Regression with a Binary Dependent Variable [1]

Jasmine(Yu) Hao

Faculty of Business and Economics
Hong Kong University

September 7, 2021

[1] This section is based on Stock and Watson (2020), Chapter 11.

Example I

- Two people, identical but for their race, walk into a bank and apply for a mortgage, a large loan so that each can buy an identical house. Does the bank treat them the same way? Are they both equally likely to have their mortgage application accepted? By law, they must receive identical treatment. But whether they actually do is a matter of great concern among bank regulators.

Example II

- Loans are made and denied for many legitimate reasons.
  - if the proposed loan payments take up most or all of the applicant's monthly income, a loan officer might justifiably deny the loan.
  - loan officers are human and they can make honest mistakes, so the denial of a single minority applicant does not prove anything about discrimination.
  - Many studies of discrimination thus look for statistical evidence of discrimination, that is, evidence contained in large data sets showing that whites and minorities are treated differently.

Example III

- But how, precisely, should one check for statistical evidence of discrimination in the mortgage market?
  - A start is to compare the fraction of minority and white applicants who were denied a mortgage. In the data examined in this chapter, gathered from mortgage applications in 1990 in the Boston, Massachusetts, area, 28% of black applicants were denied mortgages but only 9% of white applicants were denied.
  - But this comparison does not really answer the question that opened this chapter because the black applicants and the white applicants were not necessarily "identical but for their race." Instead, we need a method for comparing rates of denial, holding other applicant characteristics constant.
- This sounds like a job for multiple regression analysis, whether the applicant is denie is **binary.**
- When the dependent variable is binary, things are more difficult: What does it mean to fit a line to a dependent variable that can take on only two values, 0 and 1?

Binary Dependent I

Whether a mortgage application is accepted or denied is one example of a binary variable. Many other important questions also concern binary outcomes.

- What is the effect of a tuition subsidy on an individual's decision to go to college?
- What determines whether a teenager takes up smoking? What determines whether a country receives foreign aid? What determines whether a job applicant is successful?

In all these examples, the outcome of interest is binary: The student does or does not go to college, the teenager does or does not take up smoking, a country does or does not receive foreign aid, the applicant does or does not get a job.

Binary Dependent Variable I

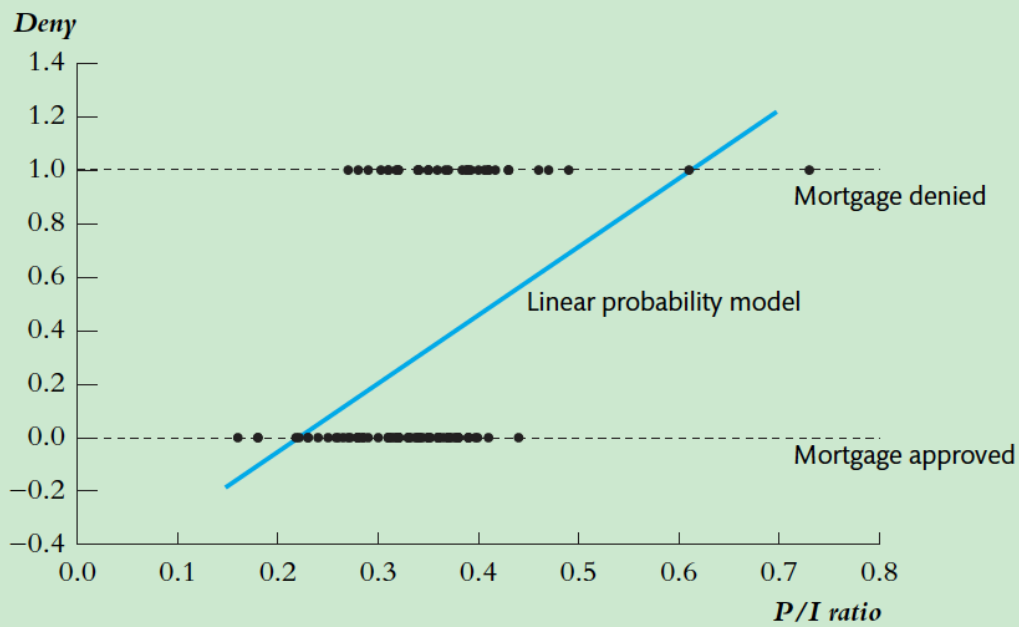
- Whether race is a factor in denying a mortgage application; the binary dependent variable is whether a mortgage application is denied.
  The data are a subset of a larger data set compiled by researchers at the Federal Reserve Bank of Boston under the Home Mortgage Disclosure Act(HMDA) and relate to mortgage applications filed in the Boston, Massachusetts, area in 1990.
- Mortgage applications are complicated. During the period covered by these data, the decision to approve a loan application typically was made by a bank loan officer.
  The loan officer must assess whether the applicant will make his or her loan payments.
  One important piece of information is the size of the required loan payments relative to the applicant's income.

Binary Dependent Variable II

- We therefore begin by looking at the relationship between two variables: the binary dependent variable deny, which equals 1 if the mortgage application was denied and equals 0 if it was accepted, and the continuous variable P/I ratio, which is the ratio of the applicant's anticipated total monthly loan payments to his or her monthly income.

FIGURE 11.1 Scatterplot of Mortgage Application Denial and the Payment-to-Income Ratio

Mortgage applicants with a high ratio of debt payments to income (*P/I ratio*) are more likely to have their application denied (*deny* = 1 if denied; *deny* = 0 if approved). The linear probability model uses a straight line to model the probability of denial, conditional on the *P/I ratio*.



Binary Dependent Variable III

Figure 11.1 presents a scatterplot of deny versus P/I ratio for 127 of the 2380 observations in the data set.
This scatterplot looks different from the scatterplots of Part II because the variable deny is binary.

- Few applicants with a payment-to-income ratio less than 0.3 have their application denied, but most applicants with a payment-to-income ratio exceeding 0.4 are denied.
- This positive relationship between P/I ratio and deny (the higher the P/I ratio, the greater the fraction of denials) is summarized in Figure 11.1 by the OLS regression line estimated using these 127 observations. As usual, this line plots the predicted value of deny as a function of the regressor, the payment-to-income ratio. For example, when P/I ratio $= 0.3$, the predicted value of deny is 0.20. But what, precisely, does it mean for the predicted value of the binary variable deny to be 0.20?

Interpretation

This interpretation follows from two facts.

- First, from Part II, the population regression function is the expected value of Y given the regressors $E[Y|X_1, \ldots, X_k]$.
- Second, if Y is a 0–1 binary variable, its expected value (or mean) is the probability that $Y = 1$; that is, $E[Y] = 0 * Pr(Y = 0) + 1 * Pr(Y = 1) = Pr(Y = 1)$. In the regression context, the expected value is conditional on the value of the regressors, so the probability is conditional on X. Thus for a binary variable,

The Linear Probability Model I

- The linear probability model is the name for the multiple regression model of Part II when the dependent variable is binary rather than continuous.
- Because the dependent variable Y is binary, the population regression function corresponds to the probability that the dependent variable equals 1 given X. The population coefficient b1 on a regressor X is the change in the probability that $Y = 1$ associated with a unit change in X.
- Similarly, the OLS predicted value, $\hat{Y}_i$, computed using the estimated regression function, is the predicted probability that the dependent variable equals 1, and the OLS estimator $\hat{\beta}_1$ estimates the change in the probability that $Y = 1$ associated with a unit change in X.
- The coefficients can be estimated by OLS. Ninety-five percent confidence intervals can be formed as ± 1.96 standard errors, hypotheses concerning several coefficients can be tested using the F-statistic discussed

The Linear Probability Model II

- Because the errors of the linear probability model are always heteroskedastic, **it is essential that heteroskedasticity-robust standard errors be used for inference.**
- One tool that does not carry over is the R^2. When the dependent variable is continuous, it is possible to imagine a situation in which the R^2 equals 1: All the data lie exactly on the regression line. This is impossible when the dependent variable is binary unless the regressors are also binary.

Application to the Boston HMDA data I

The OLS regression of the binary dependent variable, deny, against the payment-to-income ratio, P/I ratio, estimated using all 2380 observations in our data set is

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

- The estimated coefficient on P/I ratio is positive, and the population coefficient is statistically significantly different from 0 at the 1% level.
- Thus applicants with higher debt payments as a fraction of income are more likely to have their application denied. This coefficient can be used to compute the predicted change in the probability of denial given a change in the regressor.
- For example if P/I ratio increases by 0.1, the probability of denial increases by $0.604 * 0.1 \approx= 0.060$ that is, by 6.0 percentage points.

Application to the Boston HMDA data II

- The estimated linear probability model can be used to compute predicted denial probabilities as a function of P/I ratio.
  - For example, if projected debt payments are 30% of an applicant's income, P/I ratio is 0.3, the predicted value is $-0.080 + 0.604 * 0.3 = 0.101$.

What is the effect of race on the probability of denial, holding constant the P/I ratio?

To keep things simple, we focus on differences between black applicants and white applicants.

$$\widehat{deny} = \underset{(0.029)}{-0.091} + \underset{(0.089)}{0.559} P/I\ ratio + \underset{(0.025)}{0.177} black. \quad (11.3)$$

- The coefficient on black, 0.177, indicates that an African American applicant has a 17.7% higher probability of having a mortgage application denied than a white applicant, holding constant their payment-to-income ratio. This coefficient is significant at the 1% level (the t-statistic is 7.11).
- The estimate suggests that there might be racial bias in mortgage decisions, but such a conclusion would be premature.
- Although the payment-toincome ratio plays a role in the loan officer's decision, so do many other factors, such as the applicant's earning potential and his or her credit history. If any of these variables is correlated with the regressors black given the P/I ratio, its omission will cause omitted variable bias.
- Thus we must defer any conclusions about discrimination in

Shortcomings of the linear probability model I

- The linearity that makes the linear probability model easy to use is also its major flaw. Because probabilities cannot exceed 1, the effect on the probability that $Y = 1$ of a given change in X must be nonlinear:
- Although a change in P/I ratio from 0.3 to 0.4 might have a large effect on the probability of denial, once P/I ratio is so large that the loan is very likely to be denied, increasing P/I ratio further will have little effect.
- In contrast, in the linear probability model, the effect of a given change in P/I ratio is constant, which leads to predicted probabilities to drop below 0 for very low values of P/I ratio and exceed 1 for high values!
- A probability cannot be less than 0 or greater than 1. This nonsensical feature is an inevitable consequence of the linear regression.

Probit and Logit Regression I

- Probit and logit regression are nonlinear regression models specifically designed for binary dependent variables.
- A regression with a binary dependent variable Y models the probability that $Y = 1$, it makes sense to adopt a nonlinear formulation that forces the predicted values to be between 0 and 1.
- Because cumulative probability distribution functions (c.d.f.'s) produce probabilities between 0 and 1, they are used in logit and probit regressions.
  - Probit regression uses the standard normal c.d.f.
  - Logit regression, also called logistic regression, uses the logistic c.d.f.

Probit regression with a single regressor I

$Pr(Y = 1|X) = \Phi(\beta_0 + X\beta_1)$,
where Φ is the normal C.D.F.
For example, suppose that Y is the binary mortgage denial variable (deny), X is the payment-to-income ratio (P/I ratio), $\beta_0 = -2$, and $\beta_1 = 3$.
What then is the probability of denial if P/I *ratio* $= 0.4$?

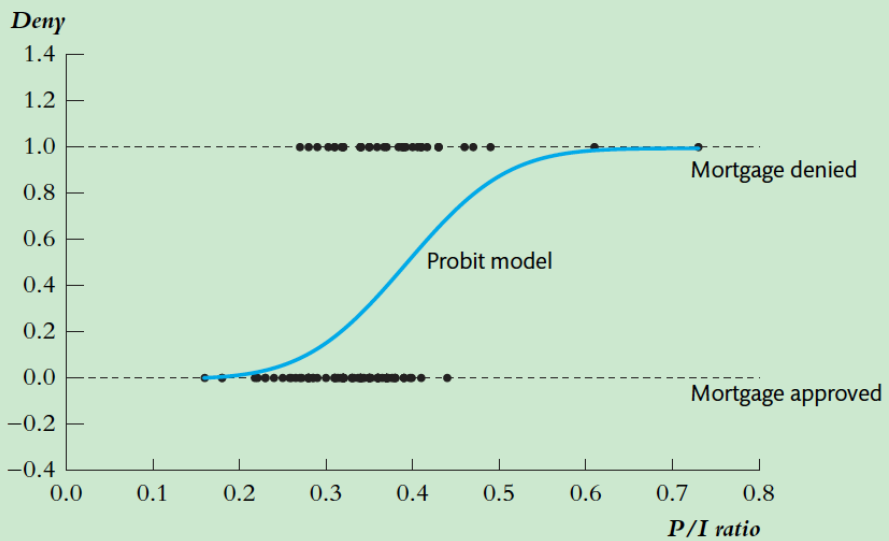
- This probability is $\Phi(-0.8) = 0.212$.
  - when P/I ratio is 0.4, the predicted probability that the application will be denied is 21.2%,
- In the probit model, the term $\beta_0 + \beta_1 X$ plays the role of "z" in the cumulative standard normal distribution table in Appendix Table 1.
- If β_1 is positive, a greater value for X increases the z-value and thus increases the probability that Y $= 1$; if β_1 is negative, a greater value for X decreases the probability that $Y = 1$.

Probit regression with a single regressor II

- Although the effect of X on the z-value is linear, its effect on the probability is nonlinear. Thus in practice the easiest way to interpret the coefficients of a probit model is to compute the predicted probability, or the change in the predicted probability, for one or more values of the regressors.

FIGURE 11.2 Probit Model of the Probability of Denial Given *P/I Ratio*

The probit model uses the cumulative normal distribution function to model the probability of denial given the payment-to-income ratio or, more generally, to model $Pr(Y = 1|X)$. Unlike the linear probability model, the probit conditional probabilities are always between 0 and 1.



Probit regression with multiple regressors

- In all the regression problems we have studied so far, leaving out a determinant of Y that is correlated with the included regressors results in omitted variable bias.
- The solution is to include the additional variable as a regressor.
- The probit model with multiple regressors extends the single-regressor probit model by adding regressors to compute the z-value. Accordingly, the probit population regression model with two regressors, X_1 and X_2, is

Probit regression with multiple regressors II

$$Pr(Y = 1|X_1, X_2) = \Phi(\beta_0 + \beta_1 X_1 + \beta_2 X_2).$$

1. For example, suppose that $\beta_0 = -1.6, \beta_1 = 2, \beta_2 = 0.5$. If $X_1 = 0.4$ and $X_2 = 1$, the z-value is $z = -1.6 + 2 * 0.4 + 0.5 * 1 = -0.3$. So the probability that $Y = 1$ given $X_1 = 0.4, X_2 = 1$ is $\Phi(-0.3) = 38\%$.
2. Effect of a change in X. In general, the regression model can be used to determine the expected change in Y arising from a change in X.

Probit regression with multiple regressors III

When Y is binary, its conditional expectation is the conditional probability that it equals 1, so the expected change in Y arising from a change in X is the change in the probability that $Y = 1$.

1. First, compute the predicted value at the original value of X using the estimated regression function;
2. next, compute the predicted value at the changed value of X, $X + \Delta X$;
3. finally, compute the difference between the two predicted values.

Application to the mortgage data I

As an illustration, we fit a probit model to the 2380 observations in our data set on mortgage denial (deny) and the payment-toincome ratio (P/I *ratio*):

$$\widehat{Pr(deny|P/I \text{ ratio})} = \Phi(\underset{(0.16)}{-2.19} + \underset{(0.47)}{2.97}\, P/I \text{ ratio}). \ (SW \ 11.7)$$

- The estimated coefficients of -2.19 and 2.97 are difficult to interpret because they
affect the probability of denial via the z-value. Indeed, the only things that can ber eadily concluded from the estimated probit regression are that the payment-to-income ratio is positively related to probability of denial (the coefficient on P/I ratio is positive) and that this relationship is statistically significant ($t = 2.97/0.47 = 6.32$).

Application to the mortgage data II

What is the change in the predicted probability that an application will be denied when the payment-to-income ratio increases from 0.3 to 0.4?

To answer this question, we follow the procedure:

- Compute the probability of denial for P/I ratio = 0.3 and for P/I ratio = 0.4, and then compute the difference.
- The probability of denial when P/I *ratio* $= 0.3$ is 0.097.
- The probability of denial when P/I *ratio* $= 0.4$ is 0.159.
- The estimated change in the probability of denial is $0.159 - 0.097 = 0.062$.
- That is, an increase in the payment-to-income ratio from 0.3 to 0.4 is associated with an increase in the probability of denial of 6.2 percentage points, from 9.7% to 15.9%.
- Because the probit regression function is nonlinear, the effect of a change in X depends on the starting value of X. For example, if P/I *ratio* $= 0.5$ the estimated denial probability is $\Phi(-2.19 + 2.97 * 0.5) = \Phi(-0.71) = 0.239$.

Application to the mortgage data III

- Thus the change in the predicted probability when P/I ratio increases from 0.4 to 0.5 is 0.239 - 0.159, or 8.0 percentage points, larger than the increase of 6.2 percentage points when P/I ratio increases from 0.3 to 0.4.

What is the effect of race on the probability of mortgage denial, holding constant the P/I ratio?
To estimate this effect, we estimate a probit regression with P/I ratio and black as regressors:

$$\widehat{Pr(deny = 1|P/I\ ratio, black)} = \Phi(\underset{(0.16)}{-2.26} + \underset{(0.44)}{2.74}\, P/I\ ratio + \underset{(0.083)}{0.71}\ black),$$

- Again, the values of the coefficients are difficult to interpret, but the sign and statistical significance are not.
- The coefficient on black is positive, indicating that an African American applicant has a higher probability of denial than a white applicant, holding constant their payment-to-income ratio.

Application to the mortgage data IV

- This coefficient is statistically significant at the 1% level (the t-statistic on the coefficient multiplying black is 8.55).
- For a white applicant with P/I ratio = 0.3, the predicted denial probability is 7.5%,
- while for a black applicant with P/I ratio = 0.3, it is 23.3%;
- the difference in denial probabilities between these two hypothetical applicants is 15.8 percentage points.

Estimation of the probit coefficients I

- The probit coefficients reported here were estimated using the method of maximum likelihood, which produces efficient (minimum variance) estimators in a wide variety of applications, including regression with a binary dependent variable.
- The maximum likelihood estimator is consistent and normally distributed in large samples, so t-statistics and confidence intervals for the coefficients can be constructed in the usual way.
- Regression software for estimating probit models typically uses maximum likelihood estimation, so this is a simple method to apply in practice.
- Standard errors produced by such software can be used in the same way as the standard errors of regression coefficients; for
  - example, a 95% confidence interval for the true probit coefficient can be constructed as the estimated coefficient $\pm$ 1.96 standard errors.
  - Similarly, F-statistics computed using maximum likelihood estimators can be used to test joint hypotheses.

Logit Regression I

The population logit model of the binary dependent variable Y with multiple regressors is

$$Pr(Y = 1|X_1, \ldots, X_k) = F(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)$$
$$= \frac{1}{1 + \exp -(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)}.$$

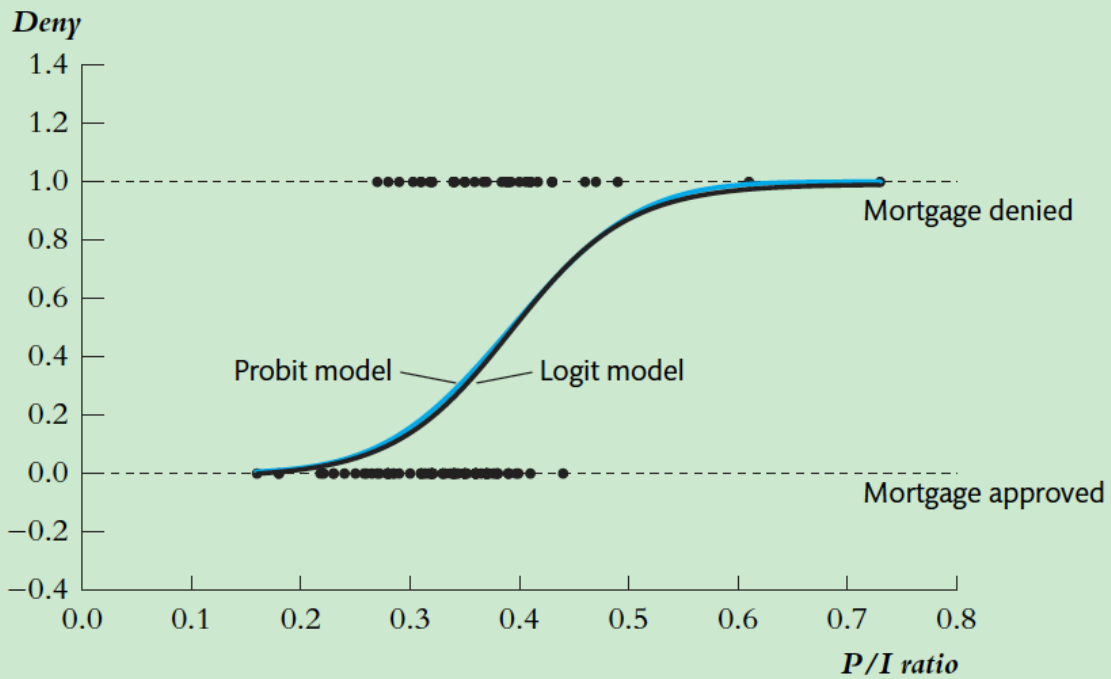
Logit regression is similar to probit regression except that the cumulative distribution function is different.

- As with probit, the logit coefficients are best interpreted by computing predicted probabilities and differences in predicted probabilities.
- The coefficients of the logit model can be estimated by maximum likelihood.
- The maximum likelihood estimator is consistent and normally distributed in large samples, so t-statistics and confidence intervals for the coefficients can be constructed in the usual way.
- The logit and probit regression functions are similar.

Logit Regression II

FIGURE 11.3 Probit and Logit Models of the Probability of Denial Given *P/I Ratio*

These logit and probit models produce nearly identical estimates of the probability that a mortgage application will be denied, given the payment-to-income ratio.



Historically, the main motivation for logit regression was that the logistic cumulative distribution function could be computed faster than the normal cumulative distribution function. With the advent of more powerful computers, this distinction is no longer important.

Application to the Boston HMDA data

A logit regression of deny against P/I ratio and black, using the 2380 observations in the data set, yields the estimated regression function

$$\begin{aligned} &Pr(deny = 1|P/I\ ratio, black) \\ &\qquad = F(\underset{(0.35)}{-4.13} + \underset{(0.96)}{5.37}\, P/I\ ratio + \underset{(0.15)}{1.27}\, black), SW(11.10) \end{aligned}$$

- The coefficient on black is positive and statistically significant at the 1% level (the t-statistic is 8.47).
- The predicted denial probability of a white applicant with P/I ratio = 0.3 is 0.074, or 7.4%.
- The predicted denial probability of an African American applicant with P/I ratio = 0.3 is 0.222, or 22.2%, so the difference between the two probabilities is 14.8 percentage points.

Comparing the Linear Probability, Probit, and Logit Model I

All three models—linear probability, probit, and logit—are just approximations to the unknown population regression function $E[Y|X] = Pr(Y = 1|X)$. The linear probability model is easiest to use and to interpret, but it cannot capture the nonlinear nature of the true population regression function.
Probit and logit regressions model this nonlinearity in the probabilities, but their regression coefficients are mored ifficult to interpret.
So which should you use in practice?

- There is no one right answer, and different researchers use different models. Probit and logit regressions frequently produce similar results.

Comparing the Linear Probability, Probit, and Logit Model II

- For example, according to the estimated probit model the difference in denial probabilities between a black applicant and a white applicant with P/I *ratio* $= 0.3$ was estimated to be 15.8 percentage points, whereas the logit estimate of this gap was 14.9 percentage points.
- For practical purposes, the two estimates are very similar. One way to choose between logit and probit is to pick the method that is easier to use in your statistical software.
- The linear probability model provides the least sensible approximation to the nonlinear population regression function. Even so, in some data sets there may be few extreme values of the regressors, in which case the linear probability model still can provide an adequate approximation.

Comparing the Linear Probability, Probit, and Logit Model III

- In the denial probability regression in Equation (11.3), the estimated black/white gap from the linear probability model is 17.7 percentage points, larger than the probit and logit estimates but still qualitatively similar.
- The only way to know this, however, is to estimate both a linear and a nonlinear model and to compare their predicted probabilities.

Estimation

- The nonlinear models studied are nonlinear functions of the independent variables but are linear functions of the unknown coefficients(parameters).
- Consequently, the unknown coefficients of those nonlinear regression functions can be estimated by OLS.
- In contrast, the probit and logitregression functions are nonlinear functions of the coefficients.
- That is, the probit coefficients appear inside the cumulative distribution functions.
- Because the population regression function is a nonlinear function of the coefficients, those coefficients cannot be estimated by OLS.

Nonlinear Least Squares Estimation I

- Nonlinear least squares is a general method for estimating the unknown parameters of a regression function when, like the probit coefficients, those parameters enter the population regression function nonlinearly.
- The nonlinear least squares estimator extends the OLS estimator to regression functions that are nonlinear functions of the parameters.
- Like OLS, nonlinear least squares finds the values of the parameters that minimize the sum of squared prediction mistakes produced by the model.

Nonlinear Least Squares Estimation II

Nonlinear least squares estimator of the parameters of the probit model. The conditional expectation of Y given the X's is $E[Y|X_1, \ldots, X_k] = Pr(Y = 1|X_1, \ldots, X_k)$.
Estimation by nonlinear least squares fits this conditional expectation function, which is a nonlinear function of the parameters, to the dependent variable.
That is, the nonlinear least squares estimator of the probit coefficients is the values of $\beta_1, \ldots, \beta_k$ that minimize the sum of squared prediction mistakes: $\sum_{i=1}^{n}[Y_i - \Phi(\beta_0 + \beta_1 X_1 + \ldots + \beta_k X_k)]^2$.

Nonlinear Least Squares Estimation III

The nonlinear least squares estimator shares two key properties with the OLS estimator in linear regression:

- It is consistent (the probability that it is close to the true value approaches 1 as the sample size gets large),
- normally distributed in large samples.
- There are, however, estimators that have a smaller variance than the nonlinear least squares estimator;
- the nonlinear least squares estimator is **inefficient**.
- For this reason, the nonlinear least squares estimator of the probit coefficients is rarely used in practice, and instead the parameters are estimated by **maximum likelihood**.

Maximum Likelihood Estimation I

- The likelihood function is the joint probability distribution of the data, treated as a function of the unknown coefficients. The maximum likelihood estimator (MLE) of the unknown coefficients consists of the values of the coefficients that maximize the likelihood function. Because the MLE chooses the unknown coefficients to maximize the likelihood function, which is in turn the joint probability distribution, in effect the MLE chooses the values of the parameters to maximize the probability of drawing the data that are actually observed. In this sense, the MLEs are the parameter values "most likely" to have produced the data.
- To illustrate maximum likelihood estimation, consider two i.i.d. observations, Y_1 and Y_2, on a binary dependent variable with no regressors. Thus Y is a Bernoulli random variable, and the only unknown parameter to estimate is the probability p that $Y = 1$, which is also the mean of Y.

Maximum Likelihood Estimation II

- To obtain the maximum likelihood estimator, we need an expression for the likelihood function, which in turn requires an expression for the joint probability distribution of the data. The joint probability distribution of the two observations Y_1 and Y_2 is $Pr(Y_1 = y_1, Y_2 = y_2)$.
  Because Y_1 and Y_2 are independently distributed, the joint distribution is the product of the individual distributions, so $Pr(Y_1 = y_1, Y_2 = y_2) = Pr(Y_1 = y_1)Pr(Y_2 = y_2)$.
  The Bernoulli distribution can be summarized in the formula $Pr(Y = y) = p^{y1}(1 - p)^{1-y}$.
  Thus the joint probability distribution of Y_1 and Y_2 is $Pr(Y_1 = y_1, Y_2 = y_2) = p^{y_1+y_2}(1 - p)^{2-y_1-y_2}$.
- The likelihood function is the joint probability distribution, treated as a function of the unknown coefficients. For $n = 2$ i.i.d. observations on Bernoulli random variables,
  the likelihood function is
  $f(p; Y_1, Y_2) = p^{Y_1+Y_2}(1 - p)^{2-Y_1-Y_2}$.(11.12)

Maximum Likelihood Estimation III

- The maximum likelihood estimator of p is the value of p that maximizes the likelihood function in Equation (11.12). As with all maximization or minimization problems, this can be done by trial and error; that is, you can try different values of p and compute the likelihood $f(p; Y_1, Y_2)$ until you are satisfied that you have maximized this function. In this example, however, maximizing the likelihood function using calculus produces a simple formula for the MLE:
  The MLE is $\hat{p} = \frac{1}{2}(Y_1 + Y_2)$.
- In other words, the MLE of p is just the sample average! In fact, for general n, the MLE pn of the Bernoulli probability p is the sample average; that is, $\hat{p} = \bar{Y}$.

Statistical inference based on the MLE I

- Because the MLE is normally distributed in large samples, statistical inference about the probit and logit coefficients based on the MLE proceeds in the same way as inference about the linear regression function coefficients based on the OLS estimator. That is, hypothesis tests are performed using the t-statistic, and 95% confidence intervals are formed as $\pm$ 1.96 standard errors.
  Tests of joint hypotheses on multiple coefficients use the F-statistic in a way similar to that discussed in Chapter 7 for the linear regression model.
- All of this is completely analogous to statistical inference in the linear regression model.
  An important practical point is that some statistical software reports tests of joint
  hypotheses using the F-statistic, while other software uses the chi-squared statistic. The chi-squared statistic is $q * F$, where q is the number of restrictions being tested.

Statistical inference based on the MLE II

Because the F-statistic is, under the null hypothesis, distributed as χ^2_q in large samples, $q * F$ is distributed as χ^2_q in large samples.

- Because the two a pproaches differ only in whether they divide by q, they produce identical inferences, but you need to know which approach is implemented in your software so that you use the correct critical values.

Measures of Fit I

- It was mentioned that the R^2 is a poor measure of fit for the linear probability model. This is also true for probit and logit regression.
- Two measures of fit for models with binary dependent variables are the fraction correctly predicted and the pseudo-R2. The fraction correctly predicted uses the following rule: If $Y_i = 1$ and the predicted probability exceeds 50% or if $Y_i = 0$ and the predicted probability is less than 50%, then Yi is said to be correctly predicted.
- Otherwise, Y_i is said to be incorrectly predicted. The fraction correctly predicted is the fraction of the n observations $Y_1, \ldots, Y_n$ that are correctly predicted.
- An advantage of this measure of fit is that it is easy to understand. A disadvantage is that it does not reflect the quality of the prediction:

Measures of Fit II

- If $Y_i = 1$, the observation is treated as correctly predicted whether the predicted probability is 51% or 90%. The pseudo-R^2 measures the fit of the model using the likelihood function.
- Because the MLE maximizes the likelihood function, adding another regressor to a probit or logit model increases the value of the maximized likelihood, just like adding a regressor necessarily reduces the sum of squared residuals in linear regression by OLS.
- This suggests measuring the quality of fit of a probit model by comparing values of the maximized likelihood function with all the regressors to the value of the likelihood with none. This is, in fact, what the pseudo-R^2 does.

Application to the Boston HMDA Data I

- Mortgage denial rates were higher for black than white applicants, holding constant their **payment-to-income ratio**.
- If any of those other factors differ systematically by race, the estimators considered so far have **omitted variable bias**.
- Statistical evidence of discrimination in the Boston HMDA data.
- **Objective**: estimate the effect of race on the probability of denial, holding constant other characteristics.

Summary statistics

TABLE 11.1 Variables Included in Regression Models of Mortgage Decisions

Variable	Definition	Sample Average
Financial Variables		
P/I ratio	Ratio of total monthly debt payments to total monthly income	0.331
housing expense-to-income ratio	Ratio of monthly housing expenses to total monthly income	0.255
loan-to-value ratio	Ratio of size of loan to assessed value of property	0.738
consumer credit score	1 if no "slow" payments or delinquencies 2 if one or two slow payments or delinquencies 3 if more than two slow payments 4 if insufficient credit history for determination 5 if delinquent credit history with payments 60 days overdue 6 if delinquent credit history with payments 90 days overdue	2.1
mortgage credit score	1 if no late mortgage payments 2 if no mortgage payment history 3 if one or two late mortgage payments 4 if more than two late mortgage payments	1.7
public bad credit record	1 if any public record of credit problems (bankruptcy, charge-offs, collection actions) 0 otherwise	0.074
Additional Applicant Characteristics		
denied mortgage insurance	1 if applicant applied for mortgage insurance and was denied, 0 otherwise	0.020
self-employed	1 if self-employed, 0 otherwise	0.116
single	1 if applicant reported being single, 0 otherwise	0.393
high school diploma	1 if applicant graduated from high school, 0 otherwise	0.984
unemployment rate	1989 Massachusetts unemployment rate in the applicant's industry	3.8
condominium	1 if unit is a condominium, 0 otherwise	0.288
black	1 if applicant is black, 0 if white	0.142
deny	1 if mortgage application denied, 0 otherwise	0.120

TABLE 11.2 Mortgage Denial Regressions Using the Boston HMDA Data

Dependent variable: *deny* = 1 if mortgage application is denied, = 0 if accepted; 2380 observations.

Regression Model	*LPM*	*Logit*	*Probit*	*Probit*	*Probit*	*Probit*
Regressor	(1)	(2)	(3)	(4)	(5)	(6)
black	0.084**	0.688**	0.389**	0.371**	0.363**	0.246
	(0.023)	(0.182)	(0.098)	(0.099)	(0.100)	(0.448)
P/I ratio	0.449**	4.76**	2.44**	2.46**	2.62**	2.57**
	(0.114)	(1.33)	(0.61)	(0.60)	(0.61)	(0.66)
housing expense-to-income ratio	−0.048	−0.11	−0.18	−0.30	−0.50	−0.54
	(0.110)	(1.29)	(0.68)	(0.68)	(0.70)	(0.74)
medium loan-to-value ratio (0.80 ≤ *loan-value ratio* ≤ 0.95)	0.031*	0.46**	0.21**	0.22**	0.22**	0.22**
	(0.013)	(0.16)	(0.08)	(0.08)	(0.08)	(0.08)
high loan-to-value ratio (*loan-value ratio* > 0.95)	0.189**	1.49**	0.79**	0.79**	0.84**	0.79**
	(0.050)	(0.32)	(0.18)	(0.18)	(0.18)	(0.18)
consumer credit score	0.031**	0.29**	0.15**	0.16**	0.34**	0.16**
	(0.005)	(0.04)	(0.02)	(0.02)	(0.11)	(0.02)
mortgage credit score	0.021	0.28*	0.15*	0.11	0.16	0.11
	(0.011)	(0.14)	(0.07)	(0.08)	(0.10)	(0.08)
public bad credit record	0.197**	1.23**	0.70**	0.70**	0.72**	0.70**
	(0.035)	(0.20)	(0.12)	(0.12)	(0.12)	(0.12)
denied mortgage insurance	0.702**	4.55**	2.56**	2.59**	2.59**	2.59**
	(0.045)	(0.57)	(0.30)	(0.29)	(0.30)	(0.29)
self-employed	0.060**	0.67**	0.36**	0.35**	0.34**	0.35**
	(0.021)	(0.21)	(0.11)	(0.11)	(0.11)	(0.11)
single				0.23**	0.23**	0.23**
				(0.08)	(0.08)	(0.08)
high school diploma				−0.61**	−0.60*	−0.62**
				(0.23)	(0.24)	(0.23)
unemployment rate				0.03	0.03	0.03
				(0.02)	(0.02)	(0.02)
condominium					−0.05	
					(0.09)	
black × *P/I ratio*						−0.58
						(1.47)
black × *housing expense-to-income ratio*						1.23
						(1.69)
additional credit rating indicator variables	no	no	no	no	yes	no
constant	−0.183**	−5.71**	−3.04**	−2.57**	−2.90**	−2.54**
	(0.028)	(0.48)	(0.23)	(0.34)	(0.39)	(0.35)

(Table 11.2 continued)

F-Statistics and p-Values Testing Exclusion of Groups of Variables

	(1)	(2)	(3)	(4)	(5)	(6)
applicant single; high school diploma; industry unemployment rate				5.85 (< 0.001)	5.22 (0.001)	5.79 (< 0.001)
additional credit rating indicator variables					1.22 (0.291)	
race interactions and black						4.96 (0.002)
race interactions only						0.27 (0.766)
difference in predicted probability of denial, white vs. black (percentage points)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

These regressions were estimated using the $n = 2380$ observations in the Boston HMDA data set described in Appendix 11.1. The linear probability model was estimated by OLS, and probit and logit regressions were estimated by maximum likelihood. Standard errors are given in parentheses under the coefficients, and *p*-values are given in parentheses under the *F*-statistics. The change in predicted probability in the final row was computed for a hypothetical applicant whose values of the regressors, other than race, equal the sample mean. Individual coefficients are statistically significant at the *5% or **1% level.

Interpretation of results based on Table 11.2 I

- In the 1990s, loan officers commonly used thresholds, or cutoff values for the loan-to-value ratio.;
- Because the coefficients of the logit and probit models in columns (2)–(6) are not directly interpretable, the table reports standard errors but not confidence intervals. The table reports whether the test that the coefficient is 0 rejects at the 5% or 1% significance level.
- (1) is a linear probability model, its coefficients are estimated changes in predicted probabilities due to a unit change in the independent variable.

Interpretation of results based on Table 11.2 II

- Applicants with a public record of credit problems, such as filing for bankruptcy, have much greater difficulty obtaining a loan: All else equal, a public bad credit record is estimated to increase the probability of denial by 0.197, or 19.7 percentage points. Being denied private mortgage insurance is estimated to be virtually decisive: The estimated coefficient of 0.702 means that being denied mortgage insurance increases your chance of being denied a mortgage by 70.2 percentage points, all else equal.
- Of the nine variables (other than race) in the regression, the coefficients on all but two are statistically significant at the 5% level, which is consistent with loan officers' considering many factors when they make their decisions. The coefficient on black in regression (1) is 0.084, indicating that the difference in denial probabilities for black and white applicants is 8.4 percentage points, holding constant the other variables in the regression. This is statistically significant at the 1% significance level $t = 3.65$.

Interpretation of results based on Table 11.2 III

- The logit and probit estimates reported in columns (2) and (3) yield similar conclusions. In the logit and probit regressions, eight of the nine coefficients on variables other than race are individually statistically significantly different from 0 at the 5% level, and the coefficient on black is statistically significant at the 1% level. As discussed in Section 11.2, because these models are nonlinear, specific values of all the regressors must be chosen to compute the difference in predicted probabilities for white applicants and black applicants.
- A conventional way to make this choice is to consider an "average" applicant who has the sample average values of all the regressors other than race. The final row in Table 11.2 reports this estimated difference in probabilities, evaluated for this average applicant. The estimated racial differentials are similar to each other: 8.4 percentage points for the linear probability model [column (1)], 6.0 percentage points for the logit model

Interpretation of results based on Table 11.2 IV

[column (2)], and 7.1 percentage points for the probit model [column (3)]. These estimated race effects and the coefficients on black are less than in the regressions of the previous sections, in which the only regressors were P/I ratio and black, indicating that those earlier estimates had omitted variable bias.

- The regressions in columns (4) through (6) investigate the sensitivity of the results in column (3) to changes in the regression specification. Column (4) modifies column (3) by including additional applicant characteristics. These characteristics help to predict whether the loan is denied; for example, having at least a high school diploma reduces the probability of denial (the estimate is negative, and the coefficient is statistically significant at the 1% level). However, controlling for these personal characteristics does not change the estimated coefficient on black or the estimated difference in denial probabilities (6.6%) in an important way.

Interpretation of results based on Table 11.2 V

- Column (5) breaks out the six consumer credit categories and four mortgage credit categories to test the null hypothesis that these two variables enter linearly; this regression also adds a variable indicating whether the property is a condominium. The null hypothesis that the credit rating variables enter the expression for the z-value linearly is not rejected, nor is the condominium indicator significant, at the 5% level. Most importantly, the estimated racial difference in denial probabilities (6.3%) is essentially the same as in columns (3) and (4).
- Column (6) examines whether there are interactions. Are different standards applied to evaluating the payment-to-income and housing expense-to-income ratios for black applicants versus white applicants? The answer appears to be no: The interaction terms are not jointly statistically significant at the 5% level. However, race continues to have a significant effect, because the race indicator and the interaction terms are jointly statistically

Interpretation of results based on Table 11.2 VI

significant at the 1% level. Again, the estimated racial difference in denial probabilities (6.5%) is essentially the same as in the other probit regressions.

- In all six specifications, the effect of race on the denial probability, holding other applicant characteristics constant, is statistically significant at the 1% level. The estimated difference in denial probabilities between black applicants and white applicants ranges from 6.0 percentage points to 8.4 percentage points.
- One way to assess whether this differential is large or small is to return to a variation on the question posed at the beginning of this chapter. Suppose two individuals apply for a mortgage, one white and one black, but otherwise having the same values of the other independent variables in regression (3); specifically, aside from race, the values of the other variables in regression (3) are the sample average values in the HMDA data set. The white applicant faces a 7.4% chance of denial, but the black applicant faces a 14.5% chance of denial. The estimated racial difference in denial probabilities, 7.1 percentage points, means that the black applicant is nearly twice as likely to be denied as the white applicant.

Conclusion I

- The results in Table 11.2 (and in the original Boston Fed study) provide statistical evidence of racial patterns in mortgage denial that.

Criticisms I

A number of the criticisms made of the original Federal Reserve Bank of Boston study concern internal validity:

- possible errors in the data,
- alternative nonlinear functional forms,
- additional interactions, and so forth.
- The original data were subjected to a careful audit, some errors were found, and the results reported here (and in the final published Boston Fed study) are based on the "cleaned" data set.
- Estimation of other specifications—different functional forms and/or additional regressors—also produces estimates of racial differentials comparable to those in Table 11.2.
- A potentially more difficult issue of internal validity is whether there is relevant nonracial financial information obtained during in-person loan interviews, but not recorded on the loan application itself, that is correlated with race; if so, there still might be omitted variable bias in the Table 11.2 regressions.

Criticisms II

- Finally, some have questioned external validity: Even if there was racial discrimination in Boston in 1990, it is wrong to implicate lenders elsewhere today.
- Moreover, racial discrimination might be less likely using modern online applications because the mortgage can be approved or denied without a face-to-face meeting. The only way to resolve the question of external validity is to consider data from other locations and years.

References I

Stock, J. H. and Watson, M. W. (2020). *Introduction to econometrics*, volume 4. Pearson New York.