DATA ANALYSIS USING REGRESSION MODELS

The Business Perspective

Edward W. Frees
University of Wisconsin–Madison

Prentice Hall, Upper Saddle River, NJ 07458

8.1 ESTIMATION USING LEAST SQUARES REGRESSION

We begin our study of regression using binary dependent variables by examining an important issue in the tax accounting literature: determining the special characteristics of taxpayers who use a professional tax preparer. Taxpayer compliance is an important issue in all industrialized nations, including the United States. Although many aspects of tax compliance are as complex as the tax forms themselves, it is known that individuals who use professional tax preparers tend to underreport income by a significant percentage when compared to what they would report if they had prepared the returns themselves. It is thus of interest from a public policy standpoint to determine what influences a taxpayer to elect to have a professional prepare the tax return. Some good background material, references and additional details can be found in a paper entitled, "Determinants of Tax Preparer Usage" by Christian, Gupta and Lin (1992).

The response is an indicator variable. For the taxpayer illustration, the response indicates whether or not a professional tax preparer was used.

To examine this issue, we analyze data from the Ernst & Young/University of Michigan Tax Research Database. Specifically, we examine the 1984 returns from 192 individuals that were randomly selected from the database. These 192 individuals represent about 2% of the 9,762 returns on the database, which is itself meant to represent a random sample of returns. For each individual, we are interested in understanding whether or not they used a professional preparer (PREP = 1 if so, and 0 otherwise) in terms of a number of explanatory variables. These variables include:

CB_PREP

marital status (MS = 1 if married, and 0 otherwise),
whether or not the taxpayer is self-employed (EMP),
(EMP = 1 if self-employed, and 0 otherwise),
whether an additional exemption for age 65 and over is claimed
(AGE 65 = 1, and 0 otherwise),
the number of dependents claimed (DEPS),
the number of schedules filed (SCHS),
the filer's marginal tax rate (MTR),
the logarithm of the total personal income (LOGTPI) and
the type of tax form filed.

8.3 LOGISTIC REGRESSION

Use p to denote the probability of $y = 1$. Then, $p/(1 - p)$ is said to give the odds of $y = 1$.

When the response y is binary, all of the information about the distribution can be summarized by knowing only the probability of a one, $p = \text{Prob}(y = 1)$. In some applications, a simple transformation of p has an important interpretation. The lead example of this is the *odds* transformation, given by $p/(1 - p)$. For example, suppose y is an indicator variable of a horse winning a race, that is, $y = 1$ if the horse wins and $y = 0$ if the horse does not. Interpret p to be the probability of the horse winning the race and, as an example, suppose that $p = 0.25$. Then, the *odds* of the horse winning the race is $0.25/(1.00 - 0.25) = 0.3333$. We might say that the odds of winning are 0.3333 to 1, or one to three. Equivalently, we can say that the probability of not winning is $1 - p = 0.75$. Thus, the odds of the horse not winning are $0.75/(1 - 0.75) = 3$. We interpret this to mean the odds against the horse are three to one.

Odds have a useful interpretation from a betting standpoint. Suppose that we are playing a fair game and that we place a bet of \$1 with odds of one to three. If the horse wins, we get our \$1 back plus winnings of \$3. If the horse loses, we lose our bet of \$1. It is a *fair game* in the sense that the expected value of the game is zero because we win \$3 with probability $p = 0.25$ and lose \$1 with probability $1 - p = 0.75$. From an economic standpoint, the odds provide the important numbers (bet of \$1 and winnings of \$3), not the probabilities. Of course, if we know p, we can always calculate the odds. Similarly, if we know the odds, we can always calculate the probability p.

The log odds transformation, or logit, is defined by $\text{logit}(p) = \ln(p/(1 - p))$.

The difficulty that we encountered in Sections 8.1 and 8.2 was that probabilities vary between zero and one although linear combinations of explanatory variables can vary between $-\infty$ and ∞. Previously, we introduced a simple transformation to convert probabilities into odds. Note that as probabilities vary between zero and one, odds vary between zero and infinity. We now consider an additional transformation, the logarithm of the odds. With this transformation, as probabilities vary between zero and one, log odds vary between $-\infty$ and ∞, thus providing us with a match of the range for linear combinations of explanatory variables. We thus consider the log odds transformation, called the *logit*, defined by

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right).$$

Using $p_i = \text{Prob}(y_i = 1)$, the *logistic regression* equation is defined as

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}. \tag{8.4}$$

Thus, we relate the parameter p to a linear combination of the explanatory variables. An alternative way of expressing equation (8.4) is

$$p_i = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})}}. \tag{8.5}$$

An important feature of this model is that, regardless of values of $\beta_0, \beta_1, \ldots, \beta_k$ and $x_1, x_2, \ldots, x_k$, the true probabilities p lie in the interval [0,1]. The key feature of the logistic regression equation is that we use a linear combination of explanatory variables to represent the log odds of probabilities, as compared to the expected response in ordinary least squares regression.

Another way to think of the logistic regression model is through a so-called threshold interpretation.

Another way to think of the logistic regression model is through a so-called *threshold interpretation*. Here, we assume there is an underlying linear regression model, with continuous random errors, given by

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + e_i.$$

The variable y_i^* is a continuous variable, although unobserved. What we observe is whether or not y_i^* passes some threshold that, for convenience, we take to be zero. We then define the observed response as

$$y_i = \begin{cases} 1 \text{ when } & y_i^* \geq 0 \\ 0 \text{ when } & y_i^* < 0. \end{cases}$$

For example, y_i^* may represent the propensity for a horse to win a race, such as the speed of a horse in a one mile race. We assume, however, that we only observe whether or not a horse wins the race, a binary outcome.

To link the threshold model to the logistic regression equation, assume that the distribution of the random errors can be described using *logistic distribution function*. That is, we assume

$$\text{Prob}(e_i \leq a) = \frac{1}{1 + e^{-a}}. \tag{8.6}$$

The distribution function defined in equation (8.6) is close to the standard normal distribution. Like the idealized histogram of the standard normal, the idealized histogram of the logistic distribution is symmetric about zero. Due to this symmetry, we have

$$\begin{aligned} p_i &= \text{Prob}(y_i = 1) = \text{Prob}(-e_i \leq \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}) \\ &= \text{Prob}(e_i \leq \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) \\ &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})}}. \end{aligned}$$

This is the same model as given in equation (8.5).

The method of *maximum likelihood estimation* is used to determine estimates of the parameters and associated standard errors. A discussion of this estimation technique can be found in Hosmer and Lemeshow (1989). Using our convention for notation, let $b_0, b_1, \ldots, b_k$ denote the maximum likelihood estimates of $\beta_0, \beta_1, \ldots, \beta_k$. With these estimates, we can calculate fitted values as

$$\hat{p}_i = \frac{1}{1 + e^{-(b0+b1xi1+b2xi2+\ldots+bkxik)}}.$$

To illustrate, consider the tax preparer example introduced in Section 8.1. Using LOGTPI, EMP and MTR as explanatory variables, the fitted model is

$$\hat{\text{PREP}} = \frac{1}{1 + e^{-(-9.3218+1.0344\text{LOGTPI}+.8956\text{EMP}-.0512\text{MTR})}}.$$

Following our previous example, for the 27th observation, we have LOGTPI = 5.7714, MTR = 0 and EMP = 0. Thus, our estimate of the probability of using a tax preparer is

$$\hat{\text{PREP}}_{27} = \frac{1}{1 + e^{-[-9.3218+1.0344(5.7714)+.8956(0)-.0512(0)]}} = \frac{1}{1 + e^{3.3518}} = 0.033834.$$

Thus, unlike linear least squares, the logistic regression fitted value is constrained to lie in the interval [0,1].

Consider all individuals who are not self-employed and who are at the zero marginal tax rate, that is, EMP = 0 and MTR = 0. Our estimated probabilities are

$$\hat{\text{PREP}} = \frac{1}{1 + e^{-(-9.3218+1.0344\text{LOGTPI})}}.$$

Figure 8.7 displays a graph of these estimated probabilities. The logistic regression fits form a curve that looks like a tilted "S." From this graph, we see that estimated probabilities lie in the [0,1] interval. Further, one can also see the nonlinear relationship between the explanatory variable LOGTPI and the fitted value. Superimposed on the logistic regression fitted curve is the weighted least squares fitted line, from equation (8.3) with EMP = MTR = 0. Here, we see that these two fits are close to one another over the middle of the range of independent variable LOGTPI. The difference is in the extremes of LOGTPI, where the linear least squares fit fails to accommodate the fact that the fitted values must lie in the interval [0,1].

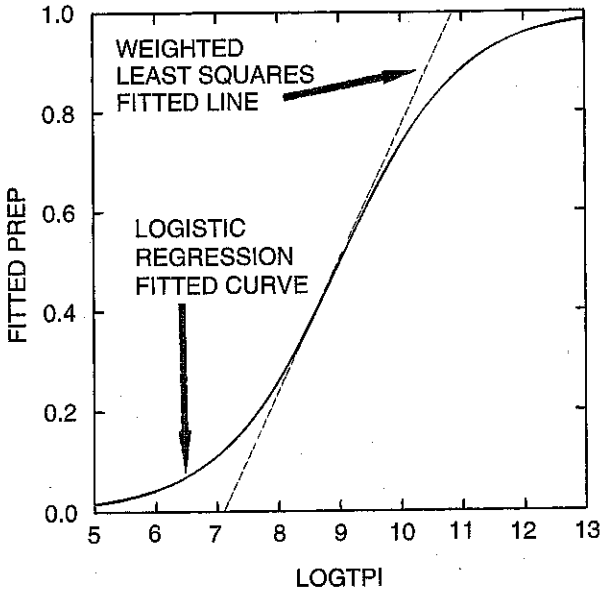

FIGURE 8.7 Comparison of estimated probabilities using the logistic and linear regression models.