A Brief Summary on Binary Choice Models for Training Incidence

The incidence of employer-supported training represents the match of the rational decisions of employers and employees. That is, it is the coincidence of the training offering from employers and the acceptance, voluntary or compulsory, on the part of employees. Hence, both firm and individual employee characteristics may explain why such training incidence occurs.

The bivariate statistics in a cross-tabulation may show the relationship between training incidence and any one of firm and employee characteristics. But this may be spurious without an appropriate control over other relevant conditional factors. Hence it is desirable to consider appropriate models which permit inference on partial correlations conditional on a fuller set of information. Given the binary nature of the variable of interest – training participation, natural choices are binary choice models such as probit and logit models.

Several models for binary choice are suitable for analyzing the independent variable y_i , for individual i, that takes on the value of either 0 or 1. Here $i = 1, 2, \ldots, N$. It is also assumed that this choice can be analyzed with reference to a vector of determinants $\boldsymbol{x}_i$, for individual i . The vector $\boldsymbol{x}_i$ will affect the binary choice y_i via a set of parameter $\boldsymbol{\beta}$. We can characterize the probability of taking either choice as

$$\Pr[y_i = 1 \mid \boldsymbol{x}_i] = F(\boldsymbol{x}_i'\boldsymbol{\beta})$$

or

$$\Pr[y_i = 0 \mid \boldsymbol{x}_i] = 1 - F(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

One simplest approach to model this binary choice is to use the linear regression model so that

$$F(\boldsymbol{x}_i'\boldsymbol{\beta}) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

This approach is to assume that

$$y_i = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i$$

or

$$E[y_i \mid \boldsymbol{x}_i] = 1 \cdot \Pr[y_i = 1 \mid \boldsymbol{x}_i] + 0 \cdot \Pr[y_i = 0 \mid \boldsymbol{x}_i] = \Pr[y_i = 1 \mid \mathbf{x}_i] = F(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

However, this approach will generate forecasts that are not limited to [0, 1]. The model of this kind typically suffers from heteroscedasticity. In the event of using this approach, the White generalized heteroscadasticitic standard errors should be used.

Two other alternatives are slightly more involved. Generally, the cumulative distribution function (cdf) $F(\boldsymbol{x}_i'\boldsymbol{\beta})$ should have the following properties:

$$F(-\infty) = 0, F(\infty) = 1, and\ f(x) = \frac{dF(x)}{dx} > 0.$$

That is, F is a nonlinear function of $\boldsymbol{x}_i$,

$$\frac{dF(\boldsymbol{x}_i'\boldsymbol{\beta})}{dx_{ik}} = f(\boldsymbol{x}_i'\boldsymbol{\beta})\beta_k$$

where x_{ik} is the k th element in $\boldsymbol{x}_i$ and β_k is the k th element in $\boldsymbol{\beta}$.

For the probit model, we use the standard normal distribution for the cdf F :

$$F(x) = \Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp(-\frac{1}{2}w^2)dw.$$

Let the latent or unobservable variable be y_i^* which can be described as

$$y_i^* = \boldsymbol{x}_i'\boldsymbol{\beta} + u_i, \quad u_i \overset{iid}{\sim} N(0,\ 1).$$

The observable choice y_i is determined as follows:

$$y_i = 1 \quad if \quad y_i^* > 0,$$
$$y_i = 0 \quad if \quad y_i^* \leq 0.$$

Note that the assumption of the distribution of u_i following the standard normal distribution $N(0,1)$ is not too restrictive. In the event that we have $u_i \overset{iid}{\sim} N(0, \sigma^2)$ we can always rewrite the model for y_i^* as y_i^* / σ so that

$$y_i^* / \sigma = \boldsymbol{x}_i'\boldsymbol{\beta} / \sigma + u_i / \sigma, \quad u_i / \sigma \overset{iid}{\sim} N(0,\ 1)\ .$$

Then we can express $\Pr[y_i = 1 | \boldsymbol{x}_i]$ as

$$\Pr[y_i = 1 | \boldsymbol{x}_i] = \Pr[y_i^* > 0] = \Pr[\boldsymbol{x}_i'\boldsymbol{\beta} + u_i > 0],$$

which can be further written as

$$\Pr[\boldsymbol{x}_i'\boldsymbol{\beta} > -u_i] = \Pr[u_i > -\boldsymbol{x}_i'\boldsymbol{\beta}] = \Pr[u_i < \boldsymbol{x}_i'\boldsymbol{\beta}] = \Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$$

where Φ is the cumulative normal distribution function.

For the logit model, we use the logistic function $F(x)$, which is defined as

$$\Lambda(x) = \frac{1}{1+e^{-x}} = \frac{e^x}{1+e^x}.$$

The first derivative of $\Lambda(x)$ is

$$\frac{\partial\Lambda(x)}{\partial x} = \frac{\partial(\frac{e^x}{1+e^x})}{\partial x} = \Lambda(x)\Lambda(-x) = \frac{e^x}{\left(1+e^x\right)^2}.$$

This implies that $\frac{\partial\Lambda(x)}{\partial x}$ is symmetric around zero, which further implies

$$\Lambda(-x) = 1 - \Lambda(x) \ .$$

The way to incorporate the logistic function into the binary choice model is to identify the logarithm of the odds ratio as a function of $\boldsymbol{x}_i'\boldsymbol{\beta}$. That is, for $P_i = \Pr(y_i = 1 | \boldsymbol{x}_i)$ and

$$1 - P_i = \Pr(y_i = 0 | \boldsymbol{x}_i),$$

$$\ln\left(\frac{P_i}{1-P_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

Then, solving for P_i gives

$$P_i = \frac{e^{\boldsymbol{x}_i'\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i'\boldsymbol{\beta}}} = \frac{1}{1+e^{-\boldsymbol{x}_i'\boldsymbol{\beta}}} = \Lambda(\boldsymbol{x}_i'\boldsymbol{\beta}).$$

Both probit and logit models can be estimated by the method of maximum likelihood. The likelihood function for the binary models is define as

$$L(\boldsymbol{\beta};\ \boldsymbol{y},\ \boldsymbol{X}) = \Pi_{i=1}^{N}\left[F(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{y_i}\left[1-F(\boldsymbol{x}_i'\boldsymbol{\beta})\right]^{(1-y_i)}$$

where $\boldsymbol{y} = [y_1,\ y_2,\ldots,\ y_N]'$ and $\boldsymbol{X} = [\boldsymbol{x}_1,\ \boldsymbol{x}_2,\ldots,\ \boldsymbol{x}_N]'$. The corresponding log-likelihood function is given by

$$\ln L(\boldsymbol{\beta};\ \boldsymbol{y},\ \boldsymbol{X}) = \sum_{i=1}^{N}\left\{y_i \ln F(\boldsymbol{x}_i'\boldsymbol{\beta}) + (1-y_i)\ln\left[1-F(\boldsymbol{x}_i'\boldsymbol{\beta})\right]\right\}.$$

The above likelihood and log-likelihood functions incorporate binary choices and their corresponding probabilities naturally; that is, when $y_i = 1$ [or $y_i = 0$], $F(\boldsymbol{x}_i'\boldsymbol{\beta})$ [or $1-F(\boldsymbol{x}_i'\boldsymbol{\beta})$] enters the likelihood function but $1-F(\boldsymbol{x}_i'\boldsymbol{\beta})$ [or $F(\boldsymbol{x}_i'\boldsymbol{\beta})$] does not. About the log-likelihood function, we can make two remarks: (1) Because $F(\boldsymbol{x}_i'\boldsymbol{\beta})$ and $1-F(\boldsymbol{x}_i'\boldsymbol{\beta})$ are probabilities which are less than 1, the logarithm of which is negative, $\ln L(\boldsymbol{\beta};\ \boldsymbol{y},\ \boldsymbol{X})$ is bounded above by zero. (2) $\ln L(\boldsymbol{\beta};\ \boldsymbol{y},\ \boldsymbol{X})$ is globally concave with respective to β . For the probit model, $F(\boldsymbol{x}_i'\boldsymbol{\beta})$ takes the form of $\Phi(\boldsymbol{x}_i'\boldsymbol{\beta})$. For logit model is used, it is replaced by $\Lambda(\boldsymbol{x}_i'\boldsymbol{\beta})$. The two models are closed to each other in terms of the predicted probabilities and the maximized values of the log-likelihood functions. Intuitively, the real difference is that two models scale $\boldsymbol{x}_i'\boldsymbol{\beta}$ differently: under the logistic function $\text{var}(\boldsymbol{x}_i'\boldsymbol{\beta}) = \frac{\pi}{3}$ but $\text{var}(\boldsymbol{x}_i'\boldsymbol{\beta}) = 1$ under the standard normal distribution. In order to determine what the variables are included in $\boldsymbol{x}_i$, both Wald (W) and the likelihood ratio (LR) tests can be used. [1]

The maximum likelihood estimates of the parameters $\boldsymbol{\beta}$, $\widehat{\boldsymbol{\beta}}$, cannot be interpreted straightforwardly as $\boldsymbol{x}_i$ affects the probability of binary choice y_i in a nonlinear fashion. Hence, using $\widehat{\boldsymbol{\beta}}$ to infer qualitative, rather than marginal, impacts is much common. However, for the logit model, we have

$$\ln\left(\frac{P_i}{1-P_i}\right) = \boldsymbol{x}_i'\boldsymbol{\beta}.$$

This is quite useful for us.

It might be useful to explain the odds ratio and its natural log here. Note that between the odds ratio of .9/.1 = 9 and the odds ratio of .1/.9 = .11, there is asymmetry. However, the log of the odds ratio .9/.1 is ln(9) = 2.217 while the log of the odds ratio .1/.9 is ln(1/9) = -2.217. Not only symmetry is preserved but also the log of the odds ratio of one choice is exactly opposite to the log of the odds ratio of another choice. When the probability of one choice and that of another choice is both .5 (both are equally likely), the odds ratio is .5/.5 = 1 and the log of the odds ratio is ln(1) = 0. The odds ratio of 1 means that no difference between the two choices in terms of probabilities.

The odds ratio can be obtained from the above equation:

[1] A formal test between the two models is the likelihood ratio test using two likelihood function values which has an asymptotic χ^2 distribution with one degree of freedom. More specifically, at the 5% significance level, if twice the difference between two log-likelihood function values is greater than $\chi^2_{0.05}(1) = 3.84$, then one poorly fit model can be rejected.

$$\frac{P_i}{1-P_i} = \exp(\boldsymbol{x}_i{}'\boldsymbol{\beta}).$$

That is, the marginal effects of changes in the explanatory variables on the odds ratio can be measured by $\exp(\boldsymbol{\beta})$. When $\exp(\beta_\kappa)$ is equal to 1, a change in the corresponding *k*th explanatory variable does not affect the odds ratio (e.g., neutral in changing the probability of workplace training participation). When $\exp(\beta_\kappa)$ is greater than 1, a change in the corresponding *k*th explanatory variable increases the odds ratio (e.g., increase the probability of workplace training participation). When $\exp(\beta_\kappa)$ is less than 1, a change in the corresponding *k*th explanatory variable decrease the odds (e.g., decrease the probability of workplace training participation).

Of course, generally it is also possible to compute the marginal probability for both probit and logit models. The marginal impact of the change of the *k*th element of $\boldsymbol{x}_i$ for the probit model is

$$\frac{\partial \Phi(\boldsymbol{x}_i)}{\partial x_{ik}} = \phi(\boldsymbol{x}_i{}'\boldsymbol{\beta})\beta_k$$

where ϕ is the normal density function. The marginal impact of the change of the kth element of $\boldsymbol{x}_i$ for the probit model is

$$\frac{\partial \Lambda(\boldsymbol{x}_i{}'\boldsymbol{\beta})}{\partial x_{ik}} = \Lambda(\boldsymbol{x}_i{}'\boldsymbol{\beta})\Lambda(-\boldsymbol{x}_i{}'\boldsymbol{\beta})\beta_k = \frac{e^{x_i'\boldsymbol{\beta}}}{\left(1+e^{x_i'\boldsymbol{\beta}}\right)^2}\beta_k$$

The terms $\phi(\boldsymbol{x}_i{}'\boldsymbol{\beta})$ and $\Lambda(\boldsymbol{x}_i{}'\boldsymbol{\beta})\Lambda(-\boldsymbol{x}_i{}'\boldsymbol{\beta}) = \frac{e^{x_i'\boldsymbol{\beta}}}{\left(1+e^{x_i'\boldsymbol{\beta}}\right)^2}$ in the above marginal probability equations are called correction factors. There are two approaches to compute the marginal probabilities. In the first approach, the correction factor could be evaluated at the sample means of $\boldsymbol{x}_i$ for all *i*. But the "average" may not be representative of the population. In the second approach, the correct factors are computed for all *i* and then the average of all correction factors is used to compute the marginal probabilities.

This paper will use the logit model in this project because it gives $\boldsymbol{\beta}$ an intuitive interpretation in terms of odds ratios. That is, we use $\exp(\boldsymbol{\beta})$ to measure the marginal effects of changes in the explanatory variables on the odds ratio.

The several statistics for the logit model need to be explained. Since that there is really no R^2 (the coefficient of determination; a measure of goodness of fit for the ordinary least squares regression), econometricians have proposed several alternatives. Based on the log-likelihood function of the model with only a constant term $-2\ln L_{null}$ and that of the model with all predictors $-2\ln L_{full}$, Cox & Snell Pseudo-R^2 is defined as

$$R^2 = 1 - \left[\frac{-2LnL_{null}}{-2LnL_{full}} \right]^{2/N}$$

Because this R^2 value cannot reach 1.0, Nagelkerke has therefore modified it. The correction increases the Cox and Snell version to make 1.0 a possible value for R^2. Nagelkerke Pseudo- R^2 is defined as

$$R^2 = \frac{1 - \left[\frac{-2LnL_{null}}{-2LnL_{full}} \right]^{2/N}}{1 - \left(-2LnL_{null} \right)^{2/N}}$$