Technical Note on Logistic Model

by

Fe Lisondra

In multiple linear regression model,

$$y_i = \beta X_i + \mu_i \text{[a/]}$$

y_i is normally distributed for fixed X_i in which the following assumptions are made:

1. $E(\mu_i) = 0$
2. $E(\mu_{i\,j}) = \sigma^2 \qquad i = j$
   $= 0 \qquad i \neq j$

In actual practice, it is not always reasonable to assume y_i to be normally distributed. Although X_i in some cases may not be normal, i.e., when some of its components are dummy variables, y_i for fixed X_i could be assumed normal and the variance-covariance matrix for y_i given X_i does not depend upon X_i. But in situations where the dependent variable y_i is dichotomous 0 or 1, the ordinary least squares method will yield $E(\mu_i) = 0$ but with $\text{var}(y_i) = \text{Var}(\mu_i) = X_i' \beta (1 - X_i' \beta)$ since y_i is a Bernoulli

---

[a/] y_i denotes the dependent variable for the i^{th} observation. $X_i = (X_{i1}, X_{i2}, X_{i3}, \ldots\ldots, X_{ip})$ is the vector of p independent variables for the i^{th} observation. $\beta = (\beta_1, \beta_2, \ldots\ldots, \beta_p)$ denotes the vector of regression parameters. μ_i denotes the i^{th} uncorrelated disturbance term.

random variable. It is clear that the resulting error variance is not constant for all observations. Thus, regression of y_i on X_i is heteroscedastic and should not be estimated using the ordinary least squares since it violates one of the basic assumptions generally made in a linear model. This heteroscedasticity (unless necessary corrections or transformations are made) will generate inefficient estimators of β. Standard errors of the sample regression coefficients would be therefore incorrect and as a result, tests of significance and confidence intervals for regression coefficients may be seriously misleading. Furthermore, in OLS, estimators of $X_i'\beta$ can have any numerical value despite the fact that $E(y_i) = X_i'\beta$ and $0 \leq y_i \leq 1$, $0 \leq X_i'\beta \leq 1$. This means that y_i being a probability, rules out the linear model because y_i may not be bounded by 0 or 1.

The logistic model provides an appropriate analysis of binary response data. The model in a logistic cumulative distribution function form,

$$P_{y_{i=1}} = F(X_i'\beta) = (1 + \exp(-X_i'\beta))^{-1} \text{ b/}$$

---

b/ This nonlinear function represents the relationship between the probability of attending school $P_{y_{i=1}}$ and the socio-economic and demographic characteristics represented by vector X_i. β represents the vector of regression parameters.

has a curve similar to the cumulative curve of the normal distribution. Its likelihood function is

$$L = \prod_{i=1}^{n} \left[\frac{1}{1 + \exp(-X_i \beta)} \right]^{y_i} \left[1 - \frac{1}{1 + \exp(-X_i \beta)} \right]^{1-y_i}$$

$$= \frac{\{\exp \beta' \sum_{i=1}^{n} X_i y_i\}}{\prod_{i=1}^{n} \left[1 + \exp(X_i' \beta) \right]}$$

where the maximum likelihood estimator[c/] of β is obtained by differentiating the logarithm of the likelihood function, setting the result equal to 0 and solving for β.

This method using the logistic cdf in solving regressions problems with qualitative dependent variable is called logit analysis. The chi-square statistic for testing the hypothesis that a parameter is zero is calculated by computing the square of the parameters estimate divided by its standard error.

[c/] In this paper, the maximum likelihood estimates (MLE) were computed using the Newton-Raphson method.