8

The logit and probit models

The dependent variable in most regression models is numerical, measured usually on a **ratio scale**. But in many applications the dependent variables are **nominal** in the sense that they denote categories, such as male or female, married or unmarried, employed or unemployed, in the labor force or not in the labor force.

Suppose we have data on adults, some of who smoke and some who do not. Further suppose we want to find out what factors determine whether a person smokes or not. So the variable smoking status is a nominal variable; you either smoke or you do not. How do we model such nominal variables? Can we use the traditional regression techniques or do we need specialized techniques?

Regression models involving nominal scale variables are an example of a broader class of models known as **qualitative response regression models**. There are a variety of such models, but in this chapter we will consider the simplest of such models, namely the **binary** or **dichotomous** or **dummy** dependent variable regression models. In subsequent chapters we will consider other types of qualitative response regression models.

The aim of this chapter is to show that although binary variable regression models can be estimated with the least-squares method, such models are usually estimated by specialized methods, such as **logit** and **probit**. First we will show why the least-squares method is not appropriate and then consider the logit and probit models. We begin with an example.

8.1 An illustrative example: to smoke or not to smoke

The data used here is a random sample of 1,196 US males.[1] These data are provided in **Table 8.1**, which can be found on the companion website.

The variables used in the analysis are as follows:

Smoker = 1 for smokers and 0 for nonsmokers

Age = age in years

Education = number of years of schooling

Income = family income

Pcigs = price of cigarettes in individual states in 1979

[1] These data are from the website of Michael P. Murray, *Econometrics: A Modern Introduction*, Addison-Wesley, Boston, 2006. See http://www.aw.-bc.com/murray. But the data were originally used by John Mullay, Instrumental-variable estimation of count data models: an application to models of cigarette smoking behavior, *The Review of Economics and Statistics*, 1997.

8.2 The linear probability model (LPM)

Since the dependent variable, smoker, is a nominal variable, it takes a value of 1 (for smoker) and 0 (for nonsmoker). Suppose we routinely apply the method of ordinary least-squares (OLS) to determine smoking behavior in relation to age, education, family income, and price of cigarettes. That is, we use the following model:

$$\begin{aligned} Y_i &= B_1 + B_2 Age_i + B_3 Educ_i + B_4 Income_i \\ &\quad + B_5 Pcigs + u_i \end{aligned} \tag{8.1}$$

which, for brevity of expression, we write as:

$$Y_i = \boldsymbol{BX} + u_i \tag{8.2}$$

where ***BX*** is the right-hand side of Eq. (8.1).

Model (8.2) is called a **linear probability model** (**LPM**) because the conditional expectation of the depending variable (smoking status), given the values of the explanatory variables, can be interpreted as the *conditional probability* that the event (i.e. smoking) will occur.[2]

Using *Eviews*, we obtained the results in Table 8.2. Let us examine the results in this table.

Notice that all the variables, except income, are individually statistically significant at least at the 10% level of significance.

Age, education, and price of cigarettes have negative impact on smoking, which may not be a surprising result. Collectively all the explanatory variables are statistically significant, for the estimated *F* value of ≈12.00 has a *p* value of almost zero. Recall that the *F* value tests the hypothesis that all the slope coefficients are simultaneously equal to zero.

Table 8.2 LPM model of to smoke or not to smoke.

Dependent Variable: SMOKER
Method: Least Squares
Date: 12/06/08 Time: 21:54
Sample: 1 1196
Included observations: 1196

	Coefficient	Std. Error	t-Statistic	Prob.
C	1.123089	0.188356	5.962575	0.0000
AGE	−0.004726	0.000829	−5.700952	0.0000
EDUC	−0.020613	0.004616	−4.465272	0.0000
INCOME	1.03E−06	1.63E−06	0.628522	0.5298
PCIGS79	−0.005132	0.002852	−1.799076	0.0723

R-squared	0.038770	Mean dependent var	0.380435
Adjusted R-squared	0.035541	S.D. dependent var	0.485697
S.E. of regression	0.476988	Akaike info criterion	1.361519
Sum squared resid	270.9729	Schwarz criterion	1.382785
Log likelihood	−809.1885	Durbin–Watson stat	1.943548
F-statistic	12.00927	Prob(F-statistic)	0.000000

[2] If $P_i = \Pr(Y_i = 1)$ and $(1 - P_i) = \Pr(Y_i = 0)$, then the expected value of $Y_i = E(Y_i) = 1.P_i + 0.(1 - P_i) = P_i$.

Since we have estimated a linear probability model, the interpretation of the regression coefficients is as follows. If we hold all other variables constant, the probability of smoking decreases at the rate of ≈ 0.005 as a person ages, probably due to the adverse impact of smoking on health. Likewise, *ceteris paribus*, an increase in schooling by one year decreases the probability of smoking by 0.02. Similarly, if the price of cigarettes goes up by a dollar, the probability of smoking decreases by ≈ 0.005, holding all other variables constant. The R^2 value of ≈ 0.038 seems very low, but one should not attach much importance to this because the dependent variable is nominal, taking only values of 1 and zero.

We can refine this model by introducing *interaction terms*, such as age multiplied by education, or education multiplied by income, or introduce a squared term in education or squared term in age to find out if there is nonlinear impact of these variables on smoking. But there is no point in doing that, because the LPM has several inherent limitations.

First, the LPM assumes that the probability of smoking moves linearly with the value of the explanatory variable, no matter how small or large that value is. *Secondly*, by logic, the probability value must lie between 0 and 1. But there is no guarantee that the *estimated* probability values from the LPM will lie within these limits. This is because OLS does not take into account the restriction that the estimated probabilities must lie within the bounds of 0 and 1. *Thirdly*, the usual assumption that the error term is normally distributed cannot hold when the dependent variable takes only values of 0 and 1. *Finally*, the error term in the LPM is heteroscedastic, making the traditional significance tests suspect.

For all these reasons, LPM is not the preferred choice for modeling dichotomous variables. The alternatives discussed in the literature are **logit** and **probit**.

8.3 The logit model

In our smoker example our primary objective is to estimate the probability of smoking, given the values of the explanatory variables. In developing such a probability function, we need to keep in mind two requirements: (1) that as X_i, the value of the explanatory variable(s) changes, the estimated probability always lies in the 0–1 interval, and (2) that the relationship between P_i and X_i is nonlinear, that is, "one which approaches zero at slower and slower rates as X_i gets small and approaches one at slower and slower rates as X_i gets very large".[3] The logit and probit models satisfy these requirements. We first consider the logit model because of its comparative mathematical simplicity.

Assume that in our example the decision of an individual to smoke or not to smoke depends on an *unobservable* ***utility index*** I_i^*, which depends on explanatory variables such as age, education, family income and price of cigarettes.[4] We express this index as:

$$I_i^* = BX + u_i \tag{8.3}$$

where i = ith individual, u = error term, and $\boldsymbol{BX}$ is as defined in Eq. (8.2).

[3] John H. Aldrich and Forrest Nelson, *Linear Probability, Logit and Probit Models*, Sage Publications, 1984, p. 26.

[4] The utility index is also known as a latent variable.

But how is the unobservable index related to the actual decision of smoking or not smoking? It is reasonable to assume that:

$Y_i = 1$ (a person smokes) if $I_i^* \geq 0$

$Y_i = 0$ (a person does not smoke) if $I_i^* < 0$

That is, if a person's utility index I exceeds the threshold level I^*, he or she will smoke but if it is less that I^*, that individual will not smoke. Note that we are not suggesting that smoking is good or bad for health, although there is extensive medical research that suggests that smoking probably is bad for health.

To make this choice operational, we can think in terms of the probability of making a choice, say the choice of smoking (i.e. $Y = 1$):

$$\begin{aligned}\Pr(Y_i = 1) &= \Pr(I^* \geq 0)\\ &= \Pr[(BX + u_i) \geq 0] \\ &= \Pr(u_i \geq -BX)\end{aligned} \tag{8.4}$$

Now this probability depends on the (probability) distribution of Y_i, which in turn depends on the probability distribution of the error term, u_i.[5] If this probability distribution is symmetric around its (zero) mean value, then Eq. (8.4) can be written as:

$$\Pr(u_i \geq -BX) = \Pr(u_i \leq BX) \tag{8.5}$$

Therefore,

$$P_i = \Pr(Y_i = 1) = \Pr(u_i \leq BX) \tag{8.6}$$

Obviously P_i depends on the particular probability distribution of u_i. Remember that the probability that a random variable takes a value less than some specified value is given by the **cumulative distribution function** (**CDF**) of that variable.[6]

The logit model assumes that the probability distribution of u_i follows the **logistic probability distribution**, which for our example can be written as:

$$P_i = \frac{1}{1 + e^{-Z_i}} \tag{8.7}$$

where P_i = probability of smoking (i.e. $Y_i = 1$) and

$$Z_i = BX + u_i \tag{8.8}$$

The probability that $Y = 0$, that is, the person is not a smoker, is given by

$$1 - P_i = \frac{1}{1 + e^{Z_i}} \tag{8.9}$$

Note: The signs of Z_i in Eqs. (8.7) and (8.9) are different.

[5] Note that B is fixed or nonrandom and X values are given. Therefore, the variation in Y_i comes from the variation in u_i.

[6] Recall from elementary statistics that the *cumulative distribution function* of a random variable X, $F(X)$, is defined as: $F(X) = \Pr(X \leq x)$, where x is a particular value of X. Also recall that if you plot CDF, it resembles an elongated S.

It can be easily verified that as Z_i ranges from $-\infty$ to $+\infty$, P_i ranges between 0 and 1 and that P_i is nonlinearly related to Z_i (i.e. X_i), thus satisfying the requirements discussed earlier.[7]

How do we estimate model (8.7), for it is nonlinear not only in ***X*** but also in the parameters, ***B***s? We can use a simple transformation to make the model linear in the *X*s and the coefficients. Taking the ratio of Eqs. (8.7) and (8.9), that is the probability that a person is a smoker against the probability that he/she is not, we obtain:

$$\frac{P_i}{1-P_i}=\frac{1+\mathrm{e}^{Z_i}}{1+\mathrm{e}^{-Z_i}}=\mathrm{e}^{Z_i} \tag{8.10}$$

Now $P_i/(1-P_i)$ is simply the **odds ratio** in favor of smoking – the ratio of the probability that a person is a smoker to the probability that he or she is not a smoker.

Taking the (natural) log of Eq. (8.10), we obtain a very interesting result, namely:

$$L_i=\ln\left(\frac{P_i}{1-P_i}\right)=Z_i=BX_i+u_i \tag{8.11}$$

In words, Eq. (8.11) states that the log of the odds ratio is a linear function of the *B*s as well as the *X*s. L_i is know as the **logit** (log of the odds ratio) and hence the name **logit model** for models like (8.11). It is interesting to observe that the linear probability model (LPM) discussed previously assumes that P_i is linearly related to X_i, whereas the logit model assumes that the log of the odds ratio is linearly related to X_i.

Some of the features of the logit model are as follows:

1. As P_i, the probability goes from 0 to 1, the logit L_i goes from $-\infty$ to $+\infty$. That is, although the probabilities lie between 0 and 1, the logits are unbounded.
2. Although L_i is linear in X_i, the probabilities themselves are not. This is contrast to with the LPM where the probabilities increase linearly with X_i.
3. If L_i, the logit, is positive, it means that when the value of the explanatory variable(s) increases, the odds of smoking increases, whereas it if is negative, the odds of smoking decreases.
4. The interpretation of the logit model in (8.11) is as follows: each slope coefficient shows how the log of the odds in favor of smoking changes as the value of the *X* variable changes by a unit.
5. Once the coefficients of the logit model are estimated, we can easily compute the probabilities of smoking, not just the odds of smoking, from (8.7).
6. In the LPM the slope coefficient measures the marginal effect of a unit change in the explanatory variable on the probability of smoking, holding other variables constant. This is not the case with the logit model, for the marginal effect of a unit change in the explanatory variable not only depends on the coefficient of that variable but also on the level of probability from which the change is measured. But the latter depends on the values of all the explanatory variables in the model.[8]

[7] The reason why P_i is nonlinearly related to, say, income is that as income increases smokers will increase their consumption of cigarettes at a decreasing rate because of the law of diminishing returns. This is true of almost all normal commodities.

[8] Calculus-minded readers can verify this if they take the (partial) derivative of Eq. (8.7) with respect to the relevant variables, noting that $Z_i = BX$. Note: use the chain rule: $\partial P_i/\partial X_i = \partial P_i/\partial Z_i \cdot \partial Z_i/\partial X_i$.

However, statistical packages such as *Eviews* and *Stata* can compute the marginal effects with simple instructions.

Now the question is: how do we estimate the parameters of the logit model?

Estimation of the logit model

Estimation of the logit model depends on the type of data available for analysis. There are two types of data available: data at the individual, or micro, level, as in the case of the smoker example, and data at the group level. We will first consider the case of individual level data.

Individual level data

For our smoker example, we have data on 1,196 individuals. Therefore, although the logit model is linear, it cannot be estimated by the usual OLS method. To see why, note that $P_i = 1$ if a person smokes, and $P_i = 0$ if a person does not smoke. But if we put these values directly in the logit L_i, we obtain expressions like $L_i = \ln(1/0)$ if a person smokes and $L_i = \ln(0/1)$ if a person does not smoke. These are undefined expressions. Therefore, to estimate the logit model we have to resort to alternative estimation methods. The most popular method with attractive statistical properties is the method of **maximum likelihood** (**ML**). We briefly discussed this method in Chapter 1, but further details of ML can be found in the references.[9] Most modern statistical packages have established routines to estimate parameters by the ML method.

We will first present the results of ML estimation for the smoker example, which are obtained from *Eviews* (Table 8.3).

Table 8.3 Logit model of to smoke or not to smoke.

Dependent Variable: SMOKER
Method: ML – Binary Logit (Quadratic hill climbing)
Sample: 1 1196
Included observations: 1196
Convergence achieved after 3 iterations
QML (Huber/White) standard errors & covariance

	Coefficient	Std. Error	z-Statistic	Prob.
C	2.745077	0.821765	3.340462	0.0008
AGE	–0.020853	0.003613	–5.772382	0.0000
EDUC	–0.090973	0.020548	–4.427431	0.0000
INCOME	4.72E–06	7.27E–06	0.649033	0.5163
PCIGS79	–0.022319	0.012388	–1.801626	0.0716

McFadden R-squared	0.029748	Mean dependent var	0.380435
S.D. dependent var	0.485697	S.E. of regression	0.477407
Akaike info criterion	1.297393	Sum squared resid	271.4495
Schwarz criterion	1.318658	Log likelihood	–770.8409
LR statistic	47.26785	Restr. log likelihood	–794.4748
Prob(LR statistic)	0.000000	Avg. log likelihood	–0.644516
Obs with Dep=0	741	Total obs	1196
Obs with Dep=1	455		

[9] For an accessible discussion of ML, see Gujarati/Porter, *op cit.*

Let us examine these results. The variables age and education are highly statistically significant and have the expected signs. As age increases, the value of the logit decreases, perhaps due to health concerns – that is, as people age, they are less likely to smoke. Likewise, more educated people are less likely to smoke, perhaps due to the ill effects of smoking. The price of cigarettes has the expected negative sign and is significant at about the 7% level. *Ceteris paribus*, the higher the price of cigarettes, the lower is the probability of smoking. Income has no statistically visible impact on smoking, perhaps because expenditure on cigarettes may be a small proportion of family income.

The interpretation of the various coefficients is as follows: holding other variables constant, if, for example, education increases by one year, the average logit value goes down by ≈ 0.09, that is, the log of odds in favor of smoking goes down by about 0.09. Other coefficients are interpreted similarly.

But the logit language is not everyday language. What we would like to know is the probability of smoking, given values of the explanatory variables. But this can be computed from Eq. (8.7). To illustrate, take smoker #2 from **Table 8.1** also. His data are as follows: age = 28, educ = 15, income = 12,500 and pcigs79 = 60.0. Inserting these values in Eq. (8.7), we obtain:

$$P = \frac{1}{1 + e^{-(-0.4935)}} \approx 0.3782$$

That is, the probability that a person with the given characteristics is a smoker is about 38%. From our data we know that this person is a smoker.

Now take a person with age, educ, income, and pcigs79 of 63, 10, 20,000, and 60.8, respectively. For this person, the probability of smoking is

$$P = \frac{1}{1 + e^{-(-0.7362)}} = 0.3227$$

That is, the probability of this person being a smoker is 32%. In our sample such a person is nonsmoker.

Table 8.1 gives the probability of smoking for each person along with the raw data.

Can we compute the marginal effect of an explanatory variable on the probability of smoking, holding all other variables constant? Suppose we want to find out $\partial P_i / \partial Age_i$, the effect of a unit change in age on the probability of smoking, holding other variables constant. This was very straightforward in the LPM, but it is not that simple with logit or probit models. This is because the change in probability of smoking if age changes by a unit (say, a year) depends not only on the coefficient of the age variable but also on the level of probability from which the change is measured. But the latter depends on values of all the explanatory variables. For details of these computations the reader is referred to the references, although *Eviews* and *Stata* can do this job readily.[10]

The conventional measure of goodness of fit, R^2, is not very meaningful when the dependent variable takes values of 1 or 0. Measures similar to R^2, called **pseudo** $\boldsymbol{R^2}$, are discussed in the literature. One such measure is the McFadden R^2, called $R^2{}_{\text{McF}}$. Like R^2, $R^2{}_{\text{McF}}$ lies between 0 and 1. For our example, its value is 0.0927.

Another goodness of fit measure is the **count** $\boldsymbol{R^2}$, which is defined as

[10] See, for instance, Gujarati/Porter, *op cit.*

$$\text{Count } R^2 = \frac{\text{number of correct predictions}}{\text{total number of observations}} \tag{8.12}$$

Since the dependent variable takes a value of 1 or 0, if the predicted probability for an observation is greater than 0.5 we classify that observation as 1, but if is less than 0.5, we classify that as 0. We then count the number of correct predictions and the count R^2 as defined above (see Exercise 8.3).

It should be emphasized that in binary regression models goodness of fit measures are of secondary importance. What matters are the expected signs of the regression coefficients and their statistical and or practical significance. From Table 8.3 we can see that except for the income coefficient, all other coefficients are individually statistically significant, at least at the 10% level. We can also test the null hypothesis that all the coefficients are simultaneously zero with the **likelihood ratio (LR) statistic**, which is the equivalent of the F test in the linear regression model.[11] Under the null hypothesis that none of the regressors are significant, the LR statistic follows the chi-square distribution with df equal to the number of explanatory variables: four in our example.

As Table 8.3 shows, the value of the LR statistic is about 47.26 and the *p value* (i.e. the exact significance level) is practically zero, thus refuting the null hypothesis. Therefore we can say that the four variables included in the logit model are important determinants of smoking habits.

▲ **Technical Note 1**: Table 8.3 gives two log likelihood statistics – unrestricted likelihood (= –770.84) and restricted likelihood (–794.47). The latter is obtained by assuming that there are no regressors in the model, only the intercept term, whereas the unrestricted likelihood is the value obtained with all the regressors (including the intercept) in the model. The likelihood ratio statistic (=λ) of about 47.27 shown in Table 8.3 is computed from the formula given in the Appendix to Chapter 1. For our example, the computed likelihood ratio of 47.27 is highly significant, for its *p* value is practically zero.[12] This is to say that it is the unrestricted model that includes all the regressors is appropriate in the present instance. To put it differently, the restricted model is not valid in the present case.

▲ **Technical Note 2**: Note that the Huber/White standard errors reported in Table 8.3 are not necessarily robust to heteroscedasticity but are robust to certain misspecification of the underlying probability distribution of the dependent variable.

Model refinement

The logit model given in Table 8.3 can be refined. For example, we can allow for the interaction effect between the explanatory variables. Individually education has negative impact and income has positive impact on the probability of smoking, although the latter effect is not statistically significant. But what is the combined influence of education and income on the probability of smoking? Do people with a higher level of

[11] In the maximum likelihood Appendix to Chapter 1 we have discussed why we use the LR statistic.

[12] As noted in the Appendix to Chapter 1, under the null hypothesis that the coefficients of all regressors in the model are zero, the LR statistic follows the chi-square distribution with df equal to the number of regressors (excluding the intercept), 4 in our example.

Table 8.4 The logit model of smoking with interaction.

Dependent Variable: SMOKER
Method: ML – Binary Logit (Quadratic hill climbing)
Sample: 1 1196
Included observations: 1196
Convergence achieved after 10 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	1.093186	0.955676	1.143887	0.2527
AGE	–0.018254	0.003794	–4.811285	0.0000
EDUC	0.039456	0.042511	0.928140	0.3533
INCOME	9.50E–05	2.69E–05	3.535155	0.0004
PCIGS79	–0.021707	0.012530	–1.732484	0.0832
EDUC*INCOME	–7.45E–06	2.13E–06	–3.489706	0.0005

McFadden R-squared	0.037738	Mean dependent var	0.380435
S.D. dependent var	0.485697	S.E. of regression	0.475290
Akaike info criterion	1.288449	Sum squared resid	268.8219
Schwarz criterion	1.313968	Log likelihood	–764.4926
LR statistic	59.96443	Restr. log likelihood	–794.4748
Prob(LR statistic)	0.000000	Avg. log likelihood	–0.639208
Obs with Dep=0	741	Total obs	1196
Obs with Dep=1	455		

education and higher level of income smoke less or more than people with other characteristics?

To allow for this, we can introduce the multiplicative or interactive effect of the two variables as an additional explanatory variable. The results are given in Table 8.4.

These results are interesting. In Table 8.3 individually education had a significant negative impact on the logit (and therefore on the probability of smoking) and income had no statistically significant impact. Now education by itself has no statistically significant impact on the logit, but income has highly significant positive impact. But if you consider the interactive term, education multiplied by income, it has significant negative impact on the logit. That is, persons with higher education who also have higher incomes are less likely to be smokers than those who are more educated only or have higher incomes only. What this suggests is that the impact of one variable on the probability of smoking may be attenuated or reinforced by the presence of other variable(s).

The reader is encouraged to see if there are any other interactions among the explanatory variables.

Logit estimation for grouped data

Suppose we group the smoker data into 20 groups of approximately 60 observations each. For each group we find out the number of smokers, say n_i. We divide n_i by 60 to get an estimate of the (empirical) probability of smokers for that group, say, p_i. Therefore, we have 20 estimated p_is. We can then use these probabilities to estimate the logit regression Eq. (8.11) by OLS.

Unless the data are already available in grouped form, forming groups in the manner suggested in the preceding paragraph has problems. *First*, we have to decide how many groups to form. If we form too few groups, we will have very few p_i to estimate Eq. (8.11). On the other hand, if we form too many groups, we will have only a few observations in each group, which might make it difficult to estimate the p_is efficiently.

Second, even if we have the "right" number of groups, one problem with the grouped logit estimation is that the error term in Eq. (8.11) is heteroscedastic. So we will have to take care of heteroscedasticity by suitable transformation or use White's robust standard errors, a topic discussed in Chapter 5.

We will not illustrate the grouped logit estimation with the smoker data for the reasons discussed above. Besides, we have data at the micro-level and we can use the ML method to estimate the logit model, as we have shown earlier (but see Exercise 8.4).

8.4 The probit model

In the LPM the error term has non-normal distribution; in the logit model the error term has the logistic distribution. Another rival model is the **probit model**, in which the error term has the normal distribution. Given the assumption of normality, the probability that I_i^* is less than or equal to I_i can be computed from the **standard normal cumulative distribution function** (**CDF**)[13] as:

$$P_i = \Pr(Y = 1 | X) = \Pr(I_i^* \leq I_i) = \Pr(Z_i \leq BX) = F(BX) \tag{8.13}$$

where $\Pr(Y|X)$ means the probability that an event occurs (i.e. smoking) given the values of the X variables and where Z is the standard normal variable (i.e. a normal variable with zero mean and unit variance). F is the standard normal CDF, which in the present context can be written as:

$$F(BX) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{BX} \mathrm{e}^{-z^2/2} \mathrm{d}z \tag{8.14}$$

Since P represents the probability that a person smokes, it is measured by the area of the standard CDF curve from $-\infty$ to I_i. In the present context, $F(I_i)$ is called the **probit function**.

Although the estimation of the utility index BX and the ***B***s is rather complicated in the probit model, the method of maximum likelihood can be used to estimate them. For our example, the ML estimates of the probit model are given in Table 8.5.

Although the numerical values of the logit and probit coefficients are different, qualitatively the results are similar: the coefficients of age, education, and price of cigarettes are individually significant at least at the 10% level. The income coefficient, however, is not significant.

There is a way of comparing the logit and probit coefficients. Although the standard logistic distribution (the basis of the logit) and the standard normal distribution (the basis of probit) both have a mean value of zero, their variances are different: 1 for

[13] If a variable X follows the normal distribution with mean μ and variance σ^2, its probability density function (PDF) is $f(X) = (1/\sigma\sqrt{2\pi})\mathrm{e}^{-(X-\mu)^2/2\sigma^2}$ and its cumulative distribution function (CDF) is $F(X_0) = \int_{-\infty}^{X_0} (1/\sigma\sqrt{2\pi})\mathrm{e}^{-(X-\mu)^2/2\sigma^2} \mathrm{d}X$, where X_0 is a specified value of X. If $\mu = 0$ and $\sigma^2 = 1$, the resulting PDF and CDF represent the standard normal PDF and CDF, respectively.

Table 8.5 Probit model of smoking.

Dependent Variable: SMOKER
Method: ML – Binary Probit (Quadratic hill climbing)
Sample: 1 1196
Included observations: 1196
Convergence achieved after 6 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	1.701906	0.510575	3.333315	0.0009
AGE	–0.012965	0.002293	–5.655439	0.0000
EDUC	–0.056230	0.012635	–4.450266	0.0000
INCOME	2.72E–06	4.40E–06	0.618642	0.5362
PCIGS79	–0.013794	0.007696	–1.792325	0.0731

McFadden R-squared	0.030066	Mean dependent var	0.380435
S.D. dependent var	0.485697	S.E. of regression	0.477328
Akaike info criterion	1.296970	Sum squared resid	271.3598
Schwarz criterion	1.318236	Log likelihood	–770.5881
LR statistic	47.77335	Restr. log likelihood	–794.4748
Prob(LR statistic)	0.000000	Avg. log likelihood	–0.644304
Obs with Dep=0	741	Total obs	1196
Obs with Dep=1	455		

the standard normal distribution and $\pi^2/3$ for the logistic distribution, where $\pi \approx 22/7$, which is about 3.14. Therefore, if we multiply the probit coefficient by about 1.81 ($\approx \pi/\sqrt{3}$), you will get approximately the logit coefficient. For example, the probit coefficient of age is –0.0235. If you multiply this coefficient by 1.81, you will get ≈ –0.0233, which is directly comparable to the age coefficient in the logit model given in Table 8.3.

How do we interpret the coefficients of the probit model given in Table 8.5? For example, what is the marginal effect on the probability of smoking if age increases by a year, holding other variables constant? This marginal effect is given by the coefficient of the age variable, –0.0130, multiplied by the value of the normal density function evaluated for all the *X* values for that individual.

To illustrate, consider the data for smoker number 1 in our sample, which are: age = 21, education = 12, income = 8,500, and pcigs 60.6. Putting these values in the standard normal density function given in footnote 13, we obtain: $f(BX) = 0.3983$. Multiplying this by –0.0130, we obtain –0.0051. This means that with the given values of the *X* variables the probability that someone smokes decreases by about 0.005 if age increases by a year. Recall that we had a similar situation in computing the marginal effect of an explanatory variable on the probability of smoking in the logit model.

As you can see, computing the marginal effect of an explanatory variable on the probability of smoking of an individual in this fashion is a tedious job, although the *Stata* and *Eviews* statistical packages can do this job relatively quickly.

Incidentally, the probit estimates of the interaction effect as in the logit model are as shown in Table 8.6.

Table 8.6 The probit model of smoking with interaction.

Dependent Variable: SMOKER
Method: ML – Binary Probit (Quadratic hill climbing)
Sample: 1 1196
Included observations: 1196
Convergence achieved after 10 iterations
Covariance matrix computed using second derivatives

	Coefficient	Std. Error	z-Statistic	Prob.
C	0.682050	0.587298	1.161336	0.2455
AGE	−0.011382	0.002332	−4.880864	0.0000
EDUC	0.024201	0.025962	0.932180	0.3512
INCOME	5.80E−05	1.62E−05	3.588406	0.0003
PCIGS79	−0.013438	0.007723	−1.739941	0.0819
EDUC*INCOME	−4.55E−06	1.28E−06	−3.551323	0.0004

McFadden R-squared	0.038139	Mean dependent var	0.380435
S.D. dependent var	0.485697	S.E. of regression	0.475190
Akaike info criterion	1.287917	Sum squared resid	268.7082
Schwarz criterion	1.313436	Log likelihood	−764.1745
Hannan–Quinn criter.	1.297531	Restr. log likelihood	−794.4748
LR statistic	60.60065	Avg. log likelihood	−0.638942
Prob(LR statistic)	0.000000		
Obs with Dep=0	741	Total obs	1196
Obs with Dep=1	455		

As you can see, the results in Tables 8.4 and 8.6 are quite similar. But you will have to use the conversion factor of about 1.81 to make the probit coefficients directly comparable with the logit coefficients.[14]

In passing it may be noted that we can also estimate the probit model for grouped data, called grouped probit, similar to the grouped logit model. But we will not pursue it here.

Logit vs. probit

Logit and probit models generally give similar results; the main difference between the two models is that the logistic distribution has slightly fatter tails; recall that the variance of a logistically distributed random variable is about $\pi^2/3$, whereas that of a (standard) normally distributed variable it is 1. That is to say, the conditional probability P_i approaches 0 or 1 at a slower rate in logit than in probit. But in practice there is no compelling reason to choose one over the other. Many researchers choose the logit over the probit because of its comparative mathematical simplicity.

8.5 Summary and conclusions

In this chapter we discussed the simplest possible qualitative response regression model in which the dependent variable is binary, taking the value of 1 if an attribute is present and the value of 0 if that attribute is absent.

[14] A similar conversion factor for comparing LPM and logit models is given in Exercise 8.1.

Although binary dependent variable models can be estimated by OLS, in which case they are known as linear probability models (LPM), OLS is not the preferred method of estimation for such models because of two limitations, namely, that the estimated probabilities from LPM do not necessarily lie in the bounds of 0 and 1 and also because LPM assumes that the probability of a positive response increases linearly with the level of the explanatory variable, which is counterintuitive. One would expect the rate of increase in probability to taper off after some point.

Binary response regression models can be estimated by the logit or probit models.

The logit model uses the logistic probability distribution to estimate the parameters of the model. Although seemingly nonlinear, the log of the odds ratio, called the logit, makes the logit model linear in the parameters.

If we have grouped data, we can estimate the logit model by OLS. But if we have micro-level data, we have to use the method of maximum likelihood. In the former case we will have to correct for heteroscedasticity in the error term.

Unlike the LPM, the marginal effect of a regressor in the logit model depends not only on the coefficient of that regressor but also on the values of all regressors in the model.

An alternative to logit is the probit model. The underlying probability distribution of probit is the normal distribution. The parameters of the probit model are usually estimated by the method of maximum likelihood.

Like the logit model, the marginal effect of a regressor in the probit model involves all the regressors in the model.

The logit and probit coefficients cannot be compared directly. But if you multiply the probit coefficients by 1.81, they are then comparable with the logit coefficients. This conversion is necessary because the underlying variances of the logistic and normal distribution are different.

In practice, the logit and probit models give similar results. The choice between them depends on the availability of software and the ease of interpretation.

Exercises

8.1 To study the effectiveness of price discount on a six-pack of soft drink, a sample of 5,500 consumers was randomly assigned to 11 discount categories as shown in Table 8.7.[15]

(*a*) Treating the redemption rate as the dependent variable and price discount as the regressor, see whether the logit model fits the data.[16]

(*b*) See whether the probit model does as well as the logit model.

(*c*) Fit the LPM model to these data.

(*d*) Compare the results of the three models. Note that the coefficients of LPM and Logit models are related as follows:

Slope coefficient of LPM = 0.25* Slope coefficient of Logit

Intercept of LPM = 0.25* slope coefficient of Logit + 0.5.

[15] The data are obtained from Douglas Montgomery and Elizabeth Peck from their book, *Introduction to Linear Regression Analysis,* John Wiley & Sons, New York, 1982, p. 243 (notation changed).

[16] The redemption rate is the number of coupons redeemed divided by the number of observations in each price discount category.

Table 8.7 The number of coupons redeemed and the price discount.

Price Discount (cents)	Sample size	Number of coupons redeemed
5	500	100
7	500	122
9	500	147
11	500	176
13	500	211
15	500	244
17	500	277
19	500	310
21	500	343
23	500	372
25	500	391

8.2 **Table 8.8** (available on the companion website) gives data on 78 homebuyers on their choice between adjustable and fixed rate mortgages and related data bearing on the choice.[17]

The variables are defined as follows:

Adjust = 1 if an adjustable mortgage is chosen, 0 otherwise.
Fixed rate = fixed interest rate
Margin = (variable rate – fixed rate)
Yield = the 10-year Treasury rate less 1-year rate
Points = ratio of points on adjustable mortgage to those paid on a fixed rate mortgage
Networth = borrower's net worth

(*a*) Estimate an LPM of adjustable rate mortgage choice.
(*b*) Estimate the adjustable rate mortgage choice using logit.
(*c*) Repeat (b) using the probit model.
(*d*) Compare the performance of the three models and decide which is a better model.
(*e*) Calculate the marginal impact of Margin on the probability of choosing the adjustable rate mortgage for the three models.

8.3 For the smoker data discussed in the chapter, estimate the count R^2.

8.4 Divide the smoker data into 20 groups. For each group compute p_i, the probability of smoking. For each group compute the average values of the regressors and estimate the grouped logit model using these average values. Compare your results with the ML estimates of smoker logit discussed in the chapter. How would you obtain the heteroscedasticity-corrected standard errors for the grouped logit?

[17] These data are obtained from the website of R. Carter Hill, William E. Griffiths and Guay C. Lim, *Principles of Econometrics*, 3rd edn, John Wiley & Sons, 2008.