possible functions (f) that impose this restriction, the following one was selected:[20]

$$P(S_i = 1 \mid Y_i) = e^{X_i\beta}/(1 + e^{X_i\beta}) \tag{4.21}$$

Because the probability that the ith firm does not work shifts is given by the logistic function, this approach is referred to in the literature as logit analysis (Theil 1971, p. 632). Equation (4.21) is nonlinear in its parameters, but the maximum-likelihood estimators of these parameters can be obtained by iterative methods, and either the likelihood ratio test or the t ratio test based on the estimates of the standard errors obtained from the estimated inverse of the information matrix can be employed to test hypotheses on the β coefficients.[21]

Our presentation of the results in this text will rely heavily (at times exclusively) on the estimates derived from the maximum-likelihood method. Nevertheless, all our models were also estimated by OLS. In our earlier work (Betancourt and Clague 1978) we found the two methods to yield similar results. This similarity was quite pronounced with respect to the signs and the statistical signficance of coefficients; however, sometimes these estimation methods differed appreciably with respect to predictive power. The wisdom of using both methods is illustrated by the conflicting evidence from the few Monte Carlo studies available. Those studies reported by McFadden (1974) and Domencich and McFadden (1975) tended to favor the maximum-likelihood method even for sample sizes as small as fifty, a typical size in our analysis; but the Goldfeld and Quandt (1972) experiments comparing OLS and probit analysis (which is quite similar to logit analysis) tended to favor OLS for sample sizes as large as 200.

*4.6 Measures of predictive performance for qualitative dependent variables

In the restricted-form specification, alternative hypotheses about the relevant theoretical factors lead to alternative regressions using a single independent variable. Therefore, the usual tests of significance on the coefficients of this independent variable are quite likely to fail to discriminate between alternative hypotheses because several of them may turn out to be significantly different from zero. Thus it is highly

[20] The main alternative available in the literature is probit analysis, but logit analysis was selected because of its simplicity and lower computational cost.

[21] A recent thorough discussion of the statistical properties of maximum-likelihood estimation of the binomial logit model is available (Domencich and McFadden 1975, pp. 110–12).

desirable to have a measure of predictive performance that indicates which of these hypotheses best explains the data. Traditionally, of course, the R^2 statistic is employed for these purposes. However, when the estimation method is nonlinear, the R^2 statistic, although it can be calculated, loses some of its desirable properties; for example, the usual decomposition of total variation is not applicable. Furthermore, there has been controversy regarding the appropriate upper bound of this statistic in the case of binary-choice models (e.g., Morrison 1972, Goldberger 1973).

Qualitative dependent-variable models, where the predicted values are probabilities, provide an opportunity to use alternative measures of predictive performance based on the concept of entropy [see the work of Theil (1971) for a discussion of entropy]. The actual measures used in this study are simple extensions of an earlier one developed by the authors (Betancourt and Clague 1978), and we now turn to a discussion of these measures.

The concept of entropy can be defined for any observation in a sample in terms of the predicted probabilities as

$$E_i = -\left(\sum_{j=1}^{k} \hat{P}_{ij} \log \hat{P}_{ij}\right) \tag{4.22}$$

where k is the number of alternatives. In the dichotomous case, equation (4.22) collapses to

$$E_i = -[\hat{P}_i \log \hat{P}_i + (1 - \hat{P}_i) \log(1 - \hat{P}_i)] \tag{4.23}$$

where, for example, $\hat{P}_i$ is the estimated probability that the ith firm will work shifts. Entropy, which may be interpreted as a measure of the amount of uncertainty associated with a distribution, takes on its maximum value at $\hat{P}_{ij} = 1/k$ for all j and its minimum value at $\hat{P}_{ij} = 1$ for any j. We seek a measure of the amount of information contained in the predicted probabilities $\hat{P}_{ij}$; this measure will be defined by

$$I_i = 1 - E_i/E_{\max} \tag{4.24}$$

where $E_{\max}$ is the maximum amount of entropy associated with the distribution.[22] For instance, $E_{\max}$ occurs at $\hat{P}_{ij} = \frac{1}{2}$ in the dichotomous case and at $\hat{P}_{ij} = \frac{1}{3}$ for three possible outcomes. I_i thus takes on its maximum value at $\hat{P}_{ij} = 1$ for any j and its minimum value at $\hat{P}_{ij} = 1/k$ for all j. In the dichotomous case, for example, the minimum occurs at $\hat{P}_i = \frac{1}{2}$.

Entropy has the attractive property that the joint entropy of two

[22] Although entropy is unique up to a factor of proportionality [set at unity in (4.22)], the measure of information in (4.24) will not be affected by the proportionality factor.

independent random variables is the sum of their individual entropies. The amount of information I_i has the same property. Therefore, defining a correct prediction as $\hat{P}_{ij} > 1/k$ when alternative j is chosen and $\hat{P}_{ij} < 1/k$ when it is not chosen,[23] we can define the amount of information contained in a set of predictions as

$$\bar{I} = (I_1 - I_2)/N \tag{4.25}$$

where I_1 is the sum of information for all the correct predictions, I_2 is the sum of misinformation for all the incorrect predictions, and N is the number of observations. $\bar{I}$ ranges from -1 (all probabilities incorrectly predicted as 1 or 0) to $+1$ (all probabilities correctly predicted as 1 or 0).

Intuitively, this summary measure scores each prediction in a given set by giving it points not only in accordance with whether the prediction is right (positive points) or wrong (negative points) but also in a way that reflects the degree of certainty of the prediction. In other words, more credit (discredit) should be and is given to a correct (incorrect) prediction if a high probability underlies the prediction than if a low probability underlies the prediction. For example, in the dichotomous case, more credit (discredit) is given to a correct (incorrect) prediction that is close to 1 or 0 than to a prediction that is close to 0.5.

The measure of information in (4.25) was applied to the dichotomous case in our earlier work (Betancourt and Clague 1978), and it has also been applied to the trichotomous case by Abusada (1975). Nevertheless, this measure has a shortcoming[24] that becomes particularly acute when the distribution of the observations over the choices is very uneven. $\bar{I}$ can be calculated for the predictions that result from bringing no information from the theory to bear on the data. That is, the proportions of observations in a sample that choose an alternative can be used as an estimate of $\hat{P}_i$ for every observation in the sample. If the observations are fairly evenly divided, the amount of information from this set of predictions for the whole sample will be relatively low; however, if the observations are very unevenly divided, the amount of information contained in these "naive" predictions can be relatively high. Because the focus of interest by most investigators is on how the introduction of the theory enhances the ability to explain and predict a

[23] When there are several alternatives, this definition of a correct prediction may be too strict. In these cases, a more suitable definition is simply $\hat{P}_{ij} > 1/k$ when alternative j is chosen (Lago 1979).

[24] The need for some normalization was brought to our attention by Tom Louis of Boston University.

particular phenomenon, we define a new measure of predictive performance, I_A, that captures the absolute amount of additional information provided by the introduction of the theory into the statistical analysis (i.e., in addition to the information already contained in the sample proportions). Thus

$$I_A = \bar{I} - \bar{I}_M \tag{4.26}$$

where $\bar{I}_M$ is the amount of information provided by the sample proportions. Because this measure will now have a different range for different samples, it is also desirable to calculate a measure of the amount of information provided by the introduction of the theory relative to the maximum amount of information that the theory can capture in a given sample.[25] Hence,

$$I_R = I_A/I_{A\max} \tag{4.27}$$

where $I_{A\max}$ is simply $1 - \bar{I}_M$. Finally, because the range of I_R is the same as the range of R^2, a degrees-of-freedom correction can be defined as follows:

$$\bar{I}_R = I_R - [K/(N - K - 1)](1 - I_R) \tag{4.28}$$

where K is the number of independent variables.[26]

The three measures defined by (4.26) through (4.28) provide a basis for evaluating predictive performance in models with qualitative dependent variables, and they will be used throughout this text. It is worth noting that in our initial application of $\bar{I}$ to the dichotomous case, as mentioned earlier, we found the behavior of $\bar{I}$ to be very similar to that of R^2, defined for the nonlinear case as $1 - \Sigma(S_i - \hat{S}_i)^2/\Sigma(S_i - \bar{S})^2$. A major advantage of all our four measures of information over R^2, however, is that they are equally applicable to situations with more than two alternatives; in contrast, R^2 cannot be defined for these cases. Moreover, we can use the relative measures of information, I_R and $\bar{I}_R$, to compare sets of predictions with different numbers of alternatives. This point will be illustrated in the next chapter.

[25] In principle, one can also define a measure relative to the minimum, but in practice a theory that predicts worse than the sample proportions is of little value.

[26] For the case involving multiple choices the definition will be

$$\bar{I}_R = I_R - [JK/(N - JK - J)](1 - I_R) \tag{4.29}$$

where J is the number of alternatives minus one.