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

Aditya Guntuboyina

16 April 2015

1 Fitting Logistic Regression to Data

Recall the logistic regression model given by

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

Here $y_1, \ldots, y_n$ are independent Bernoulli random variables with mean p_i. The log-likelihood is given by

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

If $x_i^T := (1, x_{i1}, \ldots, x_{ip})$ denotes the ith row of X, then the log-likelihood becomes

$$\ell(\beta) = \sum_{i=1}^{n} \left[y_i x_i^T \beta - \log\left(1 + \exp(x_i^T \beta)\right)\right]$$

Note that the dependence of this likelihood on the response vector $Y = (y_1, \ldots, y_n)$ is only through $\sum_{i=1}^{n} y_i x_i = X^T Y$. In other words, $X^T Y$ is a sufficient statistic for this model. This is not true if one does not use the canonical logistic link function.

We want to maximize the likelihood so we take derivatives with respect to β. This gives

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

Also the Hessian can be easily computed to be

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

Because we want to maximize the log-likelihood, we set the gradient to zero and this gives

$$X^T Y = X^T p \tag{2}$$

Note here that $p = (p_1, \ldots, p_n)^T$ is an $n \times 1$ vector that is a function of β. In other words, we have to find a vector β for which (2) is satisfied. Because $p = \mathbb{E}Y$, we can rewrite (2) as

$$X^T Y = X^T \mathbb{E}(Y).$$

It may be observed here that in the case of the normal linear regression model, $EY = X\beta$ and this equation is then identical to the normal equations. In the case of the logistic regression model $\mathbb{E}Y$ is not equal to p but a function of p and so we get a more complicated set of equations for solving β. In this way, we can think of the Maximum Likelihood Estimation of β as also a method of moments.

We have seen how to solve (2) in the last class. One basically uses Newton's method which is equivalent to IRLS (Iteratively Reweighted Least Squares).

2 Standard Errors for $\hat{\beta}$

To obtain the standard errors for $\hat{\beta}$, consider the following heuristic argument. Because $\hat{\beta}$ maximizes the log-likelihood, we have $\nabla \ell(\hat{\beta}) = 0$. Suppose the true β is denoted by simply β. By a first order Taylor expansion of $\nabla \ell(\hat{\beta})$ around β, we obtain

$$0 = \nabla \ell(\hat{\beta}) \approx \nabla \ell(\beta) + H\ell(\beta) \left(\hat{\beta} - \beta \right)$$

Using the expressions for $\nabla \ell(\beta)$ and $H\ell(\beta)$, we obtain

$$0 \approx X^T (Y - p) - X^T W X \left(\hat{\beta} - \beta \right).$$

This gives

$$\hat{\beta} - \beta \approx (X^T W X)^{-1} X^T (Y - p)$$

This gives that $\mathbb{E}\hat{\beta} \approx \beta$ which means that $\hat{\beta}$ is approximately unbiased and also that

$$Cov(\hat{\beta}) \approx (X^T W X)^{-1}.$$