Maximum Likelihood Estimation

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.

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 [Equation (2.23)], 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 $\Pr(Y = y) = p^y(1-p)^{1-y}$: When $y = 1$, $\Pr(Y = 1) = p^1(1-p)^0 = p$, and when $y = 0$, $\Pr(Y = 0) = p^0(1-p)^1 = 1 - p$. Thus the joint probability distribution of Y_1 and Y_2 is $\Pr(Y_1 = y_1, Y_2 = y_2) = [p^{y_1}(1-p)^{1-y_1}] \times [p^{y_2}(1-p)^{1-y_2}] = p^{(y_1+y_2)}(1-p)^{2-(y_1+y_2)}$.

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

$$f(p; Y_1, Y_2) = p^{(Y_1+Y_2)}(1-p)^{2-(Y_1+Y_2)}. \tag{11.12}$$

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 $\hat{p}$ of the Bernoulli probability p is the sample average; that is, $\hat{p} = \overline{Y}$ (this is shown in Appendix 11.2). In this example, the MLE is the usual estimator of p, the fraction of times $Y_i = 1$ in the sample.

This example is similar to the problem of estimating the unknown coefficients of the probit and logit regression models. In those models, the success probability p is not constant, but rather depends on X; that is, it is the success probability conditional on X, which is given in Equation (11.6) for the probit model and Equation (11.9) for the logit model. Thus the probit and logit likelihood functions are similar to the likelihood function in Equation (11.12) except that the success probability varies from one observation to the next (because it depends on X_i). Expressions for the probit and logit likelihood functions are given in Appendix 11.2.

Like the nonlinear least squares estimator, the MLE is consistent and normally distributed in large samples. Because regression software commonly computes the MLE of the probit coefficients, this estimator is easy to use in practice. All the estimated probit and logit coefficients reported in this chapter are MLEs.

Statistical inference based on the MLE. 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 ± 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 $q \times F$, where q is the number of restrictions being tested. Because the F-statistic is, under the null hypothesis, distributed as χ_q^2/q in large samples, $q \times F$ is distributed as χ_q^2 in large samples. Because the two approaches 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

In Section 11.1, 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-R^2." 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 Y_i 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: 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. A formula for the pseudo-R^2 is given in Appendix 11.2.

11.4 Application to the Boston HMDA Data

The regressions of the previous two sections indicated that denial rates were higher for black than white applicants, holding constant their payment-to-income ratio. Loan officers, however, legitimately weigh many factors when deciding on a mortgage application, and if any of those other factors differ systematically by race, the estimators considered so far have omitted variable bias.

In this section, we take a closer look at whether there is statistical evidence of discrimination in the Boston HMDA data. Specifically, our objective is to estimate the effect of race on the probability of denial, holding constant those applicant characteristics that a loan officer might legally consider when deciding on a mortgage application.

The most important variables available to loan officers through the mortgage applications in the Boston HMDA data set are listed in Table 11.1; these are the variables we will focus on in our empirical models of loan decisions. The first two variables are direct measures of the financial burden the proposed loan would place on the applicant, measured in terms of his or her income. The first of these is the *P/I ratio*; the second is the ratio of housing-related expenses to income. The next variable is the size of the loan, relative to the assessed value of the home; if the loan-to-value ratio is nearly 1, the bank might have trouble recouping the full amount of the loan if the applicant defaults on the loan and the bank forecloses. The final three financial variables summarize the applicant's credit history. If an applicant has been unreliable paying off debts in the past, the loan officer legitimately

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

might worry about the applicant's ability or desire to make mortgage payments in the future. The three variables measure different types of credit histories, which the loan officer might weigh differently. The first concerns consumer credit, such as credit card debt; the second is previous mortgage payment history; and the third