Applied Economics

Regression with a Binary Dependent Variable

Department of Economics
Universidad Carlos III de Madrid

See Stock and Watson (chapter 11)

Binary Dependent Variables: What is Different?

- So far the dependent variable (Y) has been continuous:
  - district-wide average test score
  - traffic fatality rate
- What if Y is binary?
  - $Y =$ get into college, or not; $X =$ high school grades, SAT scores, demographic variables
  - $Y =$person smokes, or not; $X =$ cigarette tax rate, income, demographic variables
  - $Y =$ mortgage application is accepted, or not; $X =$ race, income, house characteristics, marital status

Example: Mortgage Denial and Race The Boston Fed HMDA Dataset

- Individual applications for single-family mortgages made in 1990 in the greater Boston area
- 2380 observations, collected under Home Mortgage Disclosure Act (HMDA)
- Variables
  - Dependent variable: Is the mortgage denied or accepted?
  - Independent variables: income, wealth, employment status, other loan, property characteristics, and race of applicant.

Linear Probability Model

A natural starting point is the linear regression model with a single regressor:

$Y_i = \beta_0 + \beta_1 X_i + \mu_i$

But:

- What does β_1 mean when Y is binary?
- What does the line $\beta_0 + \beta_1 X$ mean when Y is binary?
- What does the predicted value $\widehat{Y}$ mean when Y is binary? For example, what does $\widehat{Y} = 0.26$ mean?

Linear Probability Model

When Y is binary:

$E(Y|X) = 1xPr(Y = 1|X) + 0xPr(Y = 0|X) = Pr(Y = 1|X)$

Under the assumption, $E(u_i|X_i) = 0$, so

$E(Y_i|X_i) = E(\beta_0 + \beta_1 X_i + u_i|X_i) = \beta_0 + \beta_1 X_i,$

so,

$E(Y|X) = Pr(Y = 1|X) = \beta_0 + \beta_1 X_i$

In the linear probability model, the predicted value of Y is interpreted as the predicted probability that $Y = 1$, and β_1 is the change in that predicted probability for a unit change in X.

Linear Probability Model

- When Y is binary, the linear regression model

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

is called the **linear probability model** because

$$Pr(Y = 1|X) = \beta_0 + \beta_1 X_i$$

- The predicted value is a **probability**:

-$E(Y|X = x) = Pr(Y = 1|X = x) =$ prob. that $Y = 1$ given $X = x$

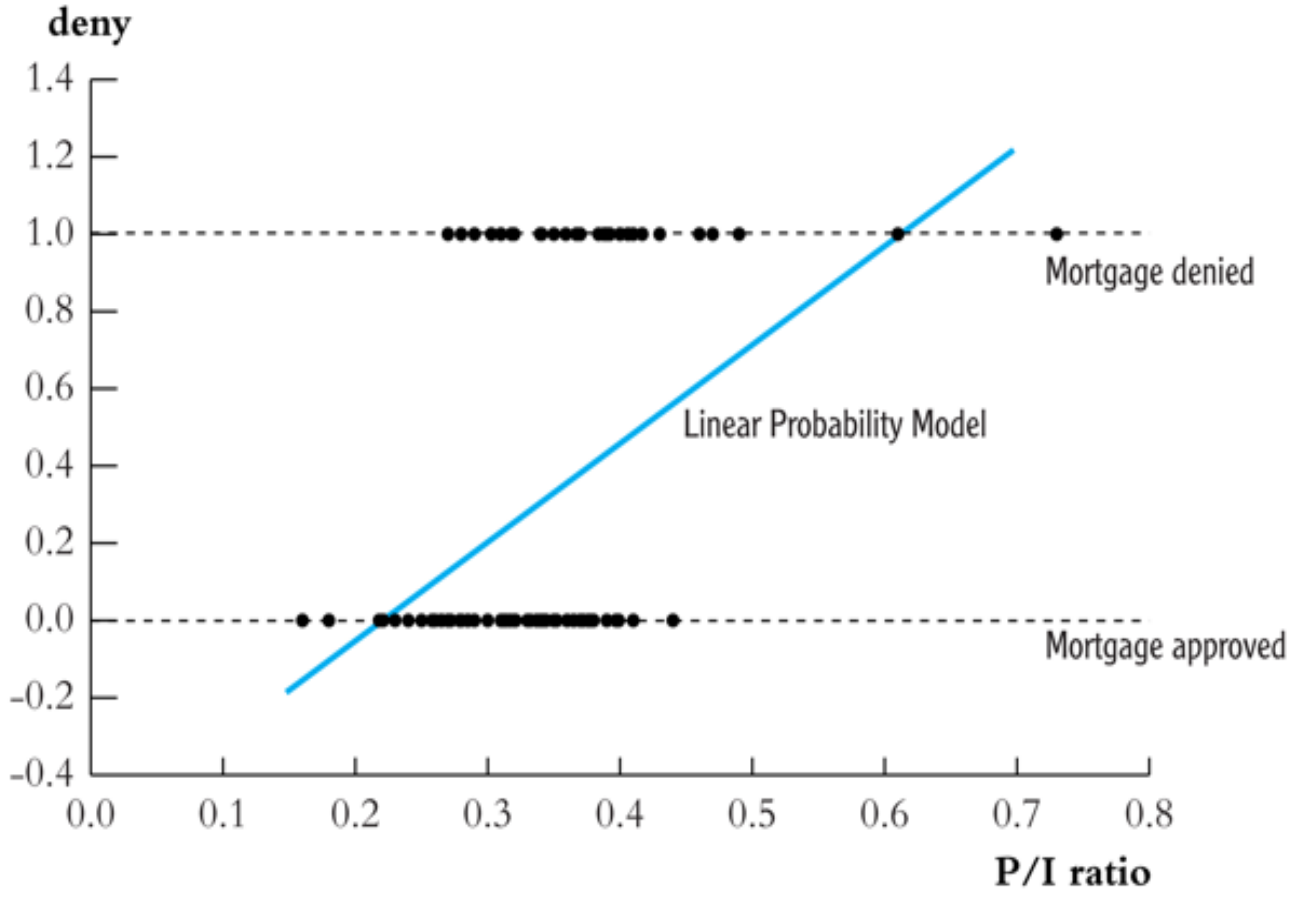
-$\widehat{Y} =$ the **predicted probability** that $Y = 1$ given X

-$\beta_1 =$ change in probability that $Y = 1$ for a unit change in x:

$$\beta_1 = \frac{Pr(Y=1|X=x+\Delta x) - Pr(Y=1|X=x)}{\Delta x}$$

Example: linear probability model, HMDA data

- Mortgage denial v. ratio of debt payments to income (P/I ratio) in a subset of the HMDA data set (n = 127)



Example: linear probability model, HMDA data

$$\widehat{deny_i} = \widehat{\beta}_0 + \widehat{\beta}_1 PI_i + \widehat{\beta}_2 black_i$$

Model 1: OLS, using observations 1–2380
Dependent variable: deny
Heteroskedasticity-robust standard errors, variant HC1

	Coefficient	Std. Error	t-ratio	p-value
const	−0.0905136	0.0285996	−3.1649	0.0016
pi_rat	0.559195	0.0886663	6.3067	0.0000
black	0.177428	0.0249463	7.1124	0.0000

Mean dependent var	0.119748	S.D. dependent var	0.324735
Sum squared resid	231.8047	S.E. of regression	0.312282
R^2	0.076003	Adjusted R^2	0.075226
$F(2, 2377)$	49.38650	P-value(F)	9.67e–22
Log-likelihood	−605.6108	Akaike criterion	1217.222
Schwarz criterion	1234.546	Hannan–Quinn	1223.527

The linear probability model: Summary

-Advantages:

- simple to estimate and to interpret
- inference is the same as for multiple regression (need heteroskedasticity-robust standard errors)

The linear probability model: Summary

- Disadvantages:

-A LPM says that the change in the predicted probability for a given change in X is the same for all values of X, but that doesn't make sense. Think about the HMDA example?

-Also, LPM predicted probabilities can be < 0 or > 1!

-These disadvantages can be solved by using a nonlinear probability model: probit and logit regression

Probit and Logit Regression

- The problem with the linear probability model is that it models the probability of Y=1 as being linear:

$$Pr(Y = 1|X) = \beta_0 + \beta_1 X$$

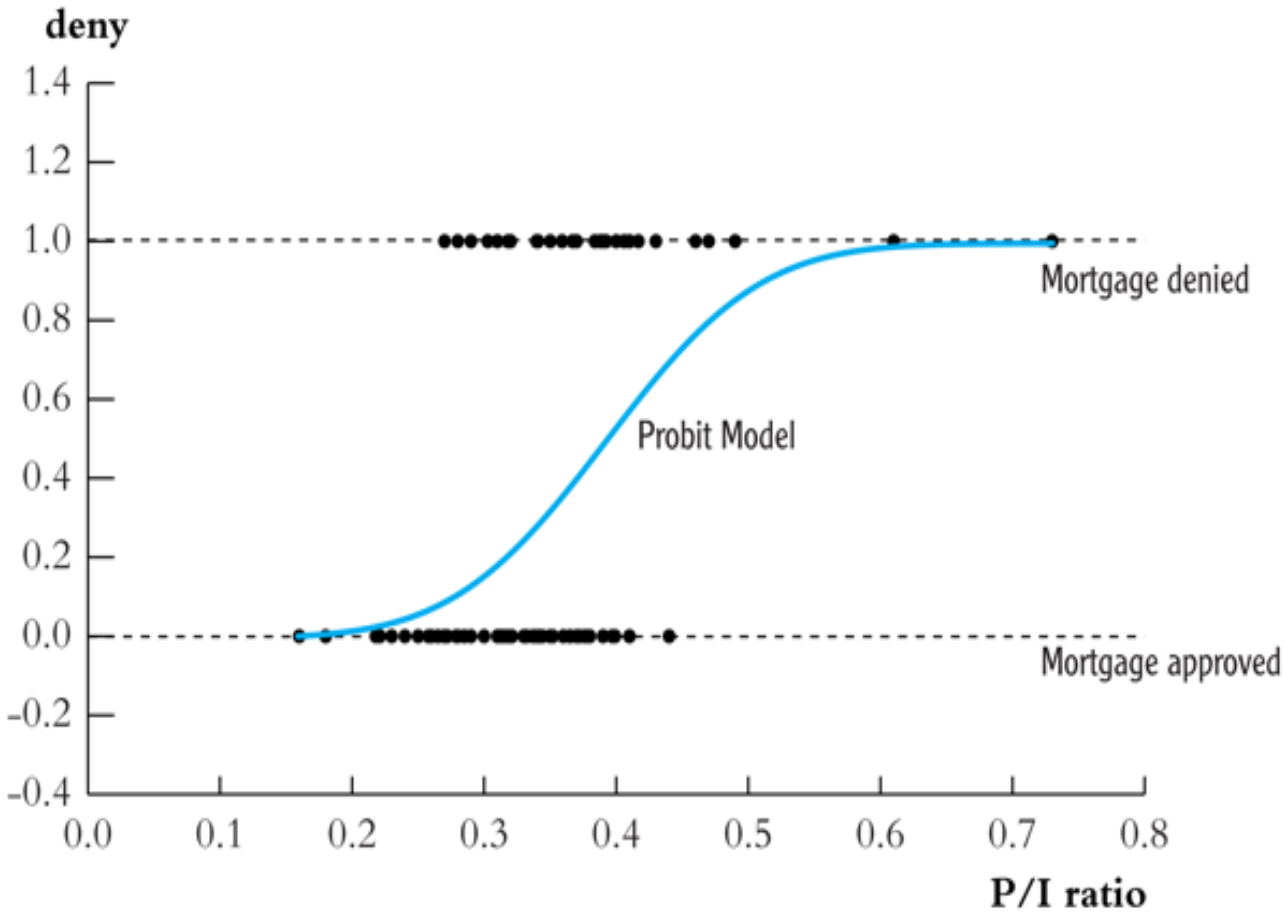
- Instead, we want:

i. $Pr(Y = 1|X)$ to be increasing in X for $\beta_1 > 0$, and

ii. $0 \leq Pr(Y = 1|X) \leq 1$ for all X

- This requires using a nonlinear functional form for the probability. How about an "S-curve"?

S-curve function candidates



Probit regression

- Probit regression models the probability that Y=1 using the cumulative standard normal distribution function, $\Phi(z)$, evaluated at $z = \beta_0 + \beta_1 X$. The probit regression model is,

$Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$

- where $\Phi(.)$ is the cumulative normal distribution function and $z = \beta_0 + \beta_1 X$. is the $z-$value or $z - index$ of the probit model.

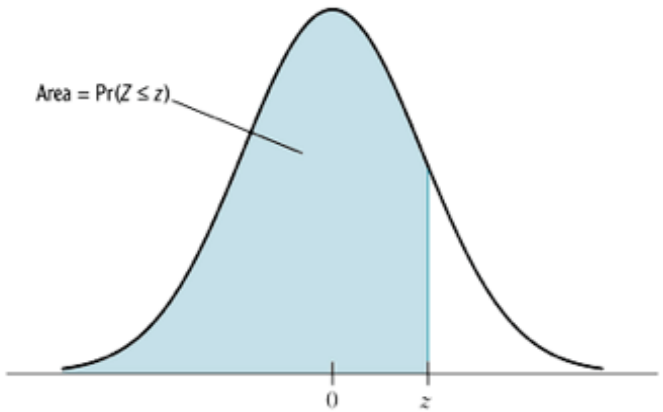
- Example: Suppose $\beta_0 = -2$, $\beta_1 = 3$, $X = .4$, so

$Pr(Y = 1|X = .4) = \Phi(-2 + 3 * .4) = \Phi(-0.8)$

$Pr(Y = 1|X = .4) =$ area under the standard normal density to left of $z = -.8$, which is ...

Probit regression

TABLE 1 The Cumulative Standard Normal Distribution Function, $\Phi(z) = \Pr(Z \leq z)$



	Second Decimal Value of z									
z	0	1	2	3	4	5	6	7	8	9
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121

$$\Pr(z \leq -0.8) = .2119$$

Logit regression

Logit regression models the probability of Y=1, given X, as the cumulative standard logistic distribution function, evaluated at $z = \beta_0 + \beta_1 X$:

$Pr(Y = 1|X) = F(\beta_0 + \beta_1 X)$

where F is the cumulative logistic distribution function:

$F(\beta_0 + \beta_1 X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$

Logit regression

- Example: Suppose $\beta_0 = -3$, $\beta_1 = 2$, $X = .4$, so

$Pr(Y = 1|X = .4) = \frac{1}{1+e^{-(-3+2*0.4)}} = 0.0998$

Why bother with logit if we have probit?

-The main reason is historical: logit is computationally faster & easier, but that does not matter nowadays

-In practice, logit and probit are very similar since empirical results typically do not hinge on the logit/probit choice, both tend to be used in practice

Understanding the Coefficients and the Slopes

In contrast to the linear model, in the probit and logit models the coefficients do not capture the marginal effect on output when a control changes

- if control x_j is continuous, $\frac{\partial Pr(y=1)}{\partial x_j} = f(\beta x)\beta_j$
- if control x_j is discrete, $\Delta Pr(y=1) = F(\beta x_1) - F(\beta x_0)$

- Where $f(.)$ and $F()$ are the density and cumulative density functions

Understanding the Coefficients and the Slopes

- Specifically with $z = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k$

  Logit:

  - $f(z) = \frac{e^{-z}}{(1+e^{-z})^2}$
  - $F(z) = \frac{1}{1+e^{-z}}$

  Probit:

  - $f(z) = \phi(z)$
  - $F(z) = \Phi(z)$

Estimation and Inference in the Logit and Probit Models

We will focus on the probit model:

$$Pr(Y = 1|X) = \Phi(\beta_0 + \beta_1 X)$$

we could use nonlinear least squares. However, a more efficient estimator (smaller variance) is **the Maximum Likelihood Estimator**

The Maximum Likelihood Estimator of the Coefficients in the Probit Model

- The likelihood function is the conditional density of $Y_1, \ldots, Y_n$ given $X_1, \ldots, X_n$, treated as a function of the unknown parameters (β's)
- The maximum likelihood estimator (MLE) is the value of the β's that maximize the likelihood function.
- The MLE is the value of the β's that best describe the full distribution of the data.
- In large samples, the MLE is:
    - consistent
    - normally distributed
    - efficient (has the smallest variance of all estimators)

The Maximum Likelihood Estimator of the Coefficients in the Probit Model

Data: $Y_1, \ldots, Y_n$, i.i.d.

Derivation of the likelihood starts with the density of Y_1:

$Pr(Y_1 = 1|X) = \Phi(\beta_0 + \beta_1 X_1)$ and $Pr(Y_1 = 0) = (1 - \Phi(\beta_0 + \beta_1 X_1))$*so*
$Pr(Y_1 = y_1|X_1) = \Phi(\beta_0 + \beta_1 X_1)^{y_1} * (1 - \Phi(\beta_0 + \beta_1 X_1))^{(1-y_1)} \qquad y_1 = 1, 0$

$Pr(Y_1 = y_1|X_1) = \Phi(z_1)^{y_1} * (1 - \Phi(z_1))^{(1-y_1)}$
with $z_1 = \beta_0 + \beta_1 X_1$

The Maximum Likelihood Estimator. Probit

The probit likelihood function is the joint density of $Y_1, \ldots, Y_n$ given $X_1, \ldots, X_n$, treated as a function of the β's:

$$f(\beta; Y_1, \ldots, Y_n | X_1, \ldots, X_n) = \{\Phi(z_1)^{y_1} * (1 - \Phi(z_1))^{(1-y_1)}\}\{\Phi(z_2)^{y_2} * (1 - \Phi(z_2))^{(1-y_2)}\} \ldots \{\Phi(z_n)^{y_n} * (1 - \Phi(z_n))^{(1-y_n)}\}$$

- $\widehat{\beta}^{MLE}$ maximize this likelihood function.

- But we cannot solve for the maximum explicitly! So the MLE must be maximized using numerical methods

- In large samples:

- $\widehat{\beta}s^{MLE}$, are consistent

- $\widehat{\beta}s^{MLE}$, are normally distributed

- $\widehat{\beta}s^{MLE}$, are asymptotically efficient among all estimators (assuming the probit model is the correct model)

The Maximum Likelihood Estimator. Probit

- Standard errors of $\widehat{\beta} s^{MLE}$ are computed automatically
- Testing, confidence intervals proceeds as usual
- Everything extends to multiple X's

The Maximum Likelihood Estimator. Logit

- The only difference between probit and logit is the functional form used for the probability: Φ is replaced by the cumulative logistic function. Otherwise, the likelihood is similar
- As with probit,
    - $\widehat{\beta}s^{MLE}$, are consistent
    - Their standard errors can be computed
    - Testing confidence intervals proceeds as usual

Measures of Fit for Logit and Probit

- The R^2 and $\bar{R}^2$ do not make sense here (why?). So, two other specialized measures are used:

- The **fraction correctly predicted** = fraction of Y's for which the predicted probability is $> 50\%$ when $Y_i = 1$, or is $< 50\%$ when $Y_i = 0$.

- The pseudo-R^2 measures the improvement in the value of the log likelihood, relative to having no Xs.

Basic Commands in gretl for Probit Estimation

- `probit`: computes Maximum Likelihood probit estimation
- `omit/add`: tests joint significance
- `$yhat`: returns probability estimates
- `$lnl`: returns the log-likelihood for the last estimated model
- `pdf(N,z)`: returns the density of normal distribution
- `cdf(N,z)`: returns the cdf normal distribution
- `logit`: computes Maximum Likelihood logit estimation

probit *depvar indvars* --robust --verbose --p-values

- *depvar* must be binary $\{0, 1\}$ (otherwise a different model is estimated or an error message is given)
- slopes are computed at the means
- by default, standard errors are computed using the negative inverse of the Hessian
- output shows χ^2_q statistic test for null that all slopes are zero
- options:
  1. `--robust`: covariance matrix robust to model misspecification
  2. `--p-values`: shows p-values instead of slope estimates
  3. `--verbose`: shows information from all numerical iterations

Impact of fertility on female labor participation

Using the data set `fertility.gdt` :

-estimate a linear probability model that explains wether or not a woman worked during the last year as a function of the variables `morekids`, `agem1`, `black`, `hispan`, and `othrace`. Give an interpretation of the parameters. - Using the previous model, what is the impact on the probability of working when a woman has more than two children?

- Using the same model and assuming that age of the mother is a continuos variable, what is the impact on the probability of a marginal change in the education of the mother?

- Answer the previous two question using a probit and logit models.