Chapter 3: Logistic Regression

1. Regression with a binary response

Consider the model

$$Y_i \;=\; \beta_0 \,+\, \beta_1 x_{i1} \,+\, \beta_2 x_{i2} \,+\, \ldots \,+\, \beta_p x_{ip} \,+\, \varepsilon_i \qquad \begin{cases} (i = 1, 2, \ldots, n) \\ Y_i = 0, 1 \end{cases}$$

1. The normality assumption does not satisfy.

   ε_i cannot be continuous since only two values are possible; namely

   $$\varepsilon_i \;=\; Y_i - [\beta_0 \,+\, \beta_1 x_{i1} \,+\, \beta_2 x_{i2} \,+\, \ldots \,+\, \beta_p x_{ip}]$$

   As a result, there can be no assumption of normality on the model errors.

2. The homogeneous assumption of variances does not satisfy.

   If we assume the usual $E(\varepsilon_i) = 0$, we have

   $$E(Y_i) \;=\; \beta_0 \,+\, \beta_1 x_{i1} \,+\, \beta_2 x_{i2} \,+\, \ldots \,+\, \beta_p x_{ip}$$

   Now, we may view $E(Y_i) = P_i$ as the population proportion of observations at x_{i1}, x_{i2}, $\ldots x_{ip}$ for which $Y = 1$. In other words,

   $$\left.\begin{array}{rcl} P_i & = & \text{Prob}(Y_i = 1) \\ Q_i = 1 - P_i & = & \text{Prob}(Y_i = 0) \end{array}\right\} \; (i = 1, 2, \ldots, n)$$

   Thus for n distinct data points there are n probabilities P_1, P_2, $\ldots P_n$, each of which is a parameter of a *Bernoulli distribution.* Thus,

   $$\text{Var}(\varepsilon_i) \;=\; P_i Q_i$$

   Since P_i varies with levels of regressor variables, the error variance is not homogeneous.

3. Fundamental problem with the assumption of linearity.

   $$E(Y_i) \;=\; \beta_0 \,+\, \beta_1 x_{i1} \,+\, \beta_2 x_{i2} \,+\, \ldots \,+\, \beta_p x_{ip}$$

   The left-hand side is restricted to the interval [0, 1] and the right-hand side may take a real value outside the interval [0, 1], especially for large or small values of x_i, because the unknown parameters $\boldsymbol{\beta}$ are allowed to vary freely. They cannot be equal in general.

To solve the problem, use a transformation (or often called link function) g to describe the relationship between the mean $\mu(\boldsymbol{x}_i)$ of Y_i and $\boldsymbol{x}_i^T\boldsymbol{\beta}$, i.e.,

$$g(\mu(\boldsymbol{x}_i)) = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

This approach indeed is the essence of the so-called generalized linear models (hereafter GLMs).

2. Logistic regression A very popular and useful model to accommodate this binary response situation is the *logistic regression model* given by

$$P(\boldsymbol{x}_i) \;=\; \frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}} \quad (i=1,2,\ldots s)$$

From it, we can write

$$\log\left[\frac{P(\boldsymbol{x}_i)}{1-P(\boldsymbol{x}_i)}\right] \;=\; \beta_0 \;+\; \beta_1 x_{i1} \;+\; \beta_2 x_{i2} \;+\; \ldots \;+\; \beta_p x_{ip}$$

We call the link function, g, logit.

Suppose there exists a set of data that takes the form

$$\left.\begin{array}{cccccc} n_1 & r_1 & x_{11} & x_{12} & \ldots & x_{1p} \\ n_2 & r_2 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots & & \vdots \\ n_s & r_s & x_{s1} & x_{s2} & \ldots & x_{sp} \end{array}\right\} \; (s > p')$$

Here, r_1, r_2, …, r_s represents the number of successes in n_1, n_2, …, n_s trials respectively.

Now, the information at each combination allows for estimation of the $P(\boldsymbol{x}_i)$ probability values, i.e.,

$$\hat{P}(\boldsymbol{x}_i) \;=\; \frac{r_i}{n_i} \quad (i=1,2,\ldots,s)$$

So, we *regress*

$$\log\left(\frac{\hat{P}(\boldsymbol{x}_i)}{1-\hat{P}(\boldsymbol{x}_i)}\right) \quad \text{against} \quad x_1, x_2, \ldots, x_p$$

Thus we may view the model as follows:

$$\log\left(\frac{\hat{P}(\boldsymbol{x}_i)}{1-\hat{P}(\boldsymbol{x}_i)}\right) \;=\; \beta_0 \;+\; \beta_1 x_{i1} \;+\; \beta_2 x_{i2} \;+\; \ldots \;+\; \beta_p x_{ip} \;+\; \varepsilon_i \quad (i=1,2,\ldots,s)$$

However, at a fixed combination $\boldsymbol{x}_i$,

$$\operatorname{Var}\log\left(\frac{\hat{P}(\boldsymbol{x}_i)}{1-\hat{P}(\boldsymbol{x}_i)}\right) \approx \frac{1}{n_i P(\boldsymbol{x}_i)(1-P(\boldsymbol{x}_i))}$$

As a result, weighted regression with weight at the ith data point

$$w_i \;=\; n_i \hat{P}(\boldsymbol{x}_i)(1-\hat{P}(\boldsymbol{x}_i))$$

is an approach that one must consider.

The preceding description of the use of weighted least squares to fit a logistic regression model is reasonable when the number of observations at the individual $\boldsymbol{x}_i$ is not small. There are two reasons.

First, weighted regression should be avoided when weights are estimated with a relatively small amount of information.

Secondly, the variance of

$$\log\left(\frac{\hat{P}(\boldsymbol{x}_i)}{1-\hat{P}(\boldsymbol{x}_i)}\right)$$

is only an approximation, and the result is most accurate when the n_i are large.

3. Weighted least squares

Let us recall the general linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

with the ordinary least squares estimator given by

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

Suppose, however, we relax the assumption that $\mathrm{Var}(\boldsymbol{\varepsilon}) = \sigma^2\boldsymbol{I}_n$ and assume instead that there is a positive definite matrix $\boldsymbol{V}$ for which

$$\mathrm{Var}(\boldsymbol{\varepsilon}) = \boldsymbol{V}$$

By minimizing

$$SS_{\mathrm{Res},\boldsymbol{V}} = (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T\,\boldsymbol{V}^{-1}\,(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})\,,$$

the appropriate estimator is the *generalized least squares estimator* given by

$$\tilde{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{y}$$

1. The estimator $\tilde{\boldsymbol{\beta}}$ is unbiased, i.e. $E(\tilde{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
2. $\tilde{\boldsymbol{\beta}}$ is a maximum likelihood estimator under normality conditions on ε, i.e. if $\varepsilon \sim N(\boldsymbol{0},\boldsymbol{V})$.
3. The estimators in $\tilde{\boldsymbol{\beta}}$ achieve minimum variance of all unbiased estimators under the condition $\varepsilon \sim N(\boldsymbol{0},\boldsymbol{V})$.
4. If the assumption of normality is relaxed, the estimators in $\tilde{\boldsymbol{\beta}}$ achieve the minimum variance of all linear unbiased estimators.
5. $\mathrm{Var}(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{X}^T\boldsymbol{V}^{-1}\boldsymbol{X})^{-1}$.

Suppose we assume that the model errors are uncorrelated but the homogeneous variance assumption does not hold. In other words, the error variances at the n data points are σ_1^2, σ_2^2, $\ldots$, σ_n^2. The $\boldsymbol{V}$ matrix is given by

$$\boldsymbol{V} = diag[\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2]$$

It is easy to verify that for this very important special case the generalized least squares estiamtor of $\boldsymbol{\beta}$ minimizes

$$SS_{\text{Res(weighted)}} = \sum_{i=1}^{n} w_i (y_i - \hat{y}_i)^2$$

where $w_i = 1/\sigma_i^2$.

4. Estimation: Method of maximum likelihood estimation

Likelihood function

1. Grouped data

   The likelihood at the ith group is given by

$$[P(\boldsymbol{x}_i)]^{r_i}[1 - P(\boldsymbol{x}_i)]^{n_i - r_i} = \left[\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right]^{r_i} \left[\frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right]^{n_i - r_i}$$

   The likelihood of the entire sample

$$\prod_{i=1}^{s} \left[\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right]^{r_i} \left[\frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right]^{n_i - r_i}$$

2. Ungrouped data

   For the case of ungrouped data, we assume that there are n observations, and n_1 of them are characterized as a success while $n - n_1$ are characterized as failures. Denoting the first n_1 observations as those with "success", we have the likelihood function,

$$\prod_{i=1}^{n_1} \left(\frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right) \prod_{i=n_1+1}^{n} \left(\frac{1}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}\right)$$

Interpretation of the regression coefficient in the logit regression model

1. Odds

   The odds in favor of the occurrence of A is $P(A)/[1 - P(A)]$, i.e.,

$$O(\boldsymbol{x}_i) = \frac{Pr(Y_i = 1|\boldsymbol{x}_i)}{1 - Pr(Y_i = 1|\boldsymbol{x}_i)}$$

2. Odds ratio

   It is the ratio of two odds. The odd ratio, i.e. the ratio of odds, that $Y = 1$ between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is

$$\frac{O(\boldsymbol{x}_i)}{O(\boldsymbol{x}_j)} = \frac{Pr(Y_i = 1|\boldsymbol{x}_i)}{1 - Pr(Y_i = 1|\boldsymbol{x}_i)} \times \frac{1 - Pr(Y_i = 1|\boldsymbol{x}_j)}{Pr(Y_i = 1|\boldsymbol{x}_j)}$$

According to the logistic regression model,

$$\log(O(\boldsymbol{x}_i)) = \beta_0 + \beta_1 x$$

In terms of odds ratio, we thus have

$$\frac{O(x+1)}{O(x)} = \exp(\beta_1) \iff \log\left(\frac{O(x+1)}{O(x)}\right) = \beta_1$$

Therefore, every one-unit increase in X_i (while holding other independent variables X_j, $j \neq i$, fixed for the multiple logitistic regression model) would lead to an amount of $\exp(\beta_i)$ change in the odds, or equivalently, an amount of β_i change in the log of the odds.

- If $\beta_1 = 0$, the odds and the probability are the same at all x ($\exp(\beta_1) = 1$).
- If $\beta_1 > 0$, the odds and the probability increase as x increases ($\exp(\beta_1) > 1$).
- If $\beta_1 < 0$, the odds and the probability decrease as x increases ($\exp(\beta_1) < 1$).

5. Hypothesis testing

1. Deviance goodness of fit test

   For testing

$$H_o : P(x_i) = \frac{e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i^T \boldsymbol{\beta}}},$$

The log of the ratio of these likelihood is an important measure of the goodness of fit of the logistic model. The statistic

$$\lambda(\boldsymbol{\beta}) = -2 \log \left[\frac{L(\hat{\boldsymbol{\beta}})}{L(\hat{\boldsymbol{P}})}\right]$$

is called a *likelihood ratio statistic* and is used to determine if the logistic model is a worthwhile description of the data as compared to the fit described by the more complete model $y_i = P_i + \varepsilon_i$. The test statistic $\lambda(\boldsymbol{\beta})$ is called the *deviance* associated with the fitted logistic regression.

- Ungrouped data case

  The likelihood is given by

$$L(\boldsymbol{P}) = \prod_{i=1}^{n} [P_i]^{y_i} [1 - P_i]^{1-y_i}$$

As a result, the maximum likelihood estimates are as expected, namely,

$$\begin{aligned} \hat{P}_i &= 1.0 \qquad \text{if} \quad y_i = 1.0 \\ &= 0 \qquad \text{if} \quad y_i = 0 \end{aligned}$$

While this model is senseless, it provides a "perfect fit" basis for determining if the regressors as a whole are effective.

We now have two likelihoods,

$$L(\hat{\boldsymbol{P}}) \;=\; \prod_{i=1}^{n}[y_i]^{y_i}[1-y_i]^{1-y_i}$$

and

$$L(\hat{\boldsymbol{\beta}}) \;=\; \frac{\prod\limits_{i=1}^{n_1}\left(e^{\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}\right)}{\prod\limits_{i=1}^{n}\left(1+e^{\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}\right)}$$

Under the null hypothesis, $\lambda(\boldsymbol{\beta})$ is approximately distributed as a χ^2 random variable with $n-(p+1)$ degrees of freedom.

If the ratio $L(\hat{\boldsymbol{\beta}})/L(\hat{\boldsymbol{P}})$ is sufficiently close to 1.0, one can reason that since $L(\hat{\boldsymbol{\beta}})$ is not significantly less than $L(\hat{\boldsymbol{P}})$, there is no significant lack of fit with the logistic regression model. Thus H_o is not rejected; i.e., the lack of fit is not significant if $\lambda(\boldsymbol{\beta})$ is sufficiently close to zero.

That is,

$$\text{Reject } H_o \text{ at the level of significant } \alpha \text{ if } \lambda(\boldsymbol{\beta}) > \chi^2_{\alpha,n-(p+1)}$$

- Grouped data

  In the case where the data are grouped, the concept is the same. However, the maximum likelihood estimate under the most complete model is $\hat{P}_i = r_i/n_i$ and hence (for s group)

$$L(\hat{\boldsymbol{P}}) \;=\; \prod_{i=1}^{s}\left(\frac{r_i}{n_i}\right)^{r_i}\left(\frac{n_i-r_i}{n_i}\right)^{n_i-r_i}$$

and

$$L(\hat{\boldsymbol{\beta}}) \;=\; \prod_{i=1}^{s}\left[\frac{e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right]^{r_i}\left[\frac{1}{1+e^{\boldsymbol{x}_i^T\boldsymbol{\beta}}}\right]^{n_i-r_i}$$

The statistic $\lambda(\boldsymbol{\beta})$ has an approximate χ^2 distribution with $s-p-1$ degrees of freedom.

The logistic model is found to be inappropriate if the deviance is sufficiently large.

2. Likelihood ratio test

   Inference concerning any regressor or subset can be computed by determining how much the presence of each regressor contributed to the reduction in deviance. Define

$$\lambda(\beta_j|\beta_0,\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p) \;=\; \lambda(\beta_0,\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p)-\lambda(\boldsymbol{\beta})$$

Thus, the difference, the amount of reduction in deviance attributed to $\beta_j x_j$, adjusted for the other regressors, is computed as

$$\lambda(\beta_j|\beta_0,\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p) \;=\; -2\log\left[\frac{L(\beta_0,\beta_1,\beta_2,\ldots,\beta_{j-1},\beta_{j+1},\ldots,\beta_p)}{L(\hat{\boldsymbol{\beta}})}\right]$$

The numerator likelihood is the maximum likelihood with $\beta_j x_j$ removed. The log likelihood ratio statistic

$$\lambda(\beta_j|\beta_0, \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$$

has an approximate χ^2-distribution with 1 degree of freedom under the appropriate hypothesis on β_j. That is, reject H_o if $\lambda(\beta_j|\beta_0, \beta_1, \beta_2, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p) > \chi^2_{\alpha,1}$.

Assume that $\boldsymbol{\beta}_1$ contains $r < p+1$ parameters. We reject H_o in the hypothesis

$$H_o : \boldsymbol{\beta}_1 = 0$$

$$H_1 : \boldsymbol{\beta}_1 \neq 0$$

if $\lambda(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) > \chi^2_{\alpha,r}$

3. <u>Wald test</u>

   We form the $\boldsymbol{C}$ matrix from negative the second partial derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$, i.e.,

$$c_{ii} = \frac{-\partial^2 \log L(\hat{\boldsymbol{\beta}})}{\partial \hat{\beta}_i^2} \quad (i = 0, 1, \ldots, p)$$

   and

$$c_{ij} = \frac{-\partial^2 \log L(\hat{\boldsymbol{\beta}})}{\partial \hat{\beta}_i \, \partial \hat{\beta}_j} \quad (i \neq j)$$

   The estimate of the variance-covariance matrix of the coefficients is the matrix $\boldsymbol{C}^{-1}$. As a result, an approximate χ^2_1 statistic for testing

$$H_o : \beta_j = 0$$

$$H_i : \beta_j \neq 0$$

   can be used by computing

$$\chi^2 = \frac{\hat{\beta}_j^2}{c^{jj}} \quad (j = 0, 1, \ldots, p)$$

   where the c^{jj} are the jth diagonal elements of $\boldsymbol{C}^{-1}$.

   Wald statistic for testing $\boldsymbol{L}^T\boldsymbol{\beta} = 0$ is defined by $(\boldsymbol{L}^T\hat{\boldsymbol{\beta}})^T(\boldsymbol{L}^T\boldsymbol{C}^{-1}\boldsymbol{L})^{-1}(\boldsymbol{L}^T\hat{\boldsymbol{\beta}})$ where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimates and $\boldsymbol{C}^{-1}$ is its estimated covariance matrix. It is χ^2_r , where r is the rank of $\boldsymbol{L}$.

4. <u>Score test</u>

   Score statistic is defined to be $\boldsymbol{U}^T(\boldsymbol{\beta}_0)\boldsymbol{I}^{-1}(\boldsymbol{\beta}_0)\boldsymbol{U}(\boldsymbol{\beta}_0)$. Under $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, score statistic $\sim \chi^2_r$ with r is the dimension of $\boldsymbol{\beta}_0$. $\boldsymbol{U}(\boldsymbol{\beta})$ is the vector of partial derivatives of the log-likelihood with respect to the parameter vector $\boldsymbol{\beta}$ and $\boldsymbol{I}(\boldsymbol{\beta})$ is the matrix of the negative second partial derivatives of the log-likelihood with respect to $\boldsymbol{\beta}$.

5. Interval estimation

Parameter	Estimate	Standard Error	Confidence Interval
$\omega = \sum_{i=0}^{p} a_i \beta_i$	$\hat{\omega} = \sum_{i=0}^{p} a_i \hat{\beta}_i$	$s.e.(\hat{\omega})$	$\hat{\omega} \pm 1.96 * s.e.(\hat{\omega}) = (\hat{\omega}_l, \hat{\omega}_u)$
$\exp(\omega)$	$\exp(\hat{\omega})$	$exp(\hat{\omega}) * s.e.(\hat{\omega})$	$(\exp(\hat{\omega}_l),\ \exp(\hat{\omega}_u))$
$P = \dfrac{\exp(\omega)}{1+\exp(\omega)}$	$\hat{P} = \dfrac{\exp(\hat{\omega})}{1+\exp(\hat{\omega})}$	$\dfrac{\exp(\hat{\omega})}{(1+\exp(\hat{\omega}))^2} * s.e.(\hat{\omega})$	$\left(\dfrac{\exp(\hat{\omega}_l)}{1+\exp(\hat{\omega}_l)}, \dfrac{\exp(\hat{\omega}_u)}{1+\exp(\hat{\omega}_u)}\right)$

where $s.e.(\hat{\omega}) \; = \; \sqrt{\sum_{i=0}^{p} a_i^2 Var(\beta_i) + \sum\sum_{j \neq k} a_j a_k Cov(\hat{\beta}_j, \hat{\beta}_k)}$

6. Measure of performance of the logistic model: Pseudo R^2

Pseudo R^2	Formula
McFadden	$R^2 = 1 - \dfrac{\log(L_M)}{\log(L_0)}$
McFadden (adjusted)	$R^2_{\text{adjusted}} = 1 - \dfrac{\log(L_M) - p}{\log(L_0)}$
Cox & Snell	$R^2 = 1 - \left(\frac{L_0}{L_M}\right)^{2/n}$
Nagelkerke / Cragg & Uhler	$R^2 = \dfrac{1 - \left(\dfrac{L_0}{L_M}\right)^{2/n}}{1 - L_0^{2/n}}$

where L_0: the value of the likelihood function for a model with no predictors and L_M be the likelihood for the model being estimated. R^2 by Cox & Snell can be converted to R^2 in linear regression model.

Another measure of performance suggested b Tjur (2009) is equal to the difference between means of predicted probabilities from each of two categories of the dependent variable. But, it cannot be generalized to ordinal or nominal logistic regression.