Probit and Logit Models

Probit and Logit Models (Binary Outcome Models)

Binary outcome examples

- Consumer economics: whether a consumer makes a purchase or not.
- Labor economics: whether an individual participates in the labor market or not.
- Agricultural economics: whether or not a farmer adopts or uses organic practices, marketing/production contracts, etc.

Binary outcome dependent variable

- The decision/choice is whether or not to have, do, use, or adopt.
- The dependent variable is a binary response
- It takes on two values: 0 and 1.

$$y = \begin{cases} 0 \ if \ no \\ 1 \ if \ yes \end{cases}$$

Binary outcome models

- Binary outcome models are among the most used in applied economics.
- A look at the OLS model: $y = \mathbf{x}'\beta + e$
- Binary outcome models estimate the probability that y=1 as a function of the independent variables.

$$p = \mathrm{pr}[y = 1|\mathbf{x}] = \mathrm{F}(\mathbf{x}'\beta)$$

- There are three different models depending on the functional form of $\mathrm{F}(\mathbf{x}'\beta)$.

Regression model (linear probability model)

- In the linear probability model, $\mathrm{F}(\mathrm{x}'\beta) = \mathrm{x}'\beta$

$$p = \mathrm{pr}[y = 1|\mathrm{x}] = \mathrm{x}'\beta$$

- A problem with the regression model is that the predicted probabilities will not be limited between 0 and 1.
- We do not use the regression model with binary outcome data.

Logit model

- For the logit model, $\mathrm{F}(\mathrm{x}'\beta)$ is the cdf of the logistic distribution.

$$\mathrm{F}(\mathbf{x}'\beta) = \Lambda(\mathbf{x}'\beta) = \frac{e^{\mathbf{x}'\beta}}{1 + e^{\mathbf{x}'\beta}} = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(\mathbf{x}'\beta)}$$

- The predicted probabilities are limited between 0 and 1.

Probit model

- For the probit model, $\mathrm{F}(\mathbf{x}'\beta)$ is the cdf of the standard normal distribution.

$$\mathrm{F}(\mathbf{x}'\beta) = \Phi(\mathbf{x}'\beta) = \int_{-\infty}^{\mathbf{x}'\beta} \phi(z)dz$$

- The predicted probabilities are limited between 0 and 1.

Model coefficients

- Probit and logit models are estimated using the maximum likelihood method.

Interpretation of coefficients

- An increase in x increases/decreases the likelihood that y=1 (makes that outcome more/less likely). In other words, an increase in x makes the outcome of 1 more or less likely.
- We interpret the *sign* of the coefficient but not the *magnitude*. The magnitude cannot be interpreted using the coefficient because different models have different scales of coefficients.

Comparison of coefficients

- Coefficients differ among models because of the functional form of the *F* function.

$$\beta_{logit} \simeq 4\beta_{OLS}$$

$$\beta_{probit} \simeq 2.5\beta_{OLS}$$

$$\beta_{logit} \simeq 1.6\beta_{probit}$$

- We should not compare the magnitude of the coefficients among different models.

Marginal effects

- When estimating probit and logit models, it is common to report the marginal effects after reporting the coefficients.
- The marginal effects reflect the change in the probability of *y*=1 given a 1 unit change in an independent variable x.

Marginal effects for the regression model

- For the OLS regression model, the marginal effects are the coefficients and they do not depend on x.

$$\partial p/\partial \mathrm{x}_j = \beta_j$$

- The index *j* refers to the *j*th independent variable.
- [When we use the index *i*, it refers to the *i*th observation.]

Marginal effects for the binary models (probit and logit)

- For the logit and probit models, the marginal effects are calculated as:

$$\partial p/\partial \mathrm{x}_j = \mathrm{F}'(\mathbf{x}'\beta)\beta_j$$

- The marginal effects depend on x, so we need to estimate the marginal effects at a specific value of x (typically the means).
- Coefficients and marginal effects have the same signs because $\mathrm{F}'(\mathbf{x}'\beta) > 0$.

Marginal effects for the logit model

$$\partial p/\partial \mathrm{x}_j = \Lambda(\mathbf{x}'\beta)[1 - \Lambda(\mathbf{x}'\beta)]\beta_j = \frac{e^{\mathbf{x}'\beta}}{\left(1 + e^{\mathbf{x}'\beta}\right)^2}\beta_j$$

Marginal effects for the probit model

$$\partial p/\partial \mathrm{x}_j = \phi(\mathbf{x}'\beta)\beta_j$$

Estimating marginal effects

Marginal effects at the mean

- The marginal effects are estimated for the average person in the sample $\bar{\mathbf{x}}$.

$$\partial p/\partial \mathrm{x}_j = \mathrm{F}'(\bar{\mathbf{x}}'\beta)\beta_j$$

- Most papers report marginal effects at the mean.
- A problem is that there may not be such a person in the sample.

Average marginal effects

- The marginal effects are estimated as the average of the individual marginal effects.

$$\partial p/\partial \mathrm{x}_j = \frac{\sum \mathrm{F}'(\mathbf{x}'\beta)}{\mathrm{n}}\ \beta_j$$

- This is a better approach of estimating marginal effects, but papers still use the previous approach.
- In practice, the two ways to estimate marginal effects produce almost identical results most of the time.

Partial effects for discrete variables

- Predict the probabilities for the two discrete values of a variable and take the difference:

$$F(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(k+1)) - F(\hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2(k))$$

Interpretation of marginal effects

- An increase in x increases (decreases) the probability that y=1 by the marginal effect expressed as a percent.
    - For dummy independent variables, the marginal effect is expressed in comparison to the base category (x=0).
    - For continuous independent variables, the marginal effect is expressed for a one-unit change in x.
- We interpret both the sign and the magnitude of the marginal effects.
- The probit and logit models produce almost identical marginal effects.

Odds ratio/relative risk for the logit model

- The odds ratio or relative risk is p/(1-p) and measures the probability that y=1 relative to the probability that y=0.

$$p = \frac{\exp(\mathbf{x}'\beta)}{1 + \exp(\mathbf{x}'\beta)}$$

$$\frac{p}{1-p} = \exp(\mathbf{x}'\beta)$$

$$ln\frac{p}{1-p} = \mathbf{x}'\beta$$

- An odds ratio of 2 means that the outcome y=1 is twice as likely as the outcome of y=0.
- Odds ratios are estimated with the logistic model.
- Reporting marginal effects instead of odds ratios is more popular in economics.

Predicted probabilities and goodness of fit measures

- After estimating the models, we can predict the probability that y=1 for each observation.

$$\hat{p} = \mathrm{pr}[y = 1|\mathbf{x}] = \mathrm{F}(\mathbf{x}'\hat{\beta})$$

- For the regression model, the predicted probabilities are *not* limited between 0 and 1.
- For the logit and probit models, the predicted probabilities are limited between 0 and 1.
- The predicted probability indicate the likelihood of y=1. If the predicted probability is greater than 0.5 we can predict that y=1, otherwise y=0.

Goodness of fit measures

Percent correctly predicted values

- If the predicted probability is greater than 0.5 we can predict that y=1, otherwise y=0.
- We can create the following table:

	Actual y=1	Actual y=0
Predicted yhat=1	True	False
Predicted yhat=0	False	True

- We have four cases of 0/1: two of them are correct predictions and two of them are wrong predictions.
- The percent correctly predicted values are the proportion of true predictions to total predictions.

Pseudo R-squared (McFadden R-squared)

- The pseudo R-square is calculated as:

$$\text{R-squared} = 1 - L_{ur}/L_r$$

- It compares the unrestricted log-likelihood L_{ur} for the model we are estimating and the restricted log-likelihood L_r with only an intercept.
- If the independent variables have no explanatory power, the restricted model will be the same as unrestricted model and R-squared will be 0.

Discussion about binary outcome models

Choice between the logit and probit model

- The choice depends on the data generating process, which is unknown.
- The models produce almost identical results (different coefficients but similar marginal effects).
- The choice is up to you.

Coding of the dependent variable

- If we reverse the categories 0 and 1, the signs of the coefficients are reversed (positive become negative and vice versa) but the magnitudes are the same.

Latent variable models

- A latent variable is a variable that is incompletely observed y^*. Latent variables can be introduced into binary outcome models in two ways: index functions and random utility models.

Probit and Logit Model Example

- We study the factors influencing the purchase of health insurance.
- Using data set from the Health and Retirement Study (HRS), wave 5 (2002) collected by the National Institute of Aging.
- Dependent variable: whether or not a person has health insurance (0 or 1).
- Independent variables: retired, age, good health status, household income, education years, married, Hispanic.
- Estimating regression model, logit, and probit models.

Health insurance	y codes	Percent frequency
Yes	1	39%
No	0	61%

Binary outcome model coefficients

Have health insurance	Regression coefficients	Logit coefficients	Probit coefficients
Retired	0.04*	0.19*	0.11*
Age	-0.002	-0.01	-0.008
Good health status	0.06*	0.31*	0.19*
HH income	0.0004*	0.002*	0.001*
Education years	0.02*	0.11*	0.07*
Married	0.12*	0.57*	0.36*
Hispanic	-0.12*	-0.81*	-0.46*
Constant	0.12	-1.71*	-1.06*
R2	0.08	0.07	0.07

* Indicates significance at the 5% level.

- Coefficient interpretation: Retired individuals (in comparison to non-retired individuals), individuals with good health status, higher household income, higher education, married are *more likely* to have health insurance, and Hispanic are *less likely* to have health insurance.
- The regression, logit and probit coefficients differ by a scale factor (and therefore we cannot interpret the magnitude of the coefficients).

Binary outcome model marginal effects

Have health insurance	Regression marginal effects	Logit marginal effects at the mean	Logit average marginal effects	Probit marginal effects at the mean	Probit average marginal effects
Retired	0.04*	0.04*	0.04*	0.04*	0.04*
Age	-0.002	-0.003	-0.003	-0.003	-0.003
Good health status	0.06*	0.07*	0.06*	0.07*	0.06*
HH income	0.0004*	0.0005*	0.0005*	0.0004*	0.0004*
Education years	0.02*	0.02*	0.02*	0.02*	0.02*
Married	0.12*	0.12*	0.12*	0.13*	0.12*
Hispanic	-0.12*	-0.16*	-0.16*	-0.16*	-0.15*

- Marginal effects interpretation: Retired individuals are 4% *more likely* to have insurance (in comparison with those that are not retired). For each additional year in education, individuals are 2% *more likely* to have insurance. Hispanics are 16% *less likely* to have insurance than non-Hispanics.
- Note that unlike the coefficients which are different, the marginal effects are almost identical in the three models.
- The marginal effects at the mean and the average marginal effects are almost identical.
- The signs of the coefficients and marginal effects are the same for the logit and probit models.

- The average of predicted probabilities for having insurance is about 38% which is similar to the actual frequency for having insurance.
- The logit and probit models correctly predict 62% of the values and the rest are misclassified.