(ii) Measures based on the *residual sum of squares*: Effron (1978) suggested using

$$R_2^2 = 1 - [\sum_{i=1}^n (y_i - \widehat{y}_i)^2 / \sum_{i=1}^n (y_i - \bar{y})^2] = 1 - [n \sum_{i=1}^n (y_i - \widehat{y}_i)^2 / n_1 n_2]$$

since $\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - n\bar{y}^2 = n_1 - n(n_1/n)^2 = n_1 n_2/n$, where $n_1 = \sum_{i=1}^n y_i$ and $n_2 = n - n_1$.

Amemiya (1981, p. 1504) suggests using $\left[\sum_{i=1}^n (y_i - \widehat{y}_i)^2 / \widehat{y}_i(1 - \widehat{y}_i)\right]$ as the residual sum of squares. This weights each squared error by the inverse of its variance.

(iii) Measures based on likelihood ratios: $R_3^2 = 1 - (\ell_r/\ell_u)^{2/n}$ where ℓ_r is the restricted likelihood and ℓ_u is the unrestricted likelihood. This tests that all the slope coefficients are zero in the standard linear regression model. For the limited dependent variable model however, the likelihood function has a maximum of 1. This means that $\ell_r \leq \ell_u \leq 1$ or $\ell_r \leq (\ell_r/\ell_u) \leq 1$ or $\ell_r^{2/n} \leq 1 - R_3^2 \leq 1$ or $0 \leq R_3^2 \leq 1 - \ell_r^{2/n}$. Hence, Cragg and Uhler (1970) suggest a pseudo-R^2 that lies between 0 and 1, and is given by $R_4^2 = (\ell_u^{2/n} - \ell_r^{2/n})/[(1 - \ell_r^{2/n})/\ell_u^{2/n}]$. Another measure suggested by McFadden (1974) is $R_5^2 = 1 - (\log\ell_u/\log\ell_r)$.

(iv) Proportion of correct predictions: After computing $\widehat{y}$, one classifies the i-th observation as a success if $\widehat{y}_i > 0.5$, and a failure if $\widehat{y}_i < 0.5$. This measure is useful but may not have enough discriminatory power.

13.9 Empirical Examples

Example 1: Union Participation

To illustrate the logit and probit models, we consider the PSID data for 1982 used in Chapter 4. In this example, we are interested in modelling union participation. Out of the 595 individuals observed in 1982, 218 individuals had their wage set by a union and 377 did not. The explanatory variables used are: years of education (ED), weeks worked (WKS), years of full-time work experience (EXP), occupation ($OCC = 1$, if the individual is in a blue-collar occupation), residence ($SOUTH = 1$, $SMSA = 1$, if the individual resides in the South, or in a standard metropolitan statistical area), industry ($IND = 1$, if the individual works in a manufacturing industry), marital status ($MS = 1$, if the individual is married), sex and race ($FEM = 1$, $BLK = 1$, if the individual is female or black). A full description of the data is given in Cornwell and Rupert (1988). The results of the linear probability, logit and probit models are given in Table 13.3. These were computed using EViews. In fact Table 13.4 gives the probit output. We have already mentioned that the probit model normalizes σ to be 1. But, the logit model has variance $\pi^2/3$. Therefore, the logit estimates tend to be larger than the probit estimates although by a factor less than $\pi/\sqrt{3}$. In order to make the logit results comparable to those of the probit, Amemiya (1981) suggests multiplying the logit coefficient estimates by 0.625.

Similarly, to make the linear probability estimates comparable to those of the probit model one needs to multiply these coefficients by 2.5 and then subtract 1.25 from the constant term. For this example, both logit and probit procedures converged quickly in 4 iterations. The log-likelihood values and McFadden's (1974) R^2 obtained for the last iteration are recorded.

Table 13.3 Comparison of the Linear Probability, Logit and Probit Models: Union Participation*

Variable	OLS	Logit	Probit
EXP	−.005 (1.14)	−.007 (1.15)	−.007 (1.21)
WKS	−.045 (5.21)	−.068 (5.05)	−.061 (5.16)
OCC	.795 (6.85)	1.036 (6.27)	.955 (6.28)
IND	.075 (0.79)	.114 (0.89)	.093 (0.76)
SOUTH	−.425 (4.27)	−.653 (4.33)	−.593 (4.26)
SMSA	.211 (2.20)	.280 (2.05)	.261 (2.03)
MS	.247 (1.55)	.378 (1.66)	.351 (1.62)
FEM	−.272 (1.37)	−.483 (1.58)	−.407 (1.47)
ED	−.040 (1.88)	−.057 (1.85)	−.057 (1.99)
BLK	.125 (0.71)	.222 (0.90)	.226 (0.99)
Const	1.740 (5.27)	2.738 (3.27)	2.517 (3.30)
Log-likelihood		−312.337	−313.380
McFadden's R^2		0.201	0.198
χ^2_{10}		157.2	155.1

* Figures in parentheses are t-statistics

Note that the logit and probit estimates yield similar results in magnitude, sign and significance. One would expect different results from the logit and probit only if there are several observations in the tails. The following variables were insignificant at the 5% level: EXP, IND, MS, FEM and BLK. The results show that union participation is less likely if the individual resides in the South and more likely if he or she resides in a standard metropolitan statistical area. Union participation is also less likely the more the weeks worked and the higher the years of education. Union participation is more likely for blue-collar than non blue-collar occupations. The linear probability model yields different estimates from the logit and probit results. OLS predicts two observations with $\widehat{y}_i > 1$, and 29 observations with $\widehat{y}_i < 0$. Table 13.5 gives the actual versus predicted values of union participation for the linear probability, logit and probit models. The percentage of correct predictions is 75% for the linear probability and probit model and 76% for the logit model.

One can test the significance of all slope coefficients by computing the LR based on the unrestricted log-likelihood value ($\log\ell_u$) reported in Table 13.3, and the restricted log-likelihood value including only the constant. The latter is the same for both the logit and probit models and is given by

$$\log\ell_r = n[\bar{y}\log\bar{y} + (1-\bar{y})\log(1-\bar{y})] \tag{13.33}$$

where $\bar{y}$ is the proportion of the sample with $y_i = 1$, see problem 2. In this example, $\bar{y} = 218/595 = 0.366$ and $n = 595$ with $\log\ell_r = -390.918$. Therefore, for the probit model,

$$LR = -2[\log\ell_r - \log\ell_u] = -2[-390.918 + 313.380] = 155.1$$

which is distributed as χ^2_{10} under the null of zero slope coefficients. This is highly significant and the null is rejected. Similarly, for the logit model this LR statistic is 157.2. For the linear probability model, the same null hypothesis of zero slope coefficients can be tested using a