BINARY DEPENDENT VARIABLE

Hüseyin Taştan[1]

[1]Yıldız Technical University
Department of Economics

Econometrics I

Binary Dependent Variable: Linear Probability Model

- So far, in all of the models we examined the dependent variable y has been a quantitative variable, e.g., wages, GPA score, prices, etc.
- Can we explain a qualitative (ie binary or dummy) variable using multiple regression?
- Binary dependent variable $y = 1$ or $y = 0$; eg it may indicate whether an adult has a high school education, whether a household owns a house, whether an adult is married, owns a car, etc.
- The case where $y = 1$ is called success whereas $y = 0$ is called failure.
- What happens if we regress a $0/1$ variable on a set of independent variables? How can we interpret regression coefficients?

Binary Dependent Variable: Linear Probability Model

- Under the standard assumptions the conditional expectation of the dependent variable can be written as follows:

$$\mathsf{E}(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

- Since y takes only values of 0 or 1 this conditional expectation can be written as follows:

$$\begin{aligned} \mathsf{E}(y|x) &= P(y = 1|x) \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k \end{aligned}$$

- The probability of success is given by $p(x) = P(y = 1|x)$. The expression above states that the success probability is a linear function of x variables.
- By definition the probability of failure is $P(y = 0|x) = 1 - P(y = 1|x)$

Binary Dependent Variable: Linear Probability Model

- Linear Probability Model (LPM):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + u$$

- x variables can be qualitative or quantitative.
- Slope coefficients are now interpreted as the change in the probability of success:

$$\Delta P(y = 1|x) = \beta_j \Delta x_j$$

- OLS sample regression function is given by

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \ldots + \hat{\beta}_k x_k$$

- $\hat{y}$ is the predicted probability of success.

Example: Women's Labor Force Participation, mroz.gdt

- y ($inlf$ - in the labor force) equals 1 if a married woman reported working for a wage outside the home in 1975, and 0 otherwise..
- Definitions of explanatory variables
- $nwifeinc$: husband's earnings (in \$1000),
- $kidslt6$: number of children less than 6 years old,
- $kidsge6$: number of children between 6-18 years of age,
- $educ, exper, age$
- Model

$$\widehat{inlf} = \hat{\beta}_0 + \hat{\beta}_1 nwifeinc + \hat{\beta}_2 educ + \ldots + \hat{\beta}_7 kidsge6$$

Women's Labor Force Participation, mroz.gdt

Model 1: OLS, using observations 1–753
Dependent variable: inlf

	Coefficient	Std. Error	t-ratio	p-value
const	0.585519	0.154178	3.7977	0.0002
nwifeinc	−0.00340517	0.00144849	−2.3508	0.0190
educ	0.0379953	0.00737602	5.1512	0.0000
exper	0.0394924	0.00567267	6.9619	0.0000
expersq	−0.000596312	0.000184791	−3.2270	0.0013
age	−0.0160908	0.00248468	−6.4760	0.0000
kidslt6	−0.261810	0.0335058	−7.8139	0.0000
kidsge6	0.0130122	0.0131960	0.9861	0.3244

Mean dependent var	0.568393	S.D. dependent var	0.495630
Sum squared resid	135.9197	S.E. of regression	0.427133
R^2	0.264216	Adjusted R^2	0.257303
$F(7, 745)$	38.21795	P-value(F)	6.90e–46

Women's Labor Force Participation, mroz.gdt

- All variables are individually statistically significant except $kidsge6$. All coefficients have expected signs using standard economic theory and intuition.
- Interpretation of coefficient estimates: For example, the coefficient estimate on $educ$, 0.038, implies that, ceteris paribus, an additional year of education increases predicted probability of labor force participation by 0.038.
- The coefficient estimate on $nwifeinc$: if husband's income increases by 10 units (ie, $\$10000$) the probability of labor force participation falls by 0.034.
- $exper$ has a quadratic relationship with $inlf$: the effect of past experience on the probability of labor force participation is diminishing.

Women's Labor Force Participation, mroz.gdt

- The number of young children has a big impact on labor force participation. The coefficient estimate on $kidslt6$ is -0.262.
- Ceteris paribus, having one additional child less than six years old reduces the probability of participation by -0.262.
- In the sample, about 20% of the women have at least one child.

Shortcomings of LPM

- Predicted probability of success is given by $\hat{y}$ and it can have values outside the range 0-1. Obviously, this contradicts the rules of probability.
- In the example out of 753 observations, 16 have $\widehat{inlf} < 0$ and 17 have $\widehat{inlf} > 1$.
- If these are relatively few, they can be interpreted as 0 and 1, respectively.
- Nevertheless, the major shortcoming of LPM is not implausible probability predictions. The major problem is that a probability cannot be linearly related to the independent variables for all their possible values.

Shortcomings of LPM

- In the example, the model predicts that the effect of going from zero children to one young child reduces the probability of working by 0.262.
- This is also the predicted drop if the woman goes from having one child to 2 or 2 to 3, etc.
- It seems more realistic that the first small child would reduce the probability by a large amount, but subsequent children would have a smaller marginal effect.
- Thus, the relationship may be nonlinear.

Shortcomings of LPM

- Despite these shortcomings LPM is useful and often applied in economics.
- It usually works well for values of the independent variables that are near the averages in the sample.
- In the previous example, 96% of the women have either no children or one child under 6. Thus, the coefficient estimate on $kidslt6$ (-0.262) practically measures the impact of the first children on the probability of labor force participation.
- Therefore, we should not use this estimate for changes from 3 to 4 or 4 to 5, etc.

Shortcomings of LPM

- LPM is heteroscedastic: The MLR.5: Constant error variance assumption is not satisfied.
- Recall that y is a binary variable following a Bernoulli distribution. Thus, the variance for a Bernoulli distribution is given by:

$$\text{Var}(u|x) = \text{Var}(y|x) = p(x) \cdot [1 - p(x)]$$

- Since $p(x)$ is a linear combination of x variables, $\text{Var}(u|x)$ is not constant.
- We learned that in this case OLS is unbiased and consistent but inefficient. The Gauss-Markov Theorem fails. Standard errors and the usual inference procedures are not valid.
- It is possible to find more efficient estimators than OLS.

Alternatives of LPM

- There are two widely used binary dependent variable models: **logit** and **probit**
- Both logit and probit model restrict the predicted values between 0 and 1. Logit model uses logistic cumulative density function (cdf) whereas probit uses normal cdf.
- The estimation is usually done by Maximum Likelihood method.
- The coefficient estimates are not directly interpretable. We need to compute marginal effects for each variable in the model after the estimation.
- Logit models can be especially useful for binary classification problems and widely used in machine learning applications.

Logit Regression

- Logit regression can be written as

$$P(Y=1|X_1,X_2,\ldots,X_k) = F(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}}.$$

  where $F(\cdot)$ is the cumulative density function of the logistic density:

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

- The beta coefficients cannot be interpreted as change in success probability.

Example: Mortgage decisions

$$deny = \beta_0 + \beta_1 \times P/I\ ratio + u \quad (1)$$

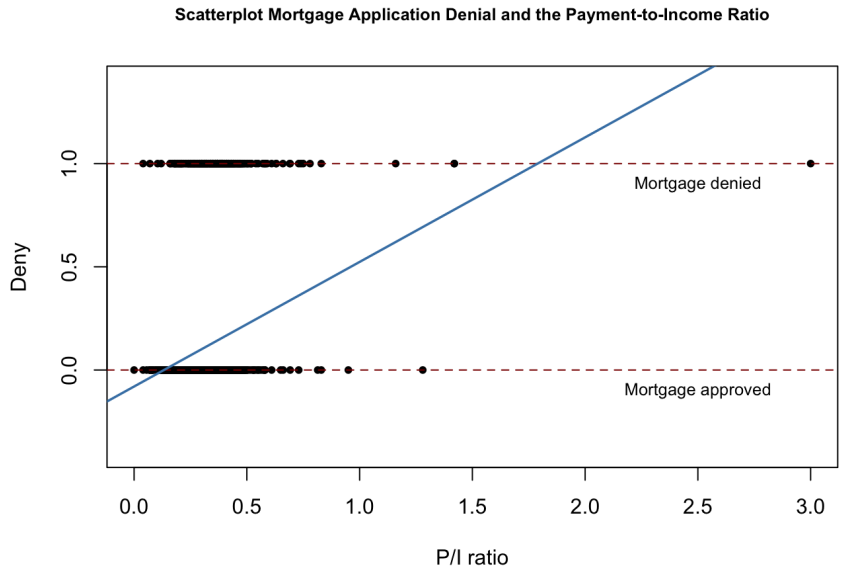
deny: binary variable with NO (=0) mortgage application is accepted, YES (=1) rejected.
P/I ratio: (pirat) the size of the anticipated total monthly loan payments relative to the the applicant's income
Data set: HDMA (from Stock and Watson text)

LPM Results

$$\widehat{deny} = \underset{(0.032)}{-0.080} + \underset{(0.098)}{0.604}(P/I\ ratio) \quad (2)$$

As the P/I ratio increases the probability of mortgage application rejection also increases. A 1 percentage point increase in P/I rate (0.01 point) leads to an increase in the rejection probability by $0.604 \times 0.01 = 0.00604 \approx 0.6\%$

Mortgage LPM model



Example: Mortgage decisions

LPM Results

$$\widehat{deny} = \underset{(0.032)}{-0.080} + \underset{(0.098)}{0.604}(P/I\ ratio) \qquad (3)$$

Add ethnicity dummy:

$$\widehat{deny} = \underset{(0.029)}{-0.091} + \underset{(0.089)}{0.559}(P/I\ ratio) + \underset{(0.025)}{0.177}black \qquad (4)$$

Given P/I ratio is the same (ceteris paribus), being black increases the probability of a mortgage application rejection by about 17.7%.

Example: Mortgage decisions LOGIT Model

Logistic regression Results

$$\widehat{P(deny = 1|P/Iratio, black)} = F(\underset{(0.35)}{-4.13} + \underset{(0.96)}{5.37}P/I\ ratio + \underset{(0.15)}{1.27}black) \qquad (5)$$

Probit results are very similar. Note that all coefficients are positive and statistically significant. Although we can interpret the sign of the coefficients, they do not measure partial effect on success probability.

What is the predicted denial probability for a white applicant with P/I ratio = 0.3?

Answer:

$$\widehat{P(deny = 1|P/Iratio} = 0.3, black = 0) = \frac{1}{1 + e^{-(-4.13+5.37\times0.3+1.27\times0)}}$$

$$= \frac{1}{1 + e^{2.52}} = 0.074$$

or 7.4%.

Example: Mortgage decisions LOGIT Model

Logistic regression Results

$$\widehat{P(deny = 1|P/Iratio, black)} = F(\underset{(0.35)}{-4.13} + \underset{(0.96)}{5.37}P/I\ ratio + \underset{(0.15)}{1.27}black) \qquad (6)$$

Similarly, What is the predicted denial probability for a black applicant with P/I ratio = 0.3?

Answer:

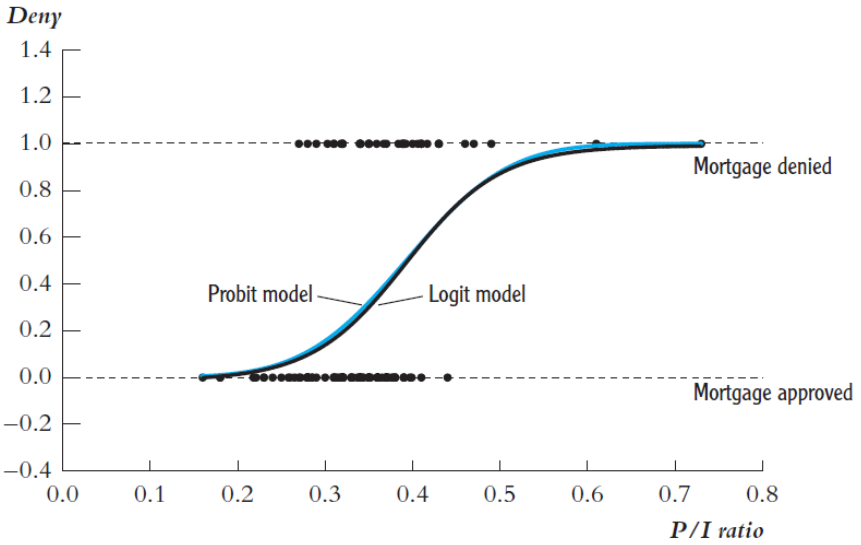
$$\widehat{P(deny = 1|P/Iratio} = 0.3, black = 1) = \frac{1}{1 + e^{-(-4.13+5.37\times0.3+1.27\times1)}}$$

$$= \frac{1}{1 + e^{1.25}} = 0.222$$

or 22.2%.

So, the difference between the two probabilities is 14.8 percentage points.

Mortgage Logit and Probit model



Probit Regression

- In probit regression, the conditional success probability is written as

$$E(Y|X) = P(Y = 1|X) = \Phi(\beta_0 + \beta_1 X). \tag{7}$$

  where $\Phi(\cdot)$ is the cumulative density of the normal distribution. Reminder: $\Phi(z) = P(Z \leq z)\ ,\ Z \sim \mathcal{N}(0, 1)$
- More generally,

$$P(Y = 1|X_1, X_2, \ldots, X_k) = \Phi(\beta_0 + \beta_1 + X_1 + \beta_2 X_2 + \cdots + \beta_k X_k)$$

- Beta coefficients cannot be interpreted as the change in success probability because the relationship is nonlinear. We need to compute marginal effects after estimation.