freedom.

Suppose that there are $Q$ blocks, and the $q$th block has $m_q$ number of observations, $q = 1, \ldots, Q$. Moreover, suppose that the $k$th group ($k = 1, \ldots, g$) is composed of the $q_1$th, …, $q_k$th blocks of observations. Then the total number of observations in the $k$th group is $s_k = \sum_{q_1}^{q_k} m_j$. The total observed frequency of events (that is, $Y = 1$) in the $k$th group, call it $O_{1k}$, is the total number of observations in the $k$th group with $Y = 1$. Let $E_{1k}$ be the total expected frequency of

the event (that is, $Y = 1$) in the *k*th group; then $E_{1k}$ is given by $E_{1k} = s_k \xi_k$, where $\xi_k$ is the average predicted event probability for the *k*th group.

$$\xi_k = \sum_{q_1}^{q_k} m_j \hat{\pi}_j \; / \; s_k$$

The Hosmer-Lemeshow goodness-of-fit statistic is computed as

$$\chi^2_{HL} = \sum_{k=1}^{g} \frac{(O_{1k} - E_{1k})^2}{E_{1k}(1 - \xi_k)}$$

The *p* value is given by $Pr\left(\chi^2 \geq \chi^2_{HL}\right)$ where $\chi^2$ is the chi-square statistic distributed with degrees of freedom $(g - 2)$.

### Information for the Variables Not in the Equation

For each of the variables not in the equation, the score statistic is calculated along with the associated degrees of freedom, significance and partial *R*. Let $X_i$ be a variable not currently in the model and $S_i$ the score statistic. The partial *R* is defined by

$$Partial\_R = \begin{cases} \sqrt{\dfrac{S_i - 2 \times df}{-2L(initial)}} & \text{if } S_i > 2 \times df \\ 0 & \text{otherwise} \end{cases}$$

where $df$ is the degrees of freedom associated with $S_i$, and $L(initial)$ is the log likelihood function for the initial model.

The residual Chi-Square printed for the variables not in the equation is defined as

$$R_{CS} = \left(L^*_\gamma\right)' B_{22} L^*_\gamma$$

where $L^*_\gamma = \left(L^*_{\gamma_1}, \ldots, L^*_{\gamma_{r_2}}\right)'$.

### Information for the Variables in the Equation

For each of the variables in the equation, the MLE of the Beta coefficients is calculated along with the standard errors, Wald statistics, degrees of freedom, significances, and partial *R*. If $X_i$ is not a categorical variable currently in the equation, the partial *R* is computed as

$$Partial\_R = \begin{cases} \text{sign}\left(\hat{\beta}_i\right)\sqrt{\dfrac{Wald_i - 2}{-2L(initial)}} & \text{if } Wald_i > 2 \\ 0 & \text{otherwise} \end{cases}$$

If $X_i$ is a categorical variable with *m* categories, the partial *R* is then

$$Partial\_R = \begin{cases} \sqrt{\dfrac{Wald_i - 2(m-1)}{-2L(initial)}} & \text{if } Wald_i > 2(m-1) \\ 0 & \text{otherwise} \end{cases}$$

### Casewise Statistics

#### (a) Individual Deviance

The deviance of the *i*th case, $G_i$, is defined as

$$G_i = \begin{cases} \sqrt{2\left(y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)\right)} & \text{if } y_i > \hat{\pi}_i \\ -\sqrt{2\left(y_i \ln(\hat{\pi}_i) + (1-y_i)\ln(1-\hat{\pi}_i)\right)} & \text{otherwise} \end{cases}$$

#### (b) Leverage

The leverage of the *i*th case, $h_i$, is the *i*th diagonal element of the matrix

$$\hat{\mathbf{V}}^{\frac{1}{2}}\mathbf{X}\left(\mathbf{X}'\mathbf{C}\hat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}'\hat{\mathbf{V}}^{\frac{1}{2}}$$

where

$$\hat{\mathbf{V}} = \text{Diag}\left\{\hat{\pi}_1(1-\hat{\pi}_1), \ldots, \hat{\pi}_n(1-\hat{\pi}_n)\right\}$$

**(c) Studentized Residual**

$$\tilde{G}_i^* = \frac{G_i}{\sqrt{1-h_i}}$$

**(d) Logit Residual**

$$\tilde{e}_i = \frac{e_i}{\hat{\pi}_i(1-\hat{\pi}_i)}$$

where $e_i = y_i - \hat{\pi}_i$.

**(e) Standardized Residual**

$$z_i = \frac{e_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}$$

**(f) Cook's Distance**

$$D_i = \frac{z_i^2 h_i}{1-h_i}$$

**(g) DFBETA**

Let $\Delta\beta_i$ be the change of the coefficient estimates from the deletion of case *i*. It is computed as

$$\Delta\beta_i = \frac{\left(\mathbf{X}'\mathbf{C}\hat{\mathbf{V}}\mathbf{X}\right)^{-1}\mathbf{X}_i' e_i}{1-h_i}$$

**(h) Predicted Group**

If $\hat{\pi}_i \geq 0.5$, predicted group = group in which $y = 1$.

Note the following:

For the unselected cases with nonmissing values for the independent variables in the analysis, the leverage $\left(\tilde{h}_i\right)$ is computed as

$$\tilde{h}_i = h_i - \frac{\hat{V}_i h_i^2}{1 + \hat{V}_i h_i}$$

where

$$h_i = \hat{V}_i \mathbf{X}_i' \left(\mathbf{X}'\mathbf{C}\hat{\mathbf{V}}\mathbf{X}\right)^{-1} \mathbf{X}_i$$

For the unselected cases, the Cook's distance and DFBETA are calculated based on $\tilde{h}_i$.