Spring 2015 Statistics 151 (Linear Models) : Lecture Twenty One

Aditya Guntuboyina

14 April 2015

1 Fitting the Logistic Regression Model to Data

Recall the logistic regression model.

We have binary responses $y_1, \ldots, y_n$ and data on p explanatory variables $x_{ij}, i = 1, \ldots, n$ and $j = 1, \ldots, p$. We assume that $y_1, \ldots, y_n$ are independent Bernoulli random variables with parameters $p_1, \ldots, p_n$. We model the relationship between the response and explanatory variables by the formula

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}. \tag{1}$$

$p_i/(1-p_i)$ denotes the odds of the event that $y_i = 1$. The interpretation of β_j is that it represents the increase in log-odds of the event that $y = 1$ for a unit increase in x_j when all other explanatory variables are held constant. In other words, e^{β_j} denotes the factor by which the odds of success (success means response equalling one) change for a unit increase in x_j (all other explanatory variables remaining unchanged).

Given data $y_1, \ldots, y_n$ and x_{ij} for $i = 1, \ldots, n$ and $j = 1, \ldots, p$, how can be estimate the parameters $\beta_0, \ldots, \beta_p$. Note that the model can alternatively be written as

$$y_i \sim Ber\left(\frac{\exp(\beta_0 + \sum_{j=1}^p \beta_j x_{ij})}{1 + \exp(\beta_0 + \sum_{j=1}^p \beta_j x_{ij})}\right)$$

with $y_1, \ldots, y_n$ being independent.

How to estimate $\beta_0, \ldots, \beta_p$? One simply uses Maximum Likelihood. The likelihood of $y_1, \ldots, y_n$ is

$$\prod_{i=1}^n p_i^{y_i}(1-p_i)^{1-y_i} \qquad \text{with } p_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip})}.$$

This likelihood is simply a function of $\beta_0, \ldots, \beta_p$ and so it can be maximized to yield estimates of $\beta_0, \ldots, \beta_p$. It is easier to work with the log-likelihood. The log-likelihood is given by

$$\begin{aligned}\ell(\beta) &= \sum_{i=1}^n \left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) \\ &= \sum_{i=1}^n \left[y_i(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}) - \log(1 + \exp\left(\beta_0 + \beta_1 x_{i1} \cdots + \beta_p x_{ip}\right))\right].\end{aligned}$$

Unfortunately, one cannot write down the minimizer for $\ell(\beta)$ in closed form. One therefore uses Newton's method.

Newton's method uses the iterative scheme

$$\beta^{(m+1)} = \beta^{(m)} - \left(H\ell(\beta^{(m)})\right)^{-1} \nabla\ell(\beta^{(m)}) \tag{2}$$

where $\nabla\ell(\beta)$ and $H\ell(\beta)$ denote the gradient and Hessian of the function $\ell(\beta)$ respectively: $\nabla\ell(\beta) := (\partial\ell(\beta)/\partial\beta_0, \ldots, \partial\ell(\beta)/\partial\ell(\beta_p))^T$ and $H\ell(\beta)$ is the $(p+1) \times (p+1)$ matrix whose entries are second order derivatives of $\ell(\beta)$. For example, the $(1,1)$th entry of $H\ell(\beta)$ is $\partial^2\ell(\beta)/\partial\beta_0^2$, the $(1,2)$th entry is $\partial^2\ell(\beta)/\partial\beta_0\partial\beta_1$ and so on.

It is quite easy to write down $\nabla\ell(\beta)$ and $H\ell(\beta)$. Check that

$$\nabla\ell(\beta) = \sum_{i=1}^{n}(y_i - p_i)(1, x_{i1}, \ldots, x_{ip})^T$$

and

$$H\ell(\beta) = -\sum_{i=1}^{n} p_i(1-p_i)(1, x_{i1}, \ldots, x_{ip})^T(1, x_{i1}, \ldots, x_{ip}).$$

These expression look much nicer in matrix notation. As before, Y denotes the vector of response values $(y_1, \ldots, y_n)^T$ and X denotes the $n \times (p+1)$ matrix whose first column is 1 and the remaining columns correspond to the explanatory variables. Let β denote the vector $(\beta_0, \ldots, \beta_p)^T$. Let p denote the vector $(p_1, \ldots, p_n)^T$ and let W denote the $n \times n$ diagonal matrix whose ith diagonal element is $p_i(1-p_i)$. Check that

$$\nabla\ell(\beta) = X^T(Y - p)$$

and

$$H\ell(\beta) = -X^T W X.$$

The iterative scheme (2) therefore becomes

$$\beta^{(m+1)} = \beta^{(m)} + (X^T W X)^{-1} X^T (Y - p).$$

This can be rewritten as

$$\beta^{(m+1)} = (X^T W X)^{-1} X^T W Z \tag{3}$$

where

$$Z = X\beta^{(m)} + W^{-1}(Y - p). \tag{4}$$

The method of estimating β therefore proceeds iteratively as follows. First have an initial estimate of $\beta_0, \ldots, \beta_p$. Call this initial estimator $\hat{\beta}^{(0)}$. Use this estimator to calculate p_i via

$$p_i = \frac{\exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \cdots + \hat{\beta}_p^{(0)} x_{ip})}{1 + \exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} \cdots + \hat{\beta}_p^{(0)} x_{ip})}.$$

Use these values of p_i to create the response variable values Z_i via (6) and also use values of p_i to construct the matrix W. With Z and W, we can estimate β via

$$\hat{\beta}^{(1)} = (X^T W X)^{-1} X^T W Z.$$

Now replace the initial estimator $\hat{\beta}^{(0)}$ by $\hat{\beta}^{(1)}$ and repeat this process. Keep repeating this until two successive estimates $\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m+1)}$ do not change much. At that point, stop and report the estimate of β in the logistic regression model by $\hat{\beta}^{(m)}$.

The expression $(X^T W X)^{-1} X^T W Z$ is reminiscent of the usual $(X^T X)^{-1} X^T Y$ which is the usual estimate of β in the linear model. In fact, this is the least squares estimate in a weighted least squares model as we shall describe next.

2 Weighted Least Squares

Consider regression data in the usual set-up. Suppose we think that the right model is:

$$Y = X\beta + e \qquad \text{where } \mathbb{E}e = 0 \text{ and } Cov(e) = \sigma^2 V$$

for some known (positive definite) matrix V. What then is a good estimator of β? The difference from the usual situation is the presence of this matrix V. It turns out the usual least squares estimator is not a good choice here for estimating β. It is better to use the weighted least squares estimator:

$$\hat{\beta}_{wls} := (X^T V^{-1} X)^{-1} X^T V^{-1} Y. \tag{5}$$

It is not too hard to see that this estimator minimizes the weighted sum of squares

$$(Y - X\beta)^T V^{-1} (Y - X\beta)$$

over all β.

Why is it sensible to use (5) for estimating β in this case? The follows reasons motivate this choice:

1. If e is multivariate normal, then (5) is the mle for β.
2. Suppose V is diagonal. Then it is obvious that

$$(Y - X\beta)^T V^{-1} (Y - X\beta) = \sum_{i=1}^{n} \frac{(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2}{v_{ii}}$$

   where v_{ii} denotes the ith diagonal entry of V. It is intuitively clear that minimizing this weighted sum of squraes as opposed to the unweighted sum of squares is the right thing to do here. For example, if v_{ii} is very high, it means that the ith observation is not very trustworthy and it therefore makes sense to give it low weight. In the same way, it makes sense to give large weight to the ith observation if v_{ii} is low.

One can show that $\hat{\beta}_{wls}$ is the BLUE for β. It is also the minimum variance unbiased estimator for β when the errors are multivariate normal.

The expectation and the covariance matrix of $\hat{\beta}_{wls}$ can be easily calculated via:

$$\mathbb{E}\hat{\beta}_{wls} = \beta \quad \text{and} \quad Cov(\hat{\beta}_{wls}) = \sigma^2 (X^T V^{-1} X)^{-1}.$$

3 Iteratively Reweighed Least Squares for Logistic Regression Fitting

Because of the similarity between (3) and (5), Newton's method for computing the maximum likelihood estimator in logistic regression can be seen as a sequence of weighted least squares estimators. That is why the iterative method is also called IRLS (Iteratively Reweighted Least Squares) or IWLS (Iteratively Weighted Least Squares).

Here is a more intuitive approach to understand IRLS. The goal is to fit the model (1) to the data. Because $p_i = \mathbb{E}y_i$, the equation (1) can be rewritten as

$$\log \frac{\mathbb{E}(y_i)}{1 - \mathbb{E}(y_i)} = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}.$$

Because of the form above, a first idea to fit this model to data might be to try to fit a linear model to the response variable $\log(y_i/(1-y_i))$ on the explanatory variables and then to estimate $\beta_0, \ldots, \beta_p$ by the estimated coefficients of that linear model. But because y_i is 0 or 1, the quantity $\log(y_i/(1-y_i))$ is either $-\infty$ or ∞ and so this response variable would make no sense.

A way to fix this is to work with a response variable that is similar in spirit to $\log(y_i/(1-y_i))$ but which actually makes sense. Let $g(x) = \log(x/(1-x))$. By a first order Taylor expansion to g around p_i, we can write

$$g(y_i) \approx g(p_i) + g'(p_i)(y_i - p_i) = \log \frac{p_i}{1-p_i} + \frac{y_i - p_i}{p_i(1-p_i)}$$

The right hand side above makes sense as opposed to $g(y_i)$. So we let

$$Z_i = \log \frac{p_i}{1-p_i} + \frac{y_i - p_i}{p_i(1-p_i)} \tag{6}$$

and we can fit a linear model to Z_i based on the explanatory variables and estimate β by the estimated coefficients in that linear model. Should we estimate the coefficients of that linear model by ordinary least squares or should we use weighted least squares? The variance of Z_i is:

$$var(Z_i) = var\left(\frac{y_i - p_i}{p_i(1-p_i)}\right) = \frac{1}{p_i(1-p_i)}.$$

Therefore if W is a diagonal matrix whose ith diagonal entry is $p_i(1-p_i)$, then

$$Cov(Z) = W^{-1}.$$

Thus, while fitting a linear model to Z_i based on the explanatory variables, it is sensible to estimate the coefficients of the linear model by

$$(X^T W X)^{-1} X^T W Z.$$

This gives us the estimate of β in the logistic regression model:

$$\hat{\beta} := (X^T W X)^{-1} X^T W Z. \tag{7}$$

The obvious problem with the above approach is that we do not know p_i (p_i depends on the parameters $\beta_0, \ldots, \beta_p$ that we are trying to estimate) and so we cannot really compute the response variable Z_i or the matrix W.

The natural solution to this is to use an iterative method. First have an initial estimate of $\beta_0, \ldots, \beta_p$. Call this initial estimator $\hat{\beta}^{(0)}$. Use this estimator to calculate p_i via

$$p_i = \frac{\exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} + \cdots + \hat{\beta}_p^{(0)} x_{ip})}{1 + \exp(\hat{\beta}_0^{(0)} + \hat{\beta}_1^{(0)} x_{i1} \cdots + \hat{\beta}_p^{(0)} x_{ip})}.$$

Use these values of p_i to create the response variable values Z_i via (6) and also use values of p_i to construct the matrix W. With Z and W, we can estimate β as in (7). Call this $\hat{\beta}^{(1)}$:

$$\hat{\beta}^{(1)} = (X^T W X)^{-1} X^T W Z.$$

Now replace the initial estimator $\hat{\beta}^{(0)}$ by $\hat{\beta}^{(1)}$ and repeat this process. Keep repeating this until two successive estimates $\hat{\beta}^{(m)}$ and $\hat{\beta}^{(m+1)}$ do not change much. At that point, stop and report the estimate of β in the logistic regression model by $\hat{\beta}^{(m)}$.

By what we have seen in Section 1, this method is equivalent to computing the MLE by Newton's method.