Discussion Paper No. 37

PREDICTIONS FROM BINARY CHOICE MODELS

by

Richard B. Westin

February 24, 1973

1922 Sheridan Road
Northwestern University
Evanston, Illinois 60201

Predictions from Binary Choice Models

by

Richard B. Westin[1]

1. Introduction

An important problem in econometrics arises when cross-sectional information is used to estimate the probability that an individual or a firm will take a particular action in a binary choice situation. Various models have been proposed and used in such cases, including the linear probability model and the probit, logistic, and Gompertz models.[2] Most work to date has been concerned with estimation of the structural parameters of these models, but an important potential use of these models is to make predictions when the characteristics which determine the choices of individuals change. An unresolved problem for prediction arises, however, because binary choice models are designed to explain the behavior of individuals, while the predictions we are typically interested in relate to the behavior of aggregates of individuals rather than of particular individuals. Therefore, an important question in using binary choice models for predictions is how to aggregate predictions for individuals to give us predictions for the population. To give some examples, we may have estimated a model that explains when individuals are most likely to purchase a durable good, but we would like to use this model to predict the total change in durables purchased if the demographic structure of the population or the available finance rates change. Closely allied with the prediction problem is a model transferability problem; i.e., how can we use binary choice models estimated on one population to make predictions about other populations? To take a mode choice example, the effect of introducing a

commuter line into a new area can only be predicted by extrapolating models estimated in other areas where such a choice existed. In this case, however, the demographic and economic structure of the new area may differ from the area for which the model was estimated, and these differences should be included in the predictions. If binary choice models are to be useful in making predictions, it is important to investigate methods of summarizing how changes in individual characteristics will affect aggregate behavior for the population.

Suppose we have estimated a binary choice model,

$$\hat{p}_i = f(X_i) \tag{1}$$

where $\hat{p}_i$ is the estimated probability that the i^{th} individual will take a given action and X_i is a row vector of economic and demographic characteristics for the i^{th} individual. If we are interested in a particular individual, we can predict the probability that he will take the action being considered by calculating $\hat{p}_i$ for the vector of characteristics corresponding to that individual. If we are predicting for a population, however, we have a frequency distribution of characteristics over all individuals, yielding a frequency distribution of predicted probabilities for the population. In making aggregate predictions, it is important that this frequency distribution of predicted probabilities be incorporated into our predictions. To illustrate, we consider the logistic model of binary choice discussed in the next section. If there is a small change in the j^{th} characteristic controlling the binary choice of the i^{th} individual, the estimated change in the probability that the i^{th} individual will take the action being considered will be approximately

$$\Delta\hat{p}_i \approx \hat{\beta}_j \hat{p}_i (1-\hat{p}_i) \Delta X_{ij}, \tag{2}$$

where ΔX_{ij} is the change in the j^{th} characteristic for the i^{th} individual and $\hat{\beta}_j$ is an estimated constant. Now, even in the case where ΔX_{ij} is constant for all individuals (such as a constant across-the-board fare increase in a mode choice model), the predicted change in the probability of action for individuals will not be constant but will depend on each individual's original probability of action. This means that if we wish to use equation 2 to predict the expected change in the aggregate number of individuals taking the action being considered, we cannot merely extrapolate the estimated change in probability for a representative individual but instead must weight the prediction of equation 2 by the relative frequency distribution of the $\hat{p}_i$'s across the population. Because of this, our emphasis in this paper will be on the estimation and use of the relative frequency distribution of probabilities of action when making predictions from binary choice models of individual behavior.

In the remainder of this paper, we discuss the particular case of a logistic model of binary choice for making predictions, and we illustrate our results with an example from transportation planning. Section 2 of this paper considers the estimation of the relative frequency distribution of probabilities for the population. Section 3 considers the use of this distribution in making aggregate predictions, and Section 4 discusses another use of the distribution in testing models. Section 5 is a short summary.

2. The Relative Frequency Distribution of Probabilities

Since the prediction formulae we discuss in Sections 3 and 4 will be based on a fitted relative frequency distribution of probabilities for the population, it is reasonable to impose some criteria that the

relative frequency distribution should satisfy. In this section, we will be looking for a family of distributions that satisfies the following three criteria:

1. Flexibility. The family of distributions should be large enough to fit most reasonable relative frequency distributions of estimated probabilities.

2. Preservation. Since we will be making predictions of the effect of changes in the characteristics determining individual choices, we should have a family of distributions for individual probabilities that is preserved under reasonable changes in the underlying characteristics. This lets us base our prediction formulae on one family of distributions; and more importantly, it implies that the family of relative frequency distributions we choose for the probabilities will completely characterize the effects of changes in the characteristics determining individual choice.

3. Parameterization. Since we will be examining changes affecting the entire population of individuals, it will be convenient to represent changes in the characteristics determining individual choices in terms of changes in the moments of the distribution of characteristics. If the relative frequency distribution of individual probabilities can be parameterized in terms of the moments of the underlying distribution of characteristics determining individual choices, the change in the relative frequency distribution of probabilities resulting from a change in individual characteristics can be immediately determined by changes in the parameters of the frequency distribution of probabilities.

In order to find a family of distributions that satisfies our three criteria of flexibility, preservation, and parameterization, we will proceed by using our binary choice model to derive the relative frequency distribution of probabilities from an explicit assumption about the distribution of the characteristics determining individual choices.[3] In this section, we will consider explicitly the problem of finding a family of distributions to represent the probabilities generated by a logistic model of binary choice; the Appendix briefly discusses extending the results of this section to other models of binary choice behavior.

The logistic model of binary choice has been applied extensively in estimation of biological models and somewhat less frequently in economic applications. Estimation of the structural parameters of this model has been discussed by Berkson (1955), Walker and Duncan (1967), and Thiel (1970), among others. To define a logistic model, let y_i be a Bernoulli random variable representing the occurrence or non-occurrence of an action for the i^{th} individual, and let $E[y_i] = p_i$, where p_i is unobservable. We see that p_i is the a priori probability that the i^{th} individual will undertake the given action. A logistic model is defined by postulating that the natural logarithm of the odds that $y_i = 1$ is a linear function of the i^{th} individual's observed characteristics, or

$$\ln\left(\frac{p_i}{1-p_i}\right) = X_i\beta, \tag{3}$$

where β is a column vector of constants. The expression on the left hand side of equation 3 is called the logit. If we estimate β as $\hat{\beta}$, the estimated a priori probability that the i^{th} individual will undertake

the given action is

$$\hat{p}_i = (1 + \exp\{-X_i\hat{\beta}\})^{-1}. \tag{4}$$

The problem we are considering in this paper is how to infer the relative frequency distribution of the vector of population probabilities, P, if we know the relative frequency distribution of X, the matrix of population characteristics. The derivation of the distribution of P involves two transformations of variables, one from X to Xβ and one from Xβ to P. If we assume the distribution of Xβ is continuous and let g(Xβ) be its probability density function, then the density function of P is

$$f(p) = g\left[\ln\left(\frac{p}{1-p}\right)\right] \frac{1}{p(1-p)}, \quad 0 < p < 1. \tag{5}$$

To obtain analytic results, suppose X is distributed multivariate normal with the row vector of means μ_X and covariance matrix Σ.[4] In this case, Xβ is distributed univariate normal with mean $\mu = \mu_X\beta$ and variance $\sigma^2 = \beta'\Sigma\beta$. By equation 5, we see

$$f(p) = \frac{1}{\sqrt{2\pi}\sigma} \frac{1}{p(1-p)} \exp\left\{-\frac{1}{2\sigma^2}\left[\ln\left(\frac{p}{1-p}\right) - \mu\right]^2\right\}, \quad 0 < p < 1. \tag{6}$$

This equation defines the S_B family of probability density functions derived and discussed by Johnson (1949).

In terms of the three criteria we proposed earlier, we first note that the S_B family of densities is very flexible and can take on most reasonable unimodal or bimodal shapes on the interval (0,1). At the endpoints of the interval, 0 and 1, S_B densities equal zero and have one-sided derivatives of all orders equal to zero at these points. With regard to our second criteria, preservation, we see from our derivation that any transformation of the population characteristics that preserves

the normal distribution of X will yield a transformed distribution of population probabilities, P, that is again a member of the S_B family. Therefore, if we wish to examine changes in the distribution of population probabilities resulting from normality-preserving changes in the distribution of individual characteristics, these effects can be completely characterized by changes in the member of the S_B family of distributions derived from the assumed distribution of individual characteristics.

Finally, with regard to parameterization, we see that the parameters of the S_B distribution are μ and σ^2, which are, respectively, a linear combination of the means of X and a linear combination of the variances and covariances of X. Therefore, any transformation of the individual characteristics that can be expressed in terms of these moments of X (which are also the only transformations that preserve the normality of X) can be immediately summarized in terms of its effect on the parameters of the distribution of population probabilities, P. To illustrate, if the distribution of P is represented by an S_B distribution with parameters (μ,σ^2), an additive change in the k^{th} component of X, say X_k+a, will yield a transformed S_B distribution for P with parameters $(\mu+\beta_k a,\sigma^2)$. Similarly, a multiplicative change in the k^{th} component of X, say $(1+a)X_k$, will yield a transformed S_B distribution for P with parameters $(\mu+a\beta_k\mu_k,\sigma^2+\beta_k^2(2a+a^2)\sigma_k^2+2\sum_{i\neq k}\beta_i\beta_k a\sigma_{ik})$, where μ_k, σ_k^2, and σ_{ik} are the mean, variance, and covariance of the corresponding components of X. Since aggregate changes in X will be expressed as changes in the moments of X, this parameterization will be very convenient when we discuss changes in the distribution of population probabilities arising from changes in the distribution of individual characteristics. We do caution,

however, that μ and σ^2 should not be identified with the mean and variance of the S_B distribution; the moments of the S_B family of distributions can be shown to be very complicated functions of these parameters.

Because the vector of probabilities, P, for our sample is unobservable, direct maximum likelihood estimation of the parameters of the S_B distribution is intractible.[5] Instead, we resort to an indirect method of estimation that is quite straightforward. In the first step, we obtain consistent estimators, $\hat{\mu}_X$ and $\hat{\Sigma}$, of μ_X and Σ from the corresponding sample moments of our observations, X. We can also obtain consistent estimates, $\hat{\beta}$, of β in equation 3 by a number of different methods. In the second step, we estimate μ and σ^2 as $\hat{\mu} = \hat{\mu}_X\hat{\beta}$ and $\hat{\sigma}^2 = \hat{\beta}'\hat{\Sigma}\hat{\beta}$, respectively. By Slutsky's theorem (see e.g., Goldberger (1964)), $\hat{\mu}$ and $\hat{\sigma}^2$ are then consistent estimators of the corresponding parameters of the S_B distribution.[6]

It should be observed that the problem we are considering in this paper implies an extra restriction that must be placed on the required data set. The parameters of the S_B distribution are dependent on both the structural parameters, β, of the binary choice model and on the unknown moments of X. This means that we require observations on individual characteristics and outcomes with the additional requirement that the population of individuals be randomly sampled; this last restriction is unnecessary if we only wish to estimate the parameters of the binary choice model, equation 3. Since it is typical, however, that data collected for binary choice models is based on a random sample of individuals, this requirement is not unduly restrictive for applied work.

Our discussion to this point has been based on the assumption that the distribution of X is distributed multivariate normal. Although this assumption may often be reasonable in applied work, other distributions

of X are of course possible and would yield other distributions for P. In the extreme case where X is discrete, as for example when X consists of all dummy variables, the distribution of P is also discrete and has positive probability for at most as many points as there are distinct combinations of the possible values of X. In this case, our integral prediction equations in the next two sections would be re-expressed as sums. A more reasonable case, however, may be to assume X is multivariate normal except for one or two dummy variable components. If X_{ik} is a dummy variable, equation 3 implies the effect of the dummy variable is to shift the logit by β_k when $X_{ik} = 1$. In the special case where the dummy variables and the other individual characteristics can be assumed to be stochastically independent, the distribution of population characteristics can then be represented by separate S_B density functions for each combination of values of the dummy variables such that the S_B densities differ only in simple shifts of the parameter μ. In general, however, since the parameters of the S_B distribution depends on all the moments of the distribution of individual characteristics, if we cannot assume stochastic independence between the dummy variables and the other characteristics, we would have to fit separate S_B density functions to the estimated moments of X stratified by distinct combinations of the dummy variables to allow for variations in the distribution of individual characteristics. Our integral prediction equations in the next section would then be replaced by weighted sums of integrals, where the weights would be taken to be the relative frequency of each combination of values of the dummy variables in the data set.

To summarize our discussion so far, we showed in the Introduction that aggregate predictions derived from binary choice models must

incorporate the relative frequency distribution of probabilities for individuals in the population. We have now shown that S_B density functions provide a method of summarizing this relative frequency distribution that incorporates certain a priori criteria. We turn now to the object of our paper, which is the use of our fitted S_B frequency distribution in making aggregate predictions.

3. Aggregate Predictions

Although changes in the distribution of individual characteristics can be represented by changes in the relative frequency distribution of probabilities for the population, the major interest in prediction problems usually will not be on the distribution of unobservable probabilities but rather on the distribution of outcomes of the binary choice problem faced by each individual in the population. Since the outcome of the binary choice problem for each individual is a Bernoulli random variable based on his individual probability, p_i, the distribution of outcomes for the population must be inferred from the relative frequency distribution of individual probabilities. Rather than examining the entire probability distribution of outcomes, however, we usually will find it sufficient to consider particular moments of this distribution, which can then be inferred directly from various moments of $f(p)$.

In particular, in binary choice problems, the most important prediction problem will involve estimating the expected proportion of individuals in the population who will undertake the action being considered (i.e., the expected proportion of individuals that will buy a durable, or the expected proportion who will take the train rather than drive their

car). Since each individual, i, will undertake the action being considered with probability p_i, the expected proportion of all individuals who will undertake the action can be found by aggregating individual p_i's by their relative frequency of occurrence in the population to get $E[p] = \int_0^1 pf(p)dp$, where $f(p)$ is the relative frequency distribution of population probabilities.

When the characteristics that determine the choices of individuals change, the natural question then would be how does that change affect the expected proportion of individuals taking the action being considered, which we see is equivalent to asking how the change in individual characteristics affects the mean, $E[p]$, of the distribution of population probabilities. In particular applications, the change in $E[p]$ would depend on the particular change in individual characteristics being considered; but in this section, we can indicate the way $f(p)$ might be used for making predictions in typical examples.

A convenient method of summarizing the sensitivity of $E[p]$ to small changes in the distribution of the characteristics determining choices is to compute elasticities of $E[p]$ with respect to the moments of X. For example, the elasticity of the expected proportion of individuals who will take the action being considered with respect to a change in the mean of the k^{th} component of X is:

$$\epsilon_{\mu_k} = \left[\frac{\partial E[p]}{\partial \mu} \frac{\partial \mu}{\partial \mu_k}\right] \frac{\mu_k}{E[p]} = \frac{\beta_k \mu_k E[p(1-p)]}{E[p]} \tag{7}$$

Similarly, a change in X that affects the variance of $X\beta$ will have elasticity[7]

$$\epsilon_c = \left[\frac{\partial E[p]}{\partial c}\right] \frac{c}{E[p]} = \frac{E\left[p(1-p)\left(\ln\left(\frac{p}{1-p}\right) - \mu\right)\right]}{E[p]} \tag{8}$$

Elasticities are useful for indicating the effect of small changes in moments of X; but if large changes in the moments of X are proposed, first order approximations may be imprecise. In this case, if the magnitude of the proposed change in the moments of X is known, the new parameters of the transformed S_B distribution of probabilities can be computed from the definition of μ and σ^2; and the means of the two distributions can be compared directly.

To illustrate the use of the procedures discussed in this paper, we will derive aggregate predictions for an actual mode split study. The data used in this example are 878 observations on the binary choice between train (y_i=1) and car (y_i=0) for travelers taking business trips in the Edinburgh-Glasgow area of Scotland.[8] The particular model we used for this paper is:

$$\ln \frac{\hat{p}_i}{1-\hat{p}_i} = \underset{}{1.644} - \underset{(0.216)}{0.198}X_{i1} - \underset{(0.040)}{0.019}X_{i2} - \underset{(0.005)}{1.524}X_{i3} - \underset{(0.290)}{0.0833}X_{i4} \qquad (9)$$
$$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (0.123)$$

where

$\hat{p}_i$ = estimated probability that the i^{th} traveler will take the train,

X_1 = difference in "journey units" between train and car,

X_2 = walking and waiting time for the train trip,

X_3 = relative difference in time required for trip by each mode,

X_4 = relative difference in cost of trip by each mode.

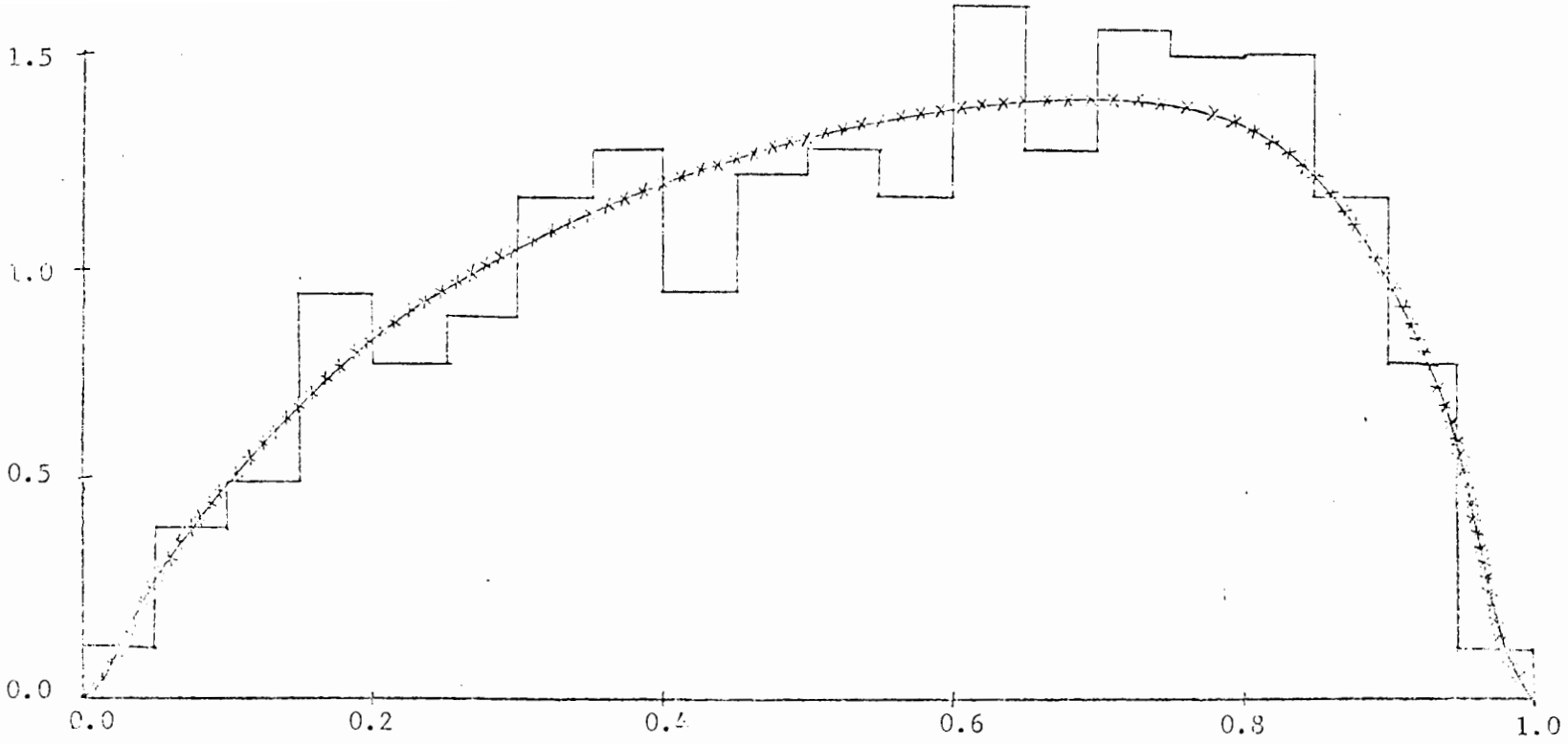
This model was fitted by the method of Walker and Duncan (1967).

In order to fit an S_B function to represent the population frequency distribution of probabilities, the procedure we used was to determine the sample mean and sample variance of the estimated logits $\left(\text{i.e. } \ln\left[\frac{\hat{p}_i}{1-\hat{p}_i}\right]\right)$ for

the data used to fit equation 9. For our sample, we found $\hat{\mu} = 0.248$ and and $\hat{\sigma}^2 = 1.435$, which we took as the estimates of the parameters μ and σ^2, respectively, for the S_B function. This fitted distribution is illustrated in Figure 1. The calculation of the prediction formulas derived in this section then required a number of moments of the S_B function. Since the moments of the S_B function are extremely intractable (see Johnson (1949)), we resorted to numerical integration techniques to calculate the prediction formulas for our fitted S_B function.

We first consider estimates of the elasticities of the proportion of business travelers taking the train with respect to changes in the moments of the independent variables. These elasticities are given in Table 1. To interpret them, consider ϵ_{μ_2}, the elasticity of E[p] with respect to the mean of walking-waiting time. The estimated value of $\epsilon_{\mu_2} = -0.17$, which implies that if walking-waiting time for travelers is increased by one percent of the current mean value, expected total train ridership will fall by 0.17%.

We have stressed that the value of the approach to aggregate predictions outlined in this paper is that it incorporates the entire relative frequency distribution of individual probabilities into the prediction process. It is interesting to note at this point what the effect on our results would be if we neglected the relative frequency distribution and extrapolated our results from the behavior of a representative individual. If we follow the common procedure of working with the mean individual, we would estimate the elasticity of his probability of taking the train with respect to a change in the characteristics determining his choice as $\epsilon^*_{x_k} = [\beta_k \mu_k (1-E[p])]/E[p]$. If we take this



——— : Relative frequency distribution of sample $\hat{p}_i$'s

-x-x-x-x- : Fitted S_B relative frequency distribution for $\hat{\mu} = 0.248$, $\hat{\sigma}^2 = 1.435$

Figure 1

ε_{μ_1}	-0.30
ε_{μ_2}	-0.17
ε_{μ_3}	-0.14
ε_{μ_4}	-0.12
ε_{σ}	-0.01

Table 1

Estimated elasticities of E[p] with respect to the moments of X.

elasticity as representative of all individuals, the difference between the estimated aggregate elasticities computed both ways will be determined by the ratio of E[p](1-E[p]) to E[p(1-p)]. For the data used to estimate equation 9, this ratio is 1.285, implying that elasticities computed by extrapolating the behavior of the mean individual will over-estimate the aggregate elasticities computed by equation 7 by 28.5% in absolute terms.[9] As an example, the elasticity of the proportion of business travelers taking the train with respect to a change in walking-waiting time is computed to be -0.22 if we extrapolate the behavior of the mean individual, a significant overestimation relative to the correct answer of -0.17 based on the entire frequency distribution of probabilities.

Although elasticities are useful for predictions of the effects of small changes in the distribution of individual characteristics, an alternative procedure if the magnitude of the changes being considered is known is to examine the mean of the transformed S_B distribution of probabilities directly. To illustrate, we consider both a five-minute decrease in walking-waiting time and a ten minute decrease in the travel time by train. The change in walking-waiting time is simpler to consider as it involves only a change in μ; but since the effect of travel time is modeled in equation 9 as depending on the difference in journey time between the two modes relative to the average of the times, the change in train travel time will affect both the mean, variance, and covariances of the population characteristics and therefore affect both μ and σ.[10]

For the five minute decrease in walking-waiting time, the parameters

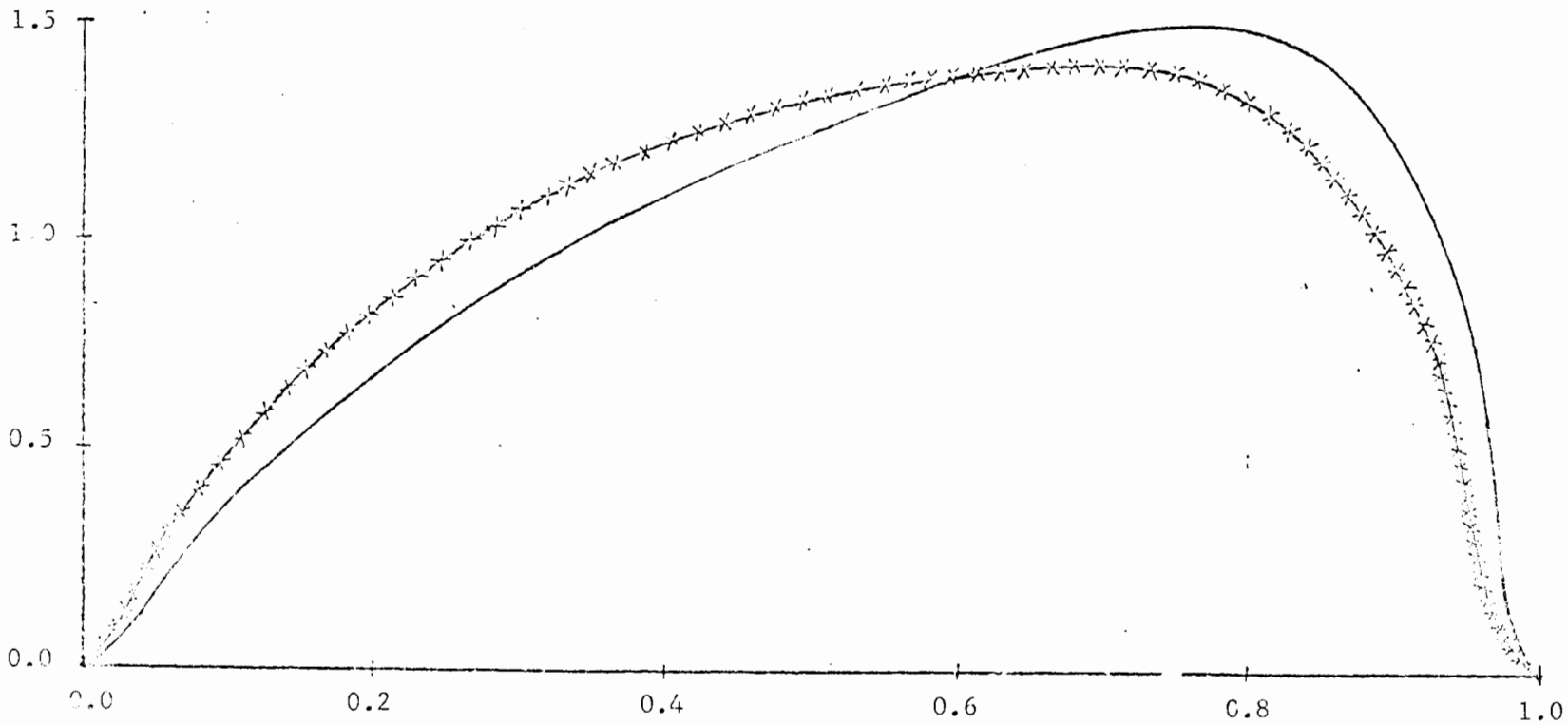
of the transformed S_B density function are $(\alpha+\beta_2 \cdot 5, \sigma^2)$ which we estimate as (0.343, 1.435). This change in the population characteristics changes E[p] from 0.5481 to 0.5664, implying an expected change of 1.83% in percentage mode split, or alternatively an expected increase of 3.34% in the expected number of business travelers taking the train. To compare this result with the result based on the first-order approximation of the elasticity, we note that a five minute decrease in walking-waiting time is a 19.3% decrease in mean walking-waiting time, implying a 3.35% increase in the expected number of business travelers taking the train if we would extrapolate the elasticity, which gives only a very slight over-estimation.

To examine the effect of a ten-minute decrease in the time required for the train trip, we need the estimated mean and covariance matrices of the population characteristics after the indicated change. Letting $\hat{\hat{\mu}}_X$ and $\hat{\hat{\Sigma}}$ be the estimated mean and covariance matrix of X after the decrease in train-travel time, we want to compare the means of two S_B distributions with estimated parameters $(\hat{\mu}_X\hat{\beta}, \hat{\beta}'\hat{\Sigma}\hat{\beta})$ before the change and estimated parameters $(\hat{\hat{\mu}}_X\hat{\beta}, \hat{\beta}'\hat{\hat{\Sigma}}\hat{\beta})$ after the change. These two density functions are compared in Figure 2. Because of the decrease in train travel time, E[p] increases from 0.548 to 0.577, implying an increase of 2.9% in expected percentage mode split, for an increase in expected train ridership by business travelers of 5.2%.

Summarizing this section, we have shown how to use the relative frequency distribution of probabilities discussed in Section 2 to aggregate individual binary choice behavior into a prediction for the aggregate population of individuals. We turn now to another use of the S_B frequency distribution of probabilities, which is in evaluating the forecasting performance of our models.

4. Forecast Performance Evaluation

In addition to aggregate predictions, there are also cases where



: S_B relative frequency distribution for probabilities of choosing the train before decrease in train trip-time

: S_B relative frequency distribution for probabilities of choosing the train after decrease in train trip-time

Figure 2

we would like to evaluate our performance at predicting the outcome of individual events over time. Models where this evaluation is of interest are competitive bidding models, weather forecasting models, and other models where predictions of a sequence of individual events is important. We would also like to use these measures as predictive tests to test our model's representation of the determination of individual probabilities. These tests are particularly valuable if we wish to use our model to make predictions for other populations, and we have a small set of data to use to test whether the model is transferable.

In a recent paper, Morrison (1972) has discussed the problem of using observed outcomes to evaluate the performance of probability predictors. In particular, Morrison shows that if we consider the sample coefficient of determination (R^2) between outcomes and predicted probabilities, the expected value of R^2 is maximum when the predicted probabilities are chosen equal to the true probabilities of the events. Furthermore, if we predict an outcome as 0 or 1 depending on whether the predicted probability of the event is less than or greater than 0.5, Morrison derives the expected proportion of correct predictions we will make if we base our predictions of the outcomes on the true probabilities of the events. Analytic results are given for these measures by Morrison under the assumption that the true distribution of probabilities is known and is distributed beta. If the frequency distribution of probabilities is generated from a logistic binary choice model, however, we have shown that the assumption of a beta distribution for the probabilities is unattractive and a more plausible assumption is to use an S_B distribution. Following Morrison, then, but assuming the distribution of probabilities is distributed as S_B, the maximum expected value of R^2 that can be obtained

by correlating binary outcomes and probabilistic predictions is

$$\tilde{R}^2 = \frac{E[p^2] - E^2[p]}{E[p] - E^2[p]} . \tag{10}$$

Furthermore, if we predict y_i as 0 or 1 depending on whether the predicted probability of the event is less than or greater than .5, we find:

$$P(C) = \int_0^{.5} (1-p)f(p)dp + \int_{.5}^{1} pf(p)dp, \tag{11}$$

where P(C) is the maximum expected proportion of correct classifications of outcomes. If we calculate these measures based on an S_B distribution fitted by the method of Section 2, we would have a base for evaluating our forecasting performance on new data under the assumption that our fitted distribution correctly represents the true distribution of probabilities.

Unfortunately, $\tilde{R}^2$ and P(C) are only expressions for the expected value of the coefficient of determination and the expected proportion of correct predictions of outcomes, assuming we calculate them using the true distribution of probabilities.[12] In order to use them as effective tests on our model's performance, we also need the sampling distribution of these statistics. Although the sampling distribution of R^2 is difficult to obtain in this model, the sampling distribution of the observed proportion of correct predictions of outcomes is easy to obtain and can be used to evaluate our predictions. In particular, if we define a new random variable as z_i=1 if we predict an outcome correctly and z_i=0 if we predict incorrectly, z_i is distributed Bernoulli with $E[z_i] = P(C)$ under the assumption that the true distribution of probabilities is used in equation 11 to calculate P(C). Under this assumption, the actual number of correct predictions in n events is distributed binomial with parameter P(C), and the sampling distribution of the proportion of correct

predictions (assuming large n) is approximately normal with mean P(C) and variance P(C)[1-P(C)]/n. A 95% confidence interval for the proportion of correct predictions of outcomes is then $P(C) \pm 1.96\sqrt{P(C)[1-P(C)]/n}$. An important attribute of this confidence interval for testing our model is that it embodies within it the assumption that the S_B distribution fitted to our data by the method of Section 2 is equal to the true distribution of probabilities for the data we predict on. Therefore, if the actual proportion of correct predictions on new data lies outside this confidence interval, it would be an indication that our fitted S_B distribution does not adequately represent the true distribution of probabilities for the population, either because of sampling variability in the fitting of the S_B distribution or because our model of binary choice or our assumption that X is normally distributed is not adequate to describe the data.[13] Since any of these occurrences is sufficient to cast doubt on the aggregate predictions derived in Section 3, this test would be very useful as a check on the validity of our assumptions.

To illustrate the use of the results of this section, we randomly split our data set in the approximate proportions of 4:1, refitted equation 9 on the larger set of data, and used the reestimated equation to predict the smaller set. Our new estimated equation, based on 692 observations randomly selected from our data set, was:

$$\ln\frac{\hat{P}_i}{1-\hat{P}_i} = \underset{(0.235)}{1.506} - \underset{(0.044)}{0.196}X_{i1} - \underset{(0.005)}{0.015}X_{i2} - \underset{(0.307)}{1.330}X_{i3} - \underset{(0.139)}{0.842}X_{i4} \qquad (12)$$

This equation implied an S_B density function with estimated parameters (.306,1.247). Using equations 10 and 11 for the S_B distribution fitted to this data, we calculated $\tilde{R}^2 = 0.201$, P(C) = 0.700, and the 95%

confidence interval for the number of correct predictions was (0.634, 0.766). To test these measures, we used equation 12 to calculate $\hat{p}_i$ for 186 observations not used to fit equation 12 and found an actual sample R^2 of 0.286 and a correct prediction percentage of 0.731. Although both the sample statistics exceeded the maximum expected values of these statistics based on the assumption that the fitted S_B distribution equals the true distribution of probabilities for the sample, the actual proportion of correct predictions lies within the 95% confidence interval for this prediction.[14] We would therefore tentatively accept the fitted S_B distribution as an adequate description of the true relative frequency distribution of probabilities necessary for our prediction formulae in Section 3.

5. Summary

In this paper, we have considered the problem of using models of individual binary choice behavior to make predictions for populations. Since binary choice models applied to individuals are an efficient method of determining what characteristics determine individual choices, it is important to be able to use this information to predict what will happen to aggregate behavior when the distribution of characteristics determining individual choices changes. We have shown that for a specific model of binary choice, changes in the distribution of individual characteristics can be characterized by their effect on the relative frequency distribution of probabilities of action for the population; and this relative frequency distribution can then be used to make predictions on aggregate behavior. In addition, we have formulated predictive tests to check on the accuracy of our fitted relative frequency distribution.

There are, of course, unfinished questions in this paper. Of particular importance is the question of the sampling distribution of the parameter estimates of the S_B distribution described in Section 2 and the effects of sampling variability on the predictions in Section 3. Another question we have not addressed is the extension of our procedure to choices involving more than two alternatives. Nevertheless, we feel that the procedure discussed in this paper can be valuably applied in many situations to extend the usefulness of binary choice models from individual behavior to describing aggregate behavior.

Appendix

Section 2 of this paper derives prediction formulas for a logistic binary choice model. This appendix briefly considers the extension of the formulas in Section 2 to other models of binary choice.

1. Linear Probability Function

The linear probability function is defined by assuming

$$p_i = X_i\beta \tag{A1}$$

As is well known, the probabilities estimated from fitting A1 may fail to lie in the interval $[0,1]$. In the context of this paper, this means that if we begin by assuming an explicit distribution of X, our possible choices are very limited because of the need that at least the theoretical distribution of P be bounded. Furthermore, if we transform X to make predictions, we would generally expect the distribution of P to be shifted so that it violated the bounds of $[0,1]$. The alternative of truncating the distribution of P appears artificial and was not pursued. Another possibility involves fitting a bounded distribution such as a beta function directly to the p_i, but this distribution will not in general be preserved under transformations of X. We therefore feel the specification of A1 is incompatible with the methods of this paper.

2. Probit Model

The probit model is defined by assuming:

$$p_i = \int_{-\infty}^{X_i\beta} \frac{1}{\sqrt{2\pi}} \exp\left\{-\tfrac{1}{2}u^2\right\} du = F(X_i\beta), \tag{A2}$$

where F is the cumulative distribution function of the standard normal dis-

tribution. If we let $h(X\beta)$ be the probability density function of $X\beta$, the probability density function of p is

$$f(p) = h(F^{-1}(p)) \left| \frac{d}{dp} F^{-1}(p) \right| . \tag{A3}$$

Since equation A3 involves the inverse cumulative normal distribution function, it has no closed form. The function f(p) can be numerically approximated, but in general the moments required in the prediction formulas will be difficult to obtain. Note that the assumption used in Section 2 that X was normally distributed has no connection with the argument usually used to justify equation A2. Equation A2 is typically justified by assuming the existence of an unobservable index that is normally distributed, but it requires no assumption on the distribution of X.

3. Gompit Model

The Gompit model is defined by assuming:

$$\ln \ln \left(\frac{1}{p_i}\right) = X_i\beta \tag{A4}$$

If we assume X is distributed multivariate normal, the density function of P is

$$f(p) = \frac{1}{\sqrt{2\pi}\,\sigma} \frac{1}{p \ln\left(\frac{1}{p}\right)} \exp\left\{-\frac{1}{2\sigma^2}\left(\left(\ln \ln \frac{1}{p}\right) - \mu\right)^2\right\} \tag{A5}$$

$$0 < p < 1.$$

To evaluate the moments of f(p), we note that if we make the change of variable $\nu = \ln\left(\frac{1}{p}\right)$, the probability density function of ν is lognormal with parameters μ and σ^2. The k^{th} moment about the origin of f(p) is then equal to the value

of the moment generating function of ν evaluated with the indicator variable of the moment generating function equal to -k. A brief check of the literature, however, failed to find an analytic expression for the moment generating function of the lognormal distribution.

Footnotes

[1]Assistant Professor at Northwestern University. I would like to thank Peter Watson for loaning me the data used in this paper and for discussing the ideas with me. Susan Westin and Thomas Cooley also provided helpful discussion, but all responsibility for error is of course mine. An earlier version of this paper was presented at the meetings of the Econometric Society in Toronto, Canada, on December 30, 1972.

[2]These models are surveyed in Zellner and Lee (1965).

[3]The alternative of directly specifying a family of distributions to represent the relative frequency distribution of probabilities will not in general yield a family of distributions that satisfies the three criteria we have specified. For example, since we are looking for a family of distributions to represent a relative frequency distribution of probabilities, an attractive possibility is the family of beta distributions. It is possible to show, however, that if probabilities are generated by a logistic model of binary choice, then even a simple additive shift in one of the characteristics determining individual choices would produce a frequency distribution of probabilities that is not distributed beta, violating the second criterion of preservation. In addition, the parameters of the beta distribution are not easily expressed in terms of the moments of the underlying characteristics, violating the third criterion; and finally, for certain ranges of the parameters of the beta family of distributions, a well-defined distribution of characteristics may not even exist.

[4]In most applications, the data matrix will include a constant column so that one component of X will have a degenerate distribution. This will not affect the results of this paper, and we have suppressed the constant term for notational convenience.

[5]One can form a likelihood function independent of the unobservable p_i's by remembering that the conditional distribution of y_i, given p_i, is Bernoulli with parameter p_i. The product of this distribution and equation 6 is then the joint density function of y_i and p_i. By integrating this joint density with respect to p_i, one gets the marginal density function of the outcomes as distributed Bernoulli with parameter equal to the mean of equation 6. The mean of equation 6, however, is an extremely intractible function of μ and σ^2, so this approach is not attractive.

[6]In actuality, the estimates of μ and σ^2 reported in this paper were found by a slightly different procedure. Since X is assumed multivariate normal, equation 3 implies the logits are distributed univariate normal with mean μ

and variance σ^2. Therefore, if we have a random sample of the population characteristics and we estimate equation 3 by standard methods on all our data, the sample mean and variance of the estimated logits, $\ln\left(\frac{\hat{p}_i}{1-\hat{p}_i}\right) = X_i\hat{\beta}$, will be estimators of μ and σ^2, respectively. It can be shown by Slutsky's Theorem that these estimators are consistent also.

[7] In general, a change that affects the variance of one of the components of X will also affect the covariances between the components of X, so this elasticity cannot be expressed simply in terms of particular moments of X.

[8] For a fuller description of this study, see Watson (1972). This study was financed by the U.K. Ministry of Transport (now Department of the Environment).

[9] Since p(1-p) is a concave function, Jensen's inequality implies that

$$E\left[p(1-p)\right] \leq E(p)\left[1-E(p)\right].$$

[10] Strictly speaking, the change considered will not preserve the normality of the distribution of relative travel times. Since the emphasis in this paper is on illustrating the use of the prediction methods, we numerically approximated the effect of a ten minute decrease in train travel time on the estimated mean and variance of relative travel time and on the estimated covariances. Since these are the moments that would be relevant to changes that preserved the normality of the population characteristics, it was felt that this procedure was sufficient to yield good predictions.

[11] A natural question at this point would be to ask for a confidence interval for this prediction and the previous predictions. There are two types of randomness that we would have to account for in this confidence interval. The first type arises because the choices of individuals are random variables, so that even if we knew the relative frequency distribution of probabilities for the population, the actual percentage mode split is still a random variable. However, since we are estimating the proportion of business travelers who will take the train rather than the absolute number, the variance in our estimate because of random behavior of individuals will be small if the population is large. On the other hand, we would also have random variation in our prediction because of sampling variation in the estimates of the parameters of the S_B distribution. Although this is liable to be important, we have not been able to treat this problem because our two step estimation does not yield a tractable sampling distribution for our estimators; and even if we could determine the sampling distribution of $\hat{\mu}$ and $\hat{\sigma}^2$, the integral prediction equations of this section are very complicated functions of these estimators.

[12] Goldberger (1973) has particularly stressed this fact and notes that Morrison's description of $\tilde{R}^2$ as an "upper bound" is misleading, since the only absolute upper bound on the sample R^2 is one.

[13] This test might be criticized on the basis that sufficient observations in the prediction set will certainly reject the S_B distribution as inadequate because of sampling variability in the original fitting. However, with a large number of observations, some observations can be used to improve the fit of the S_B distribution, so we feel this objection is not crucial.

[14] In other experiments not reported here, it was found using this test that the fitting of the S_B distribution is sensitive to data outliers in the sample. Care in checking over the data set is therefore urged, or alternatively one might wish to use estimators that are less sensitive to data outliers than the ones used in Section 2.

References

Berkson, J., 1955, "Maximum Likelihood and Minimum X^2 Estimates of the Logistic Function", Journal of the American Statistical Association, 50, 103-162.

Goldberger, A., 1964, Econometric Theory, (John Wiley and Sons, Inc., New York), p. 119.

Goldberger, A., 1973, "Correlations Between Binary Outcomes and Probabilistic Predictions", forthcoming in Journal of the American Statistical Association.

Johnson, N., 1949, "Systems of Frequency Curves Generated by Methods of Translation", Biometrika, 36, 149-176.

Morrison, D., 1972, "Upper Bounds For Correlations Between Binary Outcomes and Probabilistic Predictions", Journal of the American Statistical Association, 67, 68-70.

Thiel, H., 1970, "On the Estimation of Relationships Involving Qualitative Variables", American Journal of Sociology.

Walker, S. and D. Duncan, 1967, "Estimation of the Probability of an Event as as Function of Several Independent Variables", Biometrika, 54, 167-179.

Watson, P., 1972, The Value of Time and Behavioural Models of Modal Choice, unpublished Ph.D. dissertation, University of Edinburgh.

Zellner, A. and T. Lee, 1965, "Joint Estimation of Relationships Involving Discrete Random Variables", Econometrica, 33, 382-394.