This is to certify that the

dissertation entitled

THE VALUE OF IMPERFECT SAMPLE SEPARATION INFORMATION IN SWITCHING REGRESSION MODELS

presented by

Edwina A. Masson

has been accepted towards fulfillment
of the requirements for

Ph.D. degree in Economics

Peter Schmidt

Major professor

Peter J. Schmidt

Date July 3, 1985

THE VALUE OF IMPERFECT SAMPLE SEPARATION INFORMATION IN SWITCHING REGRESSION MODELS

By

Edwina A. Masson

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Economics

1985

ABSTRACT

THE VALUE OF IMPERFECT SAMPLE SEPARATION INFORMATION IN SWITCHING REGRESSION MODELS

By

Edwina A. Masson

The purpose of this study is to determine the value, in terms of efficiency gains, of using imperfect sample separation information in switching regression models. The imperfect information appears in the model as a regime classification, which is correct only with some probability. The importance of this study lies in the fact that knowledge of improvements in the efficiency of parameter estimation can guide one in determining whether to use sample separation information, even if it is unreliable.

We determine the value of sample separation information by comparing the asymptotic variances of the parameter estimates, under different assumptions about the available information. These assumptions range from perfect sample separation information, at the one extreme, to no such information whatever, at the other extreme. The asymptotic variances of the parameter estimates are obtained from the relevant information matrices, which are calculated by simulation over a very large sample size.

Among our findings, the following are most important. (1) There are efficiency gains when using imperfect information as compared to no information at all, and these can be

substantial in some cases. (2) Efficiency gains when using imperfect sample separation information are greatest when such information is highly reliable; and when the samples are difficult to disentangle from each other. (3) There are additional efficiency gains when the switching probabilities are modelled as probit functions of the explanatory variables. These gains occur in cases when they are most needed; specifically, when the samples are hardly distinct from each other, and when the imperfect sample separation information is not very informative.

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my adviser, Professor Peter Schmidt, for all the guidance and encouragement he gave me throughout the course of this thesis. He was always available with helpful suggestions and was very patient with me, particularly when the thesis problem had not yet been explicitly defined. I am also grateful to the other members of my dissertation committee -- Professors Christine Amsler, T.C. Anant, and Stephen Martin.

Most of all, I want to thank my family, especially my parents and my husband, for their support and encouragement during my years of study at Michigan State.

TABLE OF CONTENTS

LIST OF TABLES

Tables on Ratios of Asymptotic Variances

CHAPTER ONE

INTRODUCTION

1.1 Definition of the Problem

Switching regression models, normal mixture models, and disequilibrium models are systems characterized by discontinuous shifts in regression regimes at unknown points in the data series. The most common formulation hypothesizes that the system may switch numerous times back and forth between two particular regimes, or to successive new regimes. For the sake of simplicity, we shall restrict our discussion to the case in which it is known a priori that the number of regimes is two. These models are primarily designed to deal with samples in which sample separation information is missing. That is, we do not know whether an observed random variable is generated by one regime (which corresponds to a distinct regression model) or by another regime (which corresponds to another regression model).

An interesting issue here is the loss, measured in terms of the efficiency of parameter estimation, when sample separation (alternatively, regime classification) is unknown or is not observed. A number of papers (Goldfeld and Quandt, 1975; Kiefer, 1978; Schmidt, 1981) have addressed this question in the context of disequilibrium models and normal mixture models. All these studies found that sample separation information does have a positive value, in that estimates derived are more efficient when there is a priori knowledge

as to which regime each observation belongs to. This confirms the need to obtain reliable information about sample separation, when it is available.

The purpose of this paper is to extend the issue one more step. Sample separation information may exist, but may not be entirely reliable. Such a situation may conceivably arise in models with outliers, or when the available data is simply not entirely accurate. By how much is efficiency improved when imperfect regime classification information is used? This paper attempts to answer that question, and is therefore, an extension of Schmidt's paper, with the additional use of imperfect sample separation information. We will address the issue strictly in the context of switching regression models.

The importance of this extension lies in the fact that knowledge of improvements in efficiency of parameter estimation can guide one in deciding whether to use sample separation information, even if it is known that such information is imperfect or unreliable. In addition, even if imperfect information is not readily available, knowledge of efficiency gains will aid in determining whether such additional information is worth obtaining at all.

Before we proceed any further, a formal discussion of switching regression models is warranted at this point.

1.2 Formal Discussion of Switching Regression Models

The simplest possible formulation is a normal mixture model (actually, a switching regression model with only a constant term), where a sample of observations $y_1, y_2, \ldots, y_n$ is given on a random variable y. It is known that nature chooses between regimes with probabilities λ and $1 - \lambda$. That is,

$$y \sim N(\mu_1, \sigma_1^2) \quad \text{with probability } \lambda \quad \text{(regime 1)} \tag{1.1}$$

$$y \sim N(\mu_2, \sigma_2^2) \quad \text{with probability } (1 - \lambda) \quad \text{(regime 2)}$$

where the parameters μ_1, μ_2, σ_1^2, σ_2^2, and λ are unknown. A more complicated case arises in the switching regression model in which observations are given on a random variable y and on a vector of nonstochastic regressors x. Nature is assumed to generate each y_j from x_j by regime 1 with probability λ, and by regime 2 with probability $(1 - \lambda)$. Therefore, we have:

$$y_j = x_{1j}' \beta_1 + u_{1j} \quad \text{with probability } \lambda \quad \text{(regime 1)} \tag{1.2}$$

$$y_j = x_{2j}' \beta_2 + u_{2j} \quad \text{with probability } (1 - \lambda) \quad \text{(regime 2)}$$

where $u_{1j} \sim N(0, \sigma_1^2)$, $u_{2j} \sim N(0, \sigma_2^2)$, and the parameters β_1, β_2, σ_1^2, σ_2^2, and λ are unknown. There are also so-called disequilibrium models (which we will not discuss in this paper), in the context of demand and supply equations.

Such models are characterized by a minimum condition, as in q_j = minimum (D_j, S_j) for an ordinary demand-supply model, where the observed quantity q_j is the smaller of demand and supply. They are similar to switching regression models, since observations can come from two regimes (supply or demand equations), but the probability of an observation coming from a given regime varies over observations.

In an economic context, applications of such models are plentiful. Hamermesh (1970) used a switching regression model to examine the determination of wage bargains from observations on wage changes, changes in the consumer price index and unemployment. The dependent variable is the wage change, $\dot{w}$, and he hypothesized that the effect of cost of living changes, $\dot{c}$, on wage changes is significantly positive only when cost of living changes exceed some critical figure, which has been selected a priori. There are two wage bargain equations, each one corresponding to when $\dot{c}$ is either less than or greater than and equal to this predetermined critical figure. This is a case where regime classification is known.

Quandt and Ramsey (1978) re-estimated Hamermesh's model where there is no prior information as to the critical value of $\dot{c}$ below and above which different regression regimes are at work. They assumed that nature chooses between the two regressions for any observation, by comparing $\dot{c}$ to a critical value (known only to nature). If this critical value is $\bar{c}$, and the fraction of observations with $\dot{c} \leq \bar{c}$ is equal to λ,

then nature chooses one regime with probability λ, and the other regime where $c > \bar{c}$ with probability $(1 - \lambda)$. This is a case of no sample separation information and the regimes are unknown.

Lee and Porter (1984) used switching regression techniques to model a supply function for a railroad cartel. This supply function identifies periods in which firms are behaving non-cooperatively as opposed to cooperatively, i.e. whether price wars were occuring or not. The dependent variable is the market price for grain, so that price wars within the cartel shift the supply curve to signal reversions from collusive (higher prices) to non-collusive (lower prices) behavior. They assumed that sample separation information was available, though not perfectly reliable.

Examples of disequilibrium models can be found in the watermelon market (Suits, 1955); the market for housing starts (Fair and Jaffee, 1972); the market for chartered banks' loans to business firms (Laffont and Garcia, 1977); the U.S. labor market (Rosen and Quandt, 1978); and credit rationing in international lending (Eaton and Gersovitz, 1980).

If information on sample separation is known for switching regression models, then estimation of the parameters in the respective regimes is straightforward and is done by least squares. If information on sample separation is unknown, then we are confronted with the problem of regime classification, and estimation of the parameters is done by

either maximum likelihood, method of moments, moment generating function, or modified moment generating function. The choice of the appropriate estimation technique, however, does not concern us here, and so we will only provide a brief overview of the issues involved. A more detailed discussion of the issues may be obtained from the references cited.

We shall restrict ourselves to the basic normal mixture case of equation (1.1), since the extension to equation (1.2) is fairly straightforward. It should, first of all, be noted that parameters of finite mixtures of normal densities are identified, and that there exists no sufficient statistic for the parameters of a normal mixture (Quandt and Ramsey, 1978).

Under the assumptions of (1.1), the probability density function for y_j ($j = 1,\ldots,n$ observations) is:

$$f_j = f(y_j;\ \mu_1,\ \mu_2,\ \sigma_1^2,\ \sigma_2^2,\ \lambda) \tag{1.3}$$

$$= \lambda f_1(y_j) + (1 - \lambda) f_2(y_j)$$

$$= \frac{\lambda}{\sqrt{2\pi}\,\sigma_1} \exp\left[\frac{-(y_j - \mu_1)^2}{2\sigma_1^2}\right] +$$

$$\frac{(1 - \lambda)}{\sqrt{2\pi}\,\sigma_2} \exp\left[\frac{-(y_j - \mu_2)^2}{2\sigma_2^2}\right]$$

$f_1(y_j)$ and $f_2(y_j)$ are the normal probability density functions for observations from regime 1 and regime 2, respectively. The likelihood function for the unknown parameters is:

$$L = \prod_{j=1}^{n} f_j$$

The natural procedure for estimating the parameters using maximum likelihood is to maximize the likelihood function with respect to the parameters. This, however, runs into difficulties since as either σ_1 or σ_2 goes to zero, f_j increases without bound. It follows that the likelihood function L is unbounded, and the unboundedness of the likelihood function means that any attempt to find a global maximum will produce inconsistent estimates. To avoid this, it is possible to specify a priori knowledge of the ratio of the variances σ_1^2, σ_2^2 and to set $\sigma_1^2 = h\sigma_2^2$, or alternatively, to specify that $\sigma_2^2 \geq h\sigma_1^2$, where h is known (Goldfeld and Quandt, 1975; Kiefer, 1978). Another problem with maximum likelihood estimation is the potential singularity of the matrix of second partials of the log likelihood function, which is equivalent to a vanishing Jacobian for the set of normal equations derived from the maximum likelihood approach (Quandt and Ramsey, 1978; Hartley, 1978).

Kiefer (1978) argues that although the likelihood function is known to be unbounded at some points on the edge of the parameter space, the likelihood equations have a root which is consistent and asymptotically normally distributed. Therefore, computation of the maximum likelihood estimates should attempt to find a local maximum in the interior of the parameter space of the likelihood function. However, the attainment of such a maximum may be difficult in practice so that alternative estimators may need to be considered.

Quandt and Ramsey (1978) propose using either the

method of moments or the method of the sample moment generating function (MGF). Under the method of moments, the sample mean is equated to the theoretical first moment of equation (1.3) and the second, third, fourth and fifth sample moments about the mean to the corresponding theoretical central moments (if there are five parameters). From this, we obtain five equations from which it is possible to solve for consistent estimates of the five parameters. However, if there are K (where $K > 1$) independent variables in the switching regression model (in the normal mixture model, $K = 1$), then the number of parameters is $2K + 3$. It follows that moments of order even higher than five need to be employed, and the results are likely to be fairly unstable. While no estimates of the sampling variances are provided by this technique, it is well-known that, as a general rule, the sample variances of higher-order moments are quite large (Kendall and Stuart, 1963). For these reasons, the MGF technique is preferred over the method of moments as an estimating procedure.

The MGF method solves for the values of the parameters by minimizing a sum of squared differences between the empirical and theoretical values of the moment generating function. Define the following expression:

$$S_n(\alpha, \theta) = \bar{\epsilon}' \bar{\epsilon} \qquad (1.4)$$
$$= \sum_{t=1}^{\tau} \bar{\epsilon}_t^2$$
$$= \sum_{t=1}^{\tau} (\bar{z}_n(\alpha_t) - G(\theta, \alpha_t))^2$$

where:

$$\alpha' = (\alpha_1, \alpha_2, \ldots, \alpha_T)$$

$$\theta' = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)$$

$$\bar{\epsilon}' = (\bar{\epsilon}_1, \bar{\epsilon}_2, \ldots, \bar{\epsilon}_T)$$

$$\bar{\epsilon}_t = \frac{1}{n} \sum_{j=1}^{n} \epsilon_{jt}$$

$$\bar{z}_n(\alpha_t) = \frac{1}{n} \sum_{j=1}^{n} \exp(\alpha_t y_j)$$

$$G(\theta, \alpha_t) = \lambda \exp\left[\mu_1 \alpha_t + \frac{\alpha_t^2 \sigma_1^2}{2}\right] + (1 - \lambda) \exp\left[\mu_2 \alpha_t + \frac{\alpha_t^2 \sigma_2^2}{2}\right]$$

$$t = 1, \ldots, T; \; j = 1, \ldots, n$$

T different values of α are picked (where T $\geq$ the number of parameters, i.e. 5 in this case) and $S_n(\alpha, \theta)$ is minimized between the T estimated MGF values and their theoretical counterparts. The α_t (t = 1,...,T) are chosen so as to ensure that the corresponding normal equations derived from the minimization of $S_n(\alpha, \theta)$ with respect to θ are nonsingular. The solution to the five normal equations defines the MGF estimate, which is consistent and asymptotically normally distributed. In choosing α_t, the values which need to be avoided are those which are either very close to zero or those which are large enough so that $G(\theta, \alpha_t)$ becomes computationally intractable.

Schmidt (1982) improves on this method by postulating a modified MGF estimator, which is also consistent and where a generalized sum of squares is minimized rather than an ordinary sum of squares. The criterion in (1.4) needs to be

re-written as:

$$S_n'(\alpha, \theta) = \bar{\epsilon}' \Omega^{-1} \bar{\epsilon} \qquad (1.5)$$

where:

$$\bar{\epsilon}' = (\bar{\epsilon}_1, \bar{\epsilon}_2, \ldots, \bar{\epsilon}_T)$$

$$\bar{\epsilon}_t = \frac{1}{n} \sum_{j=1}^{n} \epsilon_{jt}$$

$$\Omega_{st} = G(\theta, \alpha_s + \alpha_t) - G(\theta, \alpha_s)G(\theta, \alpha_t)$$

$$s,t = 1,\ldots,T; \; j = 1,\ldots,n$$

The matrix Ω (of order T x T) has its st^{th} element defined as above. It comes from the covariance matrix of θ, and is proportional to the covariance matrix of the $\bar{\epsilon}_t$.

The rationale behind this approach is that the $\bar{\epsilon}_t$ ($t = 1,\ldots,T$) are correlated and have unequal variances, so that a generalized least squares criterion should be minimized, by analogy to the ordinary least squares and generalized least squares regression. When T is equal to the number of parameters (i.e. 5), the distinction between (1.4) and (1.5) does not apply because either sum of squares is minimized at zero, so that either minimization yields the same estimates. However, when T is greater than five, the estimates obtained by minimizing the generalized sum of squares are asymptotically efficient relative to those obtained by minimizing the simple sum of squares.

In comparison, the asymptotic covariance matrix of the MGF estimator can be expressed as:

$$\psi_1 = (A'A)^{-1}A'\Omega A(A'A)^{-1}$$

where A is the T x 5 matrix defined by

$$A_{jt} = \frac{\partial G(\theta, \alpha_j)}{\partial \theta_t}$$

The asymptotic covariance matrix of the modified MGF estimator is of the form

$$\Psi_2 = (A' \Omega^{-1} A)^{-1}$$

where the matrix A is defined as above. When T is equal to five, therefore,

$$\Psi_1 = \Psi_2 = A^{-1} \Omega (A')^{-1}$$

so that the modified MGF and MGF estimators are identical. However, when T is greater than five, the difference $(\Psi_1 - \Psi_2)$ is a positive semi-definite matrix, which implies that the modified MGF estimator is asymptotically efficient relative to the MGF estimator.

But, there still remains the problem of the appropriate choice of T (since asymptotically, more values are preferable to less). The values of α_t $(t = 1,\ldots,T)$, given the choice of T, may be addressed by the asymptotic covariance matrix of the resulting estimates. A useful criterion would be to choose the α's which minimize some measure of the size of the asymptotic covariance matrix, i.e. its determinant. In addition, the α_t values need to be small and need to assume different values -- this latter requirement puts some limit on how small they all can be.

1.3 Review of the Literature

Let us now turn our attention to a survey of articles which are related to this one. It has been previously stated that some studies have tried to evaluate the value of sample separation information in disequilibrium models and normal mixture models.

Goldfeld and Quandt (1975) worked on a disequilibrium model of the watermelon market (derived from Suits, 1955) and did a small sample Monte Carlo experiment based on a set of estimated parameter values. Their model is of the form:

$$Q_j = f(\text{predetermined variables})$$
$$X_j = g(P_j, Q_j, \text{predetermined variables})$$
$$P_j = h(Y_j, \text{predetermined variables})$$
$$Y_j = \text{minimum } (Q_j, X_j)$$

where Q_j, X_j, P_j and Y_j are equal to the crop of watermelons, the ex-ante or intended harvest, the price, and the actual harvest of watermelons, respectively. Two specifications were postulated -- first, where Q_j is not observed and sample separation information is therefore unknown, and second, where Q_j is observed and sample separation is known.

For their experiments, parameter values were chosen so as to reproduce approximately the levels of the dependent variables observed in the actual data. Since the first specification has less information than the second, some parameter values in the former were varied to examine the effect of the variations on the value of additional information, but the fraction of sample points was kept constant (i.e. for

$X_j \geq Q_j$ and for $X_j < Q_j$, which determines regime classification) by a compensating variation in another parameter. Over a number of cases, Goldfeld and Quandt derived the root mean-squared error ratios for the parameters (where the mean-squared error ratios can be interpreted as consistent estimates of the variance ratios) when sample separation was known relative to when it was unknown. All these ratios were less than 1.0, although the ratios were naturally larger for the parameters of the P_j equation where Q_j does not come in. A larger ratio simply means that the effect of not knowing sample separation information is minimal on the efficiency of the parameter estimates. On the whole though, knowledge of sample separation leads to smaller variances, implying that using data on Q_j has a positive value in terms of more efficient estimates.

When there is no information on Q_j, then the coefficient of Q_j in the X_j equation is zero, or the variable just drops out. However, when this coefficient should not be zero (meaning there is a significant relationship between Q_j and X_j), then there is the additional complication of the unobservable Q_j entering the X_j equation. The larger the absolute value of this coefficient, then the more valuable is information on the Q_j data for estimating the X_j equation. It follows that the larger this coefficient, then the root mean-squared errors for the parameters in the first specification with no sample separation information also increase. Experiments were conducted in this regard, where the coefficient

of Q_j was allowed to vary, and the a priori expectations were confirmed. Therefore, the superiority of the second specification rises with the value of this coefficient, since ratios of the root mean-squared errors when sample separation is known relative to when it is unknown, decline. These experiments show that efficiencies of the estimates are improved when additional information is increasingly provided to the model.

Kiefer (1979) extended the Goldfeld and Quandt results by using a large sample for a normal mixture model. The procedure involves measuring the asymptotic precision of estimates based on a marginal density (limited information estimation) and comparing it with the asymptotic precision of those based on a joint density (full information estimation). He uses the dummy variable D_j to denote regime classification information, so that the D_j variable indicates the regime which generated the j^{th} observation. Therefore, the D_j variable only appears in the full information joint density function, which can be written as:

$$f(y_j, D_j; \theta) = \lambda D_j f_1(y_j) + (1 - \lambda)(1 - D_j) f_2(y_j)$$

where θ is the vector of parameters.

The precision of a maximum likelihood estimate based on the joint and marginal density is defined, respectively, as (the subscript j was dropped for simplicity):

$$- E \frac{\partial^2 \ln f(y, D)}{\partial \theta \partial \theta'} \quad \text{and} \quad - E \frac{\partial^2 \ln f(y)}{\partial \theta \partial \theta'}$$

By definition, $\ln f(y, D) = \ln f(y) + \ln f(D/y)$, so that it follows that:

$$- E \frac{\partial^2 \ln f(y, D)}{\partial \theta \partial \theta'} = - E \frac{\partial^2 \ln f(y)}{\partial \theta \partial \theta'} - E \frac{\partial^2 \ln f(D/y)}{\partial \theta \partial \theta'}$$

The precision of the maximum likelihood estimator based on the joint density is equal to the precision of the maximum likelihood estimator based on the marginal density (here, $f(y)$ corresponds to the formulation in equation (1.3)), plus a positive definite matrix. It follows that the precision of the estimates based on the former is always greater than that for the latter. Estimates are naturally more precise when there is more information.

To confirm this relationship, Monte Carlo experiments were conducted on a normal mixture model where the only parameters being estimated are the means. Precision ratios were then taken for the full information and limited information models and converted into asymptotic variance ratios (to facilitate a comparison with the Goldfeld and Quandt results) by inversion of the information matrix. Note that the precisions of the estimates are derived from the information matrix, and that the inverse of the information matrix is a consistent estimate of the asymptotic variance-covariance matrix of the parameter estimates.

Two types of experiments were conducted -- first, when only one mean had to be estimated, and second, when two mean

values had to be estimated. Given fixed variances and the mixing parameter, the values of the means were allowed to vary. Over a series of cases, the asymptotic variance ratios for regime known relative to regime unknown were computed, and were found to be all less than 1.0, consistent with the results of Goldfeld and Quandt. Efficiency loss from using the marginal rather than the joint density could be considerable.

When the means of the samples are close together, ratios tend to be small, so the effects of implicit misclassification are serious and estimates suffer. As the means become farther apart, the probability of misclassification becomes so small, so that estimates become almost as efficient (the ratios approach a value of 1.0) as estimates based on known sample separation.

These numbers are generally a little higher than those obtained by Goldfeld and Quandt, indicating that the value of information in more complicated models (i.e. disequilibrium models) is greater than that in simpler models, as seems plausible (although this must be qualified since the Goldfeld and Quandt results are for small samples). At any rate, these results supplement the Monte Carlo evidence of the earlier study by showing that efficiency losses from not observing sample separation, found in small samples by Goldfeld and Quandt, persist and can be substantial asymptotically.

In his work, though, Kiefer assumed that the variances and the mixing parameter are known and only the means in the

normal mixture model have to be estimated. Schmidt (1981) extended Kiefer's results by also working on a normal mixture model and he derived asymptotic variance ratios (again, from the inverse of the information matrix), this time assuming that all parameters have to be estimated. The rationale behind this is that Kiefer's results understate the true value of sample separation information for the following reason. In the unknown regime case, the information matrix is not diagonal and estimates of the means are improved by knowledge of the variances and the mixing parameter, so that sample separation information is less valuable when some of the parameters are known than when all the parameters have to be estimated.

A series of experiments were conducted, each done with 100,000 replications. The values of the parameters were varied in each experiment and asymptotic variance ratios of regime unknown relative to regime known were derived. All the ratios are greater than 1.0, so the importance of having sample separation information is again verified. Among the conclusions in this study are the following: (1) the value of sample separation information depends strongly on the natural separation of the two samples, so that as the two distributions become far apart, the value of sample separation information goes to zero (ratios go to 1.0); (2) the value of sample separation information is higher for the parameters of the regime which is sampled with the lower probability; and (3) the value of sample separation information is higher

when all the parameters have to be estimated, which is why the results here show a larger value of information than in Kiefer's study, where only the means had to be estimated.

Lee and Porter (1984) also tried to evaluate the importance of sample separation information in a switching regression model. Their econometric model is different from the usual switching models in the literature in that there is additional imperfect sample separation information available and this is used as the regime indicator. Lee and Porter worked on a two-equation model with an application to cartel stability using a sample size of 328.

The model is composed of demand and supply functions for a railroad cartel, where an attempt is made to identify periods in which firms are behaving collusively, as opposed to non-cooperatively. These different behavioral rules are reflected by differing supply functions, where the supply curve can be drawn from one of two possible regimes. The cartel arrangements take the form of market share allotments. Firms then set their rates individually and the actual market share of any particular firm would depend on both the prices charged by all firms as well as on unpredictable stochastic forces. But the index of listed prices (which is the price variable in the model) is imperfect, so that member firms could not know with certainty whether secret price cutting was occuring. It is in this context that an imperfect indicator is needed to determine whether the observed price wars represent a switch from collusive to non-cooperative

behavior.

Their model consists of two equations:

$$P_j = f(I_j, \text{predetermined variables})$$
$$Q_j = g(P_j, \text{predetermined variables})$$

where P_j, Q_j, and I_j are, respectively, the price of grain; the total quantity of grain shipped; and a latent dichotomous variable which equals 1, when the industry is in a cooperative regime, and equals 0, otherwise. With no reliable information on I_j, it is measured possibly with error by W_j, a regime classification indicator. $W_j = 1$, when a trade magazine reports collusion; and $W_j = 0$, when this same trade magazine reports that a price war is occuring. This data series may not be accurate at all, but in the absence of any other information, this extra information may still help to reduce the estimated standard errors. After all, a little information (even if not entirely accurate) may be better than not having any information at all to guide in determining regime classification.

Their model was estimated twice -- first, using the partial information provided by the W_j, and second, using no information on W_j. The estimated standard errors are smaller for the former compared to the latter. However, for this particular data set, the gains in asymptotic efficiency from using the imperfect indicator are small due to the clear separation of the two underlying distributions. This is evident from the fact that the two distributions of ln P_j are far

apart, since the difference of the means is 0.48 and the variance is only 0.01. This result complements the Monte Carlo simulation results of Kiefer and Schmidt that the value of any information on regime classification becomes smaller (ratios of asymptotic variances approach 1.0) as the distributions become clearly distinct.

1.4 Plan of the Study

Our objective in this paper is to determine the value, in terms of efficiency gains, of using imperfect sample separation information, given different assumptions about the parameters and different specifications of switching regression models.

We will integrate into our study the framework of Lee and Porter regarding the use of imperfect sample separation information in switching regression models. We will also use the approach of Schmidt (1981) where all the parameters in the model have to be estimated, so as not to understate the true value of imperfect sample separation information. Similar to Kiefer's and Schmidt's procedures, we will conduct several experiments over a number of scenarios with different parameter values, each time deriving ratios of asymptotic variances, where these variances can be obtained from the corresponding information matrices. Asymptotic variance ratios will be derived twice for each experiment -- the first, showing the loss in efficiency when we have no sample separation information at all relative to full information, and the

second, showing the loss in efficiency when we have partial sample separation information (provided by an imperfect or unreliable indicator) relative to full information. A comparison of both results will show the extent of the advantages of using information even if it is inaccurate, as compared to using no regime classification information at all.

In the Lee and Porter paper, the imperfect information indicator W_j was incorporated into the switching regression model through the use of classification probabilities -- that is, the probabilities that the regime classification is right or wrong, given the true regime that the observation really belongs to. In their model, these classification probabilities were assumed to be constant for all observations.

In Chapter 2, we will deal with the simplest formulation of a switching regression model -- that of the normal mixture model. We will adopt Lee and Porter's approach of using constant probabilities of correct regime classification by our imperfect sample separation information.

In Chapter 3, we extend the previous chapter to the case where we have two explanatory variables in our switching regression model. In addition, we consider the case when the probabilities of regime classification are non-constant, and in fact, can be modelled as probit functions of the exogenous variables.

In Chapter 4, we keep the assumptions of the previous chapter but we also postulate that the mixing parameter is non-constant, so that we have varying switching probabilities.

The mixing parameter will also be modelled as a probit function of the explanatory variables.

Chapter 5 summarizes the findings of the preceding three chapters and presents the conclusions we have derived based on the series of experiments conducted.

CHAPTER TWO

THE CASE OF CONSTANT REGIME CLASSIFICATION PROBABILITIES

2.1 The Model

The first specification which we consider is the simple normal mixture model, in which a random variable y_j is drawn from $N(\mu_1, \sigma_1^2)$ with probability λ, and from $N(\mu_2, \sigma_2^2)$ with probability $(1 - \lambda)$. It can also be expressed as a switching regression model, where the only explanatory variable corresponds to the constant term. Therefore, we have the following:

$$y_j = x_{1j}\beta_1 + u_{1j} \quad \text{with probability } \lambda \quad \text{(regime 1)} \qquad (2.1)$$

$$y_j = x_{2j}\beta_2 + u_{2j} \quad \text{with probability } (1 - \lambda) \quad \text{(regime 2)}$$

For the normal mixture case, $x_{1j} = x_{2j} = 1$, and $\beta_1 = \mu_1$ and $\beta_2 = \mu_2$ are scalars. We assume that u_{1j} and u_{2j} are independently distributed, where $u_{1j} \sim N(0, \sigma_1^2)$ and $u_{2j} \sim N(0, \sigma_2^2)$. The vector of parameters $\theta' = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)$ needs to be estimated from a sample of observations on y_j. There are n observations with n_i from regime i ($i = 1,2$; $j = 1,\ldots,n$).

Suppose that there is an observed dichotomous indicator w_j for each j, which provides sample separation information. In addition, for each observation j, we define a latent dichotomous variable I_j where:

$I_j = 1$ if y_j is generated from regime 1

$I_j = 0$ otherwise

Therefore, w_j is a measure of I_j, possibly with error. The relationship between w_j and I_j can be described by the transition probability matrix given by:

	$w_j = 1$	$w_j = 0$
$I_j = 1$	p_{11}	p_{10}
$I_j = 0$	p_{01}	p_{00}

That is,

$$p_{11} = \text{Prob}(w_j = 1/I_j = 1)$$
$$p_{01} = \text{Prob}(w_j = 1/I_j = 0)$$
$$p_{10} = \text{Prob}(w_j = 0/I_j = 1)$$
$$p_{00} = \text{Prob}(w_j = 0/I_j = 0)$$

It follows that $p_{10} = 1 - p_{11}$ and $p_{00} = 1 - p_{01}$. Now, let $p = \text{Prob}(w_j = 1)$. Since $\lambda = \text{Prob}(I_j = 1)$ and $(1 - \lambda) = \text{Prob}(I_j = 0)$, then:

$$\begin{aligned} p &= \text{Prob}(I_j = 1)\text{Prob}(w_j = 1/I_j = 1) + \\ &\qquad \text{Prob}(I_j = 0)\text{Prob}(w_j = 1/I_j = 0) \\ &= \lambda p_{11} + (1 - \lambda)p_{01} \\ &= \lambda p_{11} + (1 - \lambda)(1 - p_{00}) \end{aligned}$$

The density function $f(y_j)$ for y_j when we have a mixture of two normal distributions is given in equation (1.3) as:

$$f(y_j) = \text{Prob}(I_j = 1)f_1(y_j) + \text{Prob}(I_j = 0)f_2(y_j) \qquad (2.2)$$

$$f(y_j) = \lambda f_1(y_j) + (1 - \lambda) f_2(y_j)$$

When imperfect sample separation information using the observed indicator, w_j, is incorporated into the model, then the joint density function for y_j and w_j is:

$$\begin{aligned} f(y_j, w_j) &= f_1(y_j)\text{Prob}(w_j, I_j = 1) + \\ &\quad f_2(y_j)\text{Prob}(w_j, I_j = 0) \\ &= f_1(y_j)(w_j \lambda p_{11} + (1 - w_j)\lambda(1 - p_{11})) + \\ &\quad f_2(y_j)(w_j(1 - \lambda)p_{01} + \\ &\quad (1 - w_j)(1 - \lambda)(1 - p_{01})) \\ &= \lambda f_1(y_j)(w_j p_{11} + (1 - w_j)(1 - p_{11})) + \\ &\quad (1 - \lambda) f_2(y_j)(w_j p_{01} + (1 - w_j)(1 - p_{01})) \\ &= \lambda f_1(y_j)(w_j p_{11} + (1 - w_j)(1 - p_{11})) + \\ &\quad (1 - \lambda) f_2(y_j)(w_j(1 - p_{00}) + (1 - w_j)p_{00}) \end{aligned} \tag{2.3}$$

where:

$$f_i(y_j) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[\frac{-(y_j - \mu_i)^2}{2\sigma_i^2}\right]$$

$i = 1,2;\ j = 1,\ldots,n$

The regime classification indicator w_j contains some information on sample separation if p_{11} is not equal to p_{01}. When $p_{11} = p_{01}$, or alternatively, when $p_{11} = 1 - p_{00}$, then the $\text{Prob}(w_j/I_j) = \text{Prob}(w_j)$ and the joint density function is:

$$\begin{aligned} f(y_j, w_j) &= (\lambda f_1(y_j) + (1 - \lambda) f_2(y_j)) \times \\ &\quad (w_j p + (1 - w_j)(1 - p)) \end{aligned}$$

so that the indicator w_j does not contain any information on

sample separation. This is equivalent to having no information at all, as in equation (2.2). This is, in fact, Schmidt's model and also Kiefer's marginal density function (limited information model). On the other hand, when $p_{11} = 1$ and $p_{01} = 0$ (alternatively, $p_{00} = 1$), the indicator w_j provides perfect sample separation information and the joint density function is expressed as:

$$f(y_j, w_j) = \lambda f_1(y_j)w_j + (1 - \lambda)f_2(y_j)(1 - w_j)$$

This is equivalent to Kiefer's joint density or full information model, where our w_j is his D_j, the indicator of perfect information on regime classification.

2.2 Derivation of Asymptotic Variances

We adopt the approach of Schmidt here. When the regime is known or when perfect sample separation information is available, the asymptotic variances of $\sqrt{n}(\hat{\mu}_1 - \mu_1)$, $\sqrt{n}(\hat{\mu}_2 - \mu_2)$, $\sqrt{n}(\hat{\sigma}_1^2 - \sigma_1^2)$, and $\sqrt{n}(\hat{\sigma}_2^2 - \sigma_2^2)$ are, respectively:

$$\frac{\sigma_1^2}{\lambda};$$

$$\frac{\sigma_2^2}{1 - \lambda};$$

$$\frac{2\sigma_1^4}{\lambda}; \text{ and}$$

$$\frac{2\sigma_2^4}{1 - \lambda}.$$

They are derived from the diagonal elements of the inverse of the information matrices from the corresponding likelihood functions of the respective known densities, $f_1(y_j)$ and $f_2(y_j)$. The terms λ and $(1 - \lambda)$ appear in the above expressions since they adjust for the correct sample size in each regime (n_1 or n_2), relative to the total number of observations n. This follows from the implicit relationship that $n_1 = \lambda n$ and $n_2 = (1 - \lambda)n$. λ has a binomial distribution, so the asymptotic variance of $\sqrt{n}(\hat{\lambda} - \lambda)$ is

$$\lambda(1 - \lambda).$$

When the regime is either completely unknown or is partly known (due to the partial sample separation information available), the asymptotic variances are derived in the same manner. Therefore, the asymptotic variances of $\sqrt{n}(\hat{\theta} - \theta)$ come from the diagonal elements of the inverse of the corresponding information matrices. That is, $\sqrt{n}(\hat{\theta} - \theta)$ approaches the distribution specified by $N\left(0, \lim\left(\frac{1}{n}\mathcal{I}\right)^{-1}\right)$.

The Fisher information matrix is defined as:

$$\mathcal{I} = -E\left[\frac{\partial^2 \ln L}{\partial\theta\,\partial\theta'}\right]$$

where:

$$L = \prod_{j=1}^{n} f_j$$

$$\ln L = \sum_{j=1}^{n} \ln f_j$$

When the regime is completely unknown, f corresponds to the density function laid out in (2.2) as:

$$f(y_j;\ \theta) = \lambda f_1(y_j) + (1 - \lambda) f_2(y_j)$$

θ is a (5 x 1) vector of parameters, defined by $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)'$. When the regime is partly known, f corresponds to the density function in (2.3):

$$\begin{aligned} f(y_j, w_j;\ \theta) = {} & \lambda f_1(y_j)(w_j p_{11} + (1 - w_j)(1 - p_{11})) + \\ & (1 - \lambda) f_2(y_j)(w_j p_{01} + \\ & (1 - w_j)(1 - p_{01})) \end{aligned}$$

θ is a (7 x 1) vector of parameters, defined by $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda, p_{11}, p_{01})'$.

The expected value of the expression that denotes the information matrix was intractable analytically, so that we calculate the information matrix instead by simulation techniques[1/] using the following expansion:

$$\mathcal{I} = -E \sum_{j=1}^{n} \left[\frac{\partial^2 \ln f_j}{\partial \theta \partial \theta'} \right] \tag{2.4}$$

$$= -E \sum_{j=1}^{n} \left[\frac{1}{f_j} \frac{\partial^2 f_j}{\partial \theta \partial \theta'} - \frac{1}{f_j^2} \left(\frac{\partial f_j}{\partial \theta} \right) \left(\frac{\partial f_j}{\partial \theta} \right)' \right]$$

The model we have here when there is no sample separation information is actually Schmidt's model, so we do not need to simulate the information matrix corresponding to the density function of (2.2) since that was done in his work; we will just adopt his results. All we need to simulate is

---

[1/] From the definition of the information matrix, we know that $(1/n\,\mathcal{I})$ has a limit. We therefore simulate $\lim (1/n\,\mathcal{I})$ by calculating $(1/n\,\mathcal{I})$ for some finite though large n.

the information matrix when we have imperfect sample separation information. In order to facilitate a comparison between the results, we use Schmidt's approach in our experiments.

The information matrix was evaluated by a simulation of 100,000 trials derived from a normal or Gaussian random variable generator. For any set of values assigned to θ, draws were made from the appropriate normal mixture distribution. The first and second derivatives were calculated in accordance with the expression in (2.4). The resulting 100,000 matrices were then averaged to obtain the information matrix, and the asymptotic variances are the corresponding diagonal elements of the inverse of the information matrix.

When we have imperfect sample separation information, the expressions in (2.4) are laid out below, where f comes from the density function defined by (2.3). The first derivatives of $f(y, w; \theta)$ with respect to θ are (we drop the subscript j for simplicity):

$$\frac{\partial f}{\partial \mu_1} = \lambda Q_1 \frac{\partial f_1}{\partial \mu_1}$$

$$\frac{\partial f}{\partial \mu_2} = (1 - \lambda) Q_2 \frac{\partial f_2}{\partial \mu_2}$$

$$\frac{\partial f}{\partial \sigma_1^2} = \lambda Q_1 \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial f}{\partial \sigma_2^2} = (1 - \lambda) Q_2 \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial f}{\partial \lambda} = f_1 Q_1 - f_2 Q_2$$

$$\frac{\partial f}{\partial p_{11}} = \lambda f_1(w - (1 - w))$$

$$\frac{\partial f}{\partial p_{01}} = (1 - \lambda)f_2(w - (1 - w))$$

where:

$$Q_1 = wp_{11} + (1 - w)(1 - p_{11})$$

$$Q_2 = wp_{01} + (1 - w)(1 - p_{01})$$

$$\frac{\partial f_i}{\partial \mu_i} = \frac{f_i}{\sigma_i^2}(y - \mu_i)$$

$$\frac{\partial f_i}{\partial \sigma_i^2} = \frac{f_i}{2\sigma_i^2}\left[-1 + \frac{(y - \mu_i)^2}{\sigma_i^2}\right]$$

$i = 1,2$

The non-zero second derivatives of $f(y, w; \theta)$ with respect to θ are:

$$\frac{\partial^2 f}{\partial \mu_1^2} = \lambda Q_1 \frac{\partial^2 f_1}{\partial \mu_1^2}$$

$$\frac{\partial^2 f}{\partial \mu_1 \partial \sigma_1^2} = \lambda Q_1 \frac{\partial^2 f_1}{\partial \mu_1 \partial \sigma_1^2}$$

$$\frac{\partial^2 f}{\partial \mu_1 \partial \lambda} = Q_1 \frac{\partial f_1}{\partial \mu_1}$$

$$\frac{\partial^2 f}{\partial \mu_1 \partial p_{11}} = \lambda (w - (1 - w)) \frac{\partial f_1}{\partial \mu_1}$$

$$\frac{\partial^2 f}{\partial \mu_2^2} = (1 - \lambda)Q_2 \frac{\partial^2 f_2}{\partial \mu_2^2}$$

$$\frac{\partial^2 f}{\partial \mu_2 \, \partial \sigma_2^2} = (1 - \lambda) Q_2 \frac{\partial^2 f_2}{\partial \mu_2 \, \partial \sigma_2^2}$$

$$\frac{\partial^2 f}{\partial \mu_2 \, \partial \lambda} = -Q_2 \frac{\partial f_2}{\partial \mu_2}$$

$$\frac{\partial^2 f}{\partial \mu_2 \partial p_{01}} = (1 - \lambda)(w - (1 - w)) \frac{\partial f_2}{\partial \mu_2}$$

$$\frac{\partial^2 f}{\partial (\sigma_1^2)^2} = \lambda Q_1 \frac{\partial^2 f_1}{\partial (\sigma_1^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2 \partial \lambda} = Q_1 \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2 \partial p_{11}} = \lambda (w - (1 - w)) \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial^2 f}{\partial (\sigma_2^2)^2} = (1 - \lambda) Q_2 \frac{\partial^2 f_2}{\partial (\sigma_2^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2 \partial \lambda} = -Q_2 \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2 \partial p_{01}} = (1 - \lambda)(w - (1 - w)) \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial^2 f}{\partial \lambda \partial p_{11}} = (w - (1 - w)) f_1$$

$$\frac{\partial^2 f}{\partial \lambda \partial p_{01}} = - (w - (1 - w)) f_2$$

where:

$$\frac{\partial^2 f_i}{\partial \mu_i^2} = \frac{f_i}{\sigma_i^2}\left[-1 + \frac{(y - \mu_i)^2}{\sigma_i^2}\right]$$

$$\frac{\partial^2 f_i}{\partial (\sigma_i^2)^2} = \frac{\partial f_i}{\partial \sigma_i^2}\left[\frac{(y - \mu_i)^2}{2\sigma_i^4} - \frac{1}{2\sigma_i^2}\right] + f_i\left[\frac{1}{2\sigma_i^4} - \frac{(y - \mu_i)^2}{\sigma_i^6}\right]$$

$$\frac{\partial^2 f_i}{\partial \mu_i \, \partial \sigma_i^2} = \frac{\partial f_i}{\partial \sigma_i^2}\frac{(y - \mu_i)}{\sigma_i^2} - f_i\frac{(y - \mu_i)}{\sigma_i^4}$$

$i = 1,2$

When there is no sample separation information, such that regime is unknown, θ is of dimension (5 x 1) or (2K + 3) x 1 where K is the total number of explanatory variables (i.e. K = 1 in this case). It follows that $\mathcal{I}$ is of dimension (2K + 3) x (2K + 3). When there is partial sample separation information, such that regime is partly known, θ is of dimension (7 x 1) or (4K + 3) x 1, where K is still equal to 1. Therefore, $\mathcal{I}$ is of dimension (4K + 3) x (4K + 3).

2.3 The Value of Imperfect Information

We derive here one set of ratios -- asymptotic variances with regime partly known relative to regime known. We also need the asymptotic variance ratios with regime unknown relative to regime known, but these will be simply adopted from Schmidt's study. The results are comparable, since the same simulation techniques have been employed in evaluating the information matrix.

All the results are presented in Tables 1 through 4, for a variety of cases. It is to be noted that asymptotic variance ratios are not given for the parameters p_{11} and p_{01}, since these parameters are not estimated at all, when the regime is known. All the figures in the tables are greater than or equal to one, and the extent to which they differ from one measures the value of imperfect sample separation information or just sample separation information as the case may be. That is, they measure how much we lose on efficiency grounds in parameter estimation, when we use imperfect information, or no information at all, as compared to perfect information when assigning regime classification.

The main interest here concerns the effects of the parameters p_{11} and p_{01}, which represent the level of reliability or accuracy of the available information. In Schmidt's study, when there is no sample separation information at all, only the first five parameters were used. With the partial information provided by our additional parameters p_{11} and p_{01}, we expect our asymptotic variance ratios to be less than or equal to his asymptotic variance ratios. After all, any piece of information on sample separability, even if not entirely accurate, may facilitate identification of regime membership for the observations, and thereby improve the efficiency of the parameter estimates, as compared to when no information is used at all. For purposes of comparison, we present Schmidt's figures in parentheses underneath the figures we derived.

We conduct four types of experiments here. First, for a given set of parameter values, we vary our probabilities of regime classification[2/]. The different values assigned to p_{11} and p_{01} values represent the range from highly imperfect information to almost perfect information on regime classification. In the other three types of experiments, we choose a particular p_{11} and p_{01} mix, and allow the following to vary -- the difference between the means, the variances, and the mixing parameter.

Table 1 presents the results when we couple different regime classification probabilities with fixed values of the other parameters -- $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$, and $\lambda = .5$. The values assigned to the means and variances are not as restrictive as they might seem, in the sense that they are invariant to translation ($\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$ give the same results as $\mu_1 = -6$, $\mu_2 = -4$, $\sigma_1 = \sigma_2 = 1$) and to scale ($\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$ give the same results as $\mu_1 = 0$, $\mu_2 = 4$, $\sigma_1 = \sigma_2 = 2$).

When p_{11} is equal to p_{01}, then we have the special case of there being no sample separation information at all, and the ratios derived here should be the same as Schmidt's. The difference (i.e. 41.7 versus 41.1 for μ_1) is presumable due to randomness in the simulation of the information matrix. When p_{11} is equal to p_{00} (where $p_{00} = 1 - p_{01}$), that is, when

---

[2/] These regime classification probabilities are assumed constant for all observations in each case, and can be estimated (as Lee and Porter did) by maximum likelihood.

Table 1. Ratios of Asymptotic Variances

Varying p_{11} and p_{01} when $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$

Parameters		Ratios: Partly Known (Unknown)/Known				
p_{11}	p_{01}	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
.5	.5	41.7	40.4	13.0	12.0	79.4
.4	.6	20.1	19.3	7.39	7.03	36.8
.3	.7	7.68	7.27	4.09	3.87	12.7
.2	.8	3.92	3.71	2.94	2.81	5.57
.1	.9	2.36	2.26	2.28	2.21	2.83
.05	.95	1.81	1.76	1.92	1.90	1.99
		(41.1)	(40.4)	(12.7)	(12.6)	(78.8)
		Additional Results				
.4	.5	32.8	31.7	10.7	10.2	61.8
.7	.6	28.9	28.1	9.62	9.35	54.4
.4	.3	28.6	27.9	9.55	9.27	53.9
.1	.2	26.8	25.4	9.26	8.48	49.5
.9	.8	25.2	24.3	8.67	8.36	47.1
.4	.1	8.10	8.43	4.03	4.35	14.2
		(41.1)	(40.4)	(12.7)	(12.6)	(78.8)

n = 100000. Figures in parentheses are the ratios of asymptotic variances when regime is unknown relative to when regime is known. This is true for the other tables in this chapter.

there are equal probabilities of correct classification into each regime, then the ratios diminish considerably, with the figures being lowest (efficiency is highest) when there is greater certainty about rightly or wrongly assigning the observation into each regime. The ratios approach one when p_{11} goes to zero, and p_{01} goes to one (alternatively, when p_{11} goes to one, and p_{01} goes to zero); that is, when there is almost perfect sample separation information. In this sense, the use of imperfect sample separation information leads to estimates which are almost as efficient as those derived when regime classification is completely known.

An interesting observation here is that the value of information is unchanged when p_{11} and p_{01} are symmetric (i.e. $p_{11} = .2$, $p_{01} = .8$ give the same results as $p_{11} = .8$, $p_{01} = .2$; alternatively, $p_{11} = p_{00} = .2$ give the same results as $p_{11} = p_{00} = .8$). This is a consequence of the identification issue referred to in Lee and Porter, such that when $x_{1j} = x_{2j}$ for all j, as they are here (they are both equal to one), then the names of the two regimes can simply be interchanged, and this holds true when there is no sample separation information and even when there is imperfect sample separation information. This does not really come as a surprise since in the normal mixture model, the only parameters being estimated in a regression sense are the means; therefore, it makes no difference at all about having the same probabilities for right or wrong regime classification, since we can merely switch the names of the regimes.

Additional results show different p_{11} and p_{01} values paired together. When p_{11} and p_{01} are close together but are in the intermediate range (.4 to .6), then the ratios are highest. This occurs when uncertainty about regime classification is at its peak, since the imperfect information indicates that there are almost equal chances of misclassification into both regimes (p_{11} and p_{01} are close to .5). Note that at the extreme, when $p_{11} = p_{01} = .5$, we have no information at all. When p_{11} and p_{01} are close together but are out of the intermediate range, then the ratios go down. This means that when there is greater certainty of correct regime classification into the two regimes, or when the partial sample separation information is quite reliable for both regimes, then the ratios decline and efficiency improves.

Tables 2, 3 and 4 illustrate the case of a particular p_{11}, p_{01} mix -- we choose $p_{11} = .8$, $p_{01} = .2$. In Table 2, μ_2 is allowed to vary. The results are similar to Schmidt's findings that the value of sample separation information depends on the natural separation of the two regimes. As the distributions become far apart (μ_2 increases while μ_1 is constant), the ratios diminish and tend to approach one. When the means are very close together, the resulting ratios show the substantial gains in efficiency when information is quite accurate as compared to using no information at all.

Table 3 takes the cases where $\lambda = .2$ and $\lambda = .5$. The results are again similar to the earlier findings that when the distributions are fairly close to each other

Table 2. Ratios of Asymptotic Variances

Varying μ_2 when $\mu_1 = 0$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$, $p_{11} = p_{00} = .8$

Parameter	Ratios				
	Partly Known (Unknown)/Known				
μ_2	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
2	3.68	3.72	2.77	2.87	5.35
	(41.1)	(40.4)	(12.7)	(12.6)	(78.8)
3	1.93	1.91	1.95	1.96	1.75
	(4.21)	(4.23)	(3.51)	(3.58)	(3.78)
4	1.28	1.27	1.48	1.48	1.13
	(1.50)	(1.52)	(1.77)	(1.82)	(1.25)
5	1.08	1.07	1.19	1.20	1.02
	(1.11)	(1.11)	(1.26)	(1.29)	(1.04)
6	1.02	1.02	1.06	1.07	1.00
	(1.02)	(1.03)	(1.07)	(1.10)	(1.01)
8	1.00	.999	.992	1.00	1.00
	(1.00)	(1.00)	(1.00)	(1.00)	(1.00)

Table 3. Ratios of Asymptotic Variances

Varying λ when $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = \sigma_2 = 1$, $p_{11} = p_{00} = .8$

Parameter	Ratios				
	Partly Known (Unknown)/Known				
λ	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
.2	6.52	2.55	3.96	2.22	6.98
	(66.3)	(20.9)	(17.1)	(8.08)	(89.1)
.5	3.68	3.72	2.77	2.87	5.35
	(41.1)	(40.4)	(12.7)	(12.6)	(78.8)

Table 4. Ratios of Asymptotic Variances

Varying σ_2 when $\mu_1 = 0$, $\mu_2 = 2$, $\sigma_1 = 1$, $\lambda = .5$, $p_{11} = p_{00} = .8$

Parameter	Ratios Partly Known (Unknown)/Known				
σ_2	$\hat{\mu}_1$	$\hat{\mu}_2$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
1	3.68	3.72	2.77	2.87	5.35
	(41.1)	(40.4)	(12.7)	(12.6)	(78.8)
2	1.98	3.76	2.77	1.38	7.52
	(3.56)	(7.77)	(7.64)	(1.94)	(18.9)
4	1.52	1.23	2.38	1.21	2.84
	(1.96)	(1.34)	(3.97)	(1.31)	(3.80)

($\mu_1 = 0$, $\mu_2 = 2$) and the imperfect information is quite reliable ($p_{11} = p_{00} = .8$), then there are large efficiency gains in using partial information relative to using no information at all. In addition, we observe that the value of sample separation information is higher for the parameters of the regime which is observed with the lower probability, in this case, regime 1.

Table 4 gives results when the variances are not equal. The larger the difference between the variances of the two samples, the lower the ratios become. It is apparent that not only does mean disparity between the regimes contribute to distinct sample separation, but also disparity of the variances. Another observation here is that the ratios are higher for the variance parameter of the sample which has the smaller variance and the reason behind this is fairly intuitive. A surprising finding here though, is that the decline in the ratios as the difference between the variances widens is not monotonic for the mixing parameter when partial information is available, and the reason for this is not clear.

2.4 Summary

We have studied the value of imperfect sample separation information in a simple normal mixture model, where all the parameters have to be estimated. This was done under different values for the probabilities of correct regime classification. The ratios of asymptotic variances for

regime partly known relative to the asymptotic variances for regime known were computed. These ratios are highest when there is greater uncertainty about regime classification (p_{11} and p_{00} are in the intermediate range) and the ratios are lowest when there is almost perfect certainty about right or wrong classification for both regimes (p_{11} and p_{00} are in the extreme range). In between is a continuum of values depending on the reliability of the sample separation information for each regime.

A variety of experiments were also conducted and these show that the value of sample separation information largely depends on how much alike the two samples are. When the samples are hard to distinguish from one another, then the value of information is highest. At any rate, the presence of the partial sample separation information tends to diminish the value of any other additional information, since the figures derived are considerably lower than those when there is no sample separation information at all.

These results suggest that any information should be used, even if there is uncertainty about its reliability or accuracy, since even imperfect sample separation information improves the efficiency of the estimates. Of course, the more reliable the imperfect sample separation information, the greater the gains in efficiency.

CHAPTER THREE

THE CASE OF NON-CONSTANT REGIME CLASSIFICATION PROBABILITIES

3.1 Introduction

In the previous chapter, we considered and evaluated the value of imperfect sample separation information in a normal mixture model, where the imperfect information is reflected through constant probabilities of regime classification. We concluded that the more reliable the imperfect information, the greater the gains in efficiency, since there is greater certainty of right or wrong regime classification.

We now extend that model to a switching regression case where there are at least two independent variables -- a constant term and one or more other explanatory variables. In addition, we consider the case when the classification probabilities are non-constant, and in fact, can be modelled as probit functions of the exogenous variables. The rationale behind this is that the values of the explanatory variables are highly likely to affect the regime classification of the dependent variable, increasing the reliability of the imperfect information indicator. Consequently, treatment of the probabilities as non-constant for each observation adds more reliable information to the model and will hopefully improve the efficiency of estimation.

The framework for the use of imperfect sample separation information was derived from Lee and Porter who used

switching regression techniques to model a supply function for a railroad cartel. In their model, the observed regime classifications were obtained from data from a trade magazine (presumably reported with error) on whether there were price wars or not. The probabilities that these regime classifications were in fact, correct were assumed constant, and therefore independent of the exogenous variables.

Their model can be improved upon by postulating that the classification probabilities are dependent on the exogenous variables and will differ for each time period. Taking the Lee and Porter application as a case in point, we note that their explanatory variables include a Great Lakes dummy variable and several dummy variables on structural changes. The Great Lakes dummy variable documents when the Great Lakes were made open to navigation so that the cartel faced its main source of competition. The structural changes dummy variables are used to proxy changes caused by the entry, acquisitions or additions to existing networks in the railroad industry. When the Great Lakes were made open to navigation, or when there were instances of entry and new acquisitions, we expect that there will be price cutting or non-cooperative behavior among the firms in the cartel, due to the presence of other competitors in the industry, and this will be reflected in the imperfect indicators of information -- data from the trade magazine.

Using this information during each time period adds to the certainty on regime classification, as to whether there

were indeed price wars or not. This raises the probabilities of correct classification and also leads to higher efficiencies of parameter estimation, as opposed to the case when constant probabilities are applied for each time period as Lee and Porter did. Our suggested treatment of the derivation of classification probabilities as non-constant seems to be a plausible alternative to theirs in the sense that we use more information (at no extra cost of obtaining this information) in solving for these probabilities, which presumably improves efficiency. Also, their model is a special case of ours, so we can test the adequacy of their model against the alternative of our model.

3.2 The Model

We extend the model of the previous chapter to the case when there are at least two explanatory variables, and when the probabilities of regime classification (i.e. p_{11} and p_{00}) are not fixed. Suppose for simplicity that $x_{1j} = x_{2j}$; we call it x_j so the basic switching regression model is:

$$y_j = x_j' \beta_1 + u_{1j} \quad \text{with probability } \lambda \quad \text{(regime 1)} \tag{3.1}$$

$$y_j = x_j' \beta_2 + u_{2j} \quad \text{with probability } (1 - \lambda) \quad \text{(regime 2)}$$

β_1 and β_2 are vectors of parameters. The error terms u_{1j} and u_{2j} are assumed to be independently and normally distributed with means 0 and variances σ_1^2 and σ_2^2, respectively.

When there is imperfect sample separation information or the regime is partly known, we can then consider an observability model on probability classification like:

$$p_{11j} = F(x_j' \gamma_1) \quad \text{where } p_{11j} = \text{Prob}(w_j = 1/I_j = 1) \text{ for each } j$$

$$p_{00j} = F(x_j' \gamma_0) \quad \text{where } p_{00j} = \text{Prob}(w_j = 0/I_j = 0) \text{ for each } j$$

$$p_{10j} = 1 - p_{11j}$$

$$p_{01j} = 1 - p_{00j}$$

where $F(\ )$ is a standard normal cumulative distribution function, and γ_1 and γ_0 are vectors of parameters. w_j is the observed dichotomous indicator which provides sample separation information, while I_j is the latent dichotomous indicator of the actual regime classification. In essence, the regime classification probabilities are probit models of observability. This contains the Lee and Porter model as a special case, that is, all the elements of γ_1 and γ_0 are zero, except for those corresponding to the constant term. The joint density function for y_j and w_j is then re-written from (2.3) and given as:

$$\begin{aligned} f_j &= f(y_j, w_j; \theta) \qquad (3.2) \\ &= \lambda f_1(y_j)(w_j p_{11j} + (1 - w_j)(1 - p_{11j})) + \\ &\quad (1 - \lambda) f_2(y_j)(w_j p_{01j} + (1 - w_j)(1 - p_{01j})) \\ &= \lambda f_1(y_j)(w_j F(x_j' \gamma_1) + (1 - w_j) F(-x_j' \gamma_1)) + \\ &\quad (1 - \lambda) f_2(y_j)(w_j F(-x_j' \gamma_0) + (1 - w_j) F(x_j' \gamma_0)) \end{aligned}$$

where:

$$f_i(y_j) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[\frac{-(y_j - x_j'\beta_i)^2}{2\sigma_i^2}\right]$$

$$p_{11j} = \int_{-\infty}^{x_j'\gamma_1} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v_j^2}{2}\right] dv_j$$

$$p_{00j} = \int_{-\infty}^{x_j'\gamma_0} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v_j^2}{2}\right] dv_j$$

$i = 1,2;\ j = 1,\ldots,n$

That is, $f_1(y_j)$ and $f_2(y_j)$ are normal probability density functions with means and variances given by $N(x_j'\beta_1, \sigma_1^2)$ and $N(x_j'\beta_2, \sigma_2^2)$, respectively; and p_{11j} and p_{00j} are probabilities of correct regime classification denoted as probit models.

3.3 Derivation of Asymptotic Variances

When the regime is known, the asymptotic variances of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$, $\sqrt{n}(\hat{\beta}_2 - \beta_2)$, $\sqrt{n}(\hat{\sigma}_1^2 - \sigma_1^2)$, $\sqrt{n}(\hat{\sigma}_2^2 - \sigma_2^2)$, and $\sqrt{n}(\hat{\lambda} - \lambda)$ are, respectively:

$$\frac{\sigma_1^2}{\lambda} \cdot \left(\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} x_j x_j'\right)^{-1};$$

$$\frac{\sigma_2^2}{1-\lambda} \left(\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} x_j x_j'\right)^{-1};$$

$$\frac{2\sigma_1^4}{\lambda};$$

$$\frac{2\sigma_2^4}{1-\lambda}; \text{ and}$$

$$\lambda(1-\lambda).$$

As in the previous chapter, there are no asymptotic variances for the parameters γ_1 and γ_0 (which enter the p_{11j} and p_{00j} probability functions of regime classification) since these parameters are irrelevant when regimes are completely known. The above expressions for the asymptotic variances of $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}_1^2$ and $\hat{\sigma}_2^2$ are derived from the inverse of the information matrices from the corresponding likelihood functions of the known densities associated with the respective regimes. The asymptotic variance of $\hat{\lambda}$ comes from that of the binomial distribution.

For cases when the regime is either completely unknown or is partly known, the asymptotic variances of $\sqrt{n}(\hat{\theta} - \theta)$ are derived in the same way -- from the diagonal elements of the inverse of the Fisher information matrix. Therefore, $\sqrt{n}(\hat{\theta} - \theta)$ approaches the distribution specified by the following expression -- $N\left(0, \lim\left(\frac{1}{n}\mathcal{I}\right)^{-1}\right)$.

The information matrix is defined as:

$$\mathcal{I} = -E\ \frac{\partial^2 \ln L}{\partial\theta\,\partial\theta'}$$

where:

$$L = \prod_{j=1}^{n} f_j$$

$$\ln L = \sum_{j=1}^{n} \ln f_j$$

Therefore, when the regime is completely unknown, f corresponds to the density function of (2.2) given as:

$$f(y_j;\ \theta) = \lambda f_1(y_j) + (1 - \lambda) f_2(y_j) \qquad (3.3)$$

θ is a $(2K + 3) \times 1$ vector of parameters, where K is the number of explanatory variables. In particular, $\theta = (\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \lambda)'$ where $\beta_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1K})'$ and $\beta_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2K})'$. When the regime is partly known, f corresponds to the density function in (3.2):

$$\begin{aligned} f(y_j, w_j; \theta) = \lambda & f_1(y_j)(w_j F(x_j'\gamma_1) + \\ & (1 - w_j)(1 - F(x_j'\gamma_1))) + \\ & (1 - \lambda) f_2(y_j)(w_j(1 - F(x_j'\gamma_0)) + \\ & (1 - w_j)F(x_j'\gamma_0)) \end{aligned}$$

θ is a $(4K + 3) \times 1$ vector of parameters given by $(\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \lambda, \gamma_1', \gamma_0')'$ where β_1 and β_2 are defined as previously; and $\gamma_1 = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1K})'$ and $\gamma_0 = (\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0K})'$ are additional parameters.

As in the earlier chapter, the expected value of the expression that represents the information matrix was analytically intractable so that we instead calculate the information matrix by simulation techniques in either of two ways:

$$(1)\quad \mathcal{I} = -E \sum_{j=1}^{n} \left[\frac{\partial^2 \ln f_j}{\partial \theta \partial \theta'} \right]$$

$$= -E \sum_{j=1}^{n} \left[\frac{1}{f_j} \frac{\partial^2 f_j}{\partial \theta \partial \theta'} - \frac{1}{f_j^2} \left(\frac{\partial f_j}{\partial \theta} \right) \left(\frac{\partial f_j}{\partial \theta} \right)' \right]$$

$$(2)\quad \mathcal{I} = E \sum_{j=1}^{n} \left[\left(\frac{\partial \ln f_j}{\partial \theta} \right) \left(\frac{\partial \ln f_j}{\partial \theta} \right)' \right]$$

$$= E \sum_{j=1}^{n} \left[\frac{1}{f_j^2} \left(\frac{\partial f_j}{\partial \theta} \right) \left(\frac{\partial f_j}{\partial \theta} \right)' \right]$$

The second method of calculation follows from the first method, in the sense that, in the limit, the expression $-E\sum_{j=1}^{n}\left[\frac{1}{f_j}\frac{\partial^2 f_j}{\partial\theta\,\partial\theta'}\right]$ goes to zero. In addition, the second method has the added advantage of being positive definite always, and not just in the limit. For this reason, we choose the second method of calculation, and throughout the experiments we will be conducting, the information matrix[3/] is to be calculated as follows:

$$\mathcal{I} = E\sum_{j=1}^{n}\left[\frac{1}{f_j^2}\left(\frac{\partial f_j}{\partial\theta}\right)\left(\frac{\partial f_j}{\partial\theta}\right)'\right]$$

For the case where regime classification is completely unknown or when there is no sample separation information at all, the first derivatives of $f(y;\theta)$ (we drop the subscript j for simplicity) with respect to θ are:

$$\frac{\partial f}{\partial\beta_{1k}} = \lambda\frac{\partial f_1}{\partial\beta_{1k}}$$

$$\frac{\partial f}{\partial\beta_{2k}} = (1-\lambda)\frac{\partial f_2}{\partial\beta_{2k}}$$

$$\frac{\partial f}{\partial\sigma_1^2} = \lambda\frac{\partial f_1}{\partial\sigma_1^2}$$

[3/] The expressions for the elements of the information matrix when the first method of calculation is used are also derived and are given in Appendix A.

$$\frac{\partial f}{\partial \sigma_2^2} = (1 - \lambda) \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial f}{\partial \lambda} = f_1 - f_2$$

where:

$$\frac{\partial f_i}{\partial \beta_{ik}} = \frac{f_i}{\sigma_i^2} (y - x' \beta_i) x_k$$

$$\frac{\partial f_i}{\partial \sigma_i^2} = \frac{f_i}{2\sigma_i^2} \left[\frac{(y - x' \beta_i)^2}{\sigma_i^2} - 1 \right]$$

$i = 1,2;\ k = 1,2,\ldots,K$

Since θ is of dimension $(2K + 3) \times 1$, then it follows that $\mathcal{J}$ is of dimension $(2K + 3) \times (2K + 3)$.

For the case where regime classification is partly known or when there is imperfect sample separation information, the first derivatives of $f(y, w; \theta)$ with respect to θ are:

$$\frac{\partial f}{\partial \beta_{1k}} = \lambda Q_1 \frac{\partial f_1}{\partial \beta_{1k}}$$

$$\frac{\partial f}{\partial \beta_{2k}} = (1 - \lambda) Q_2 \frac{\partial f_2}{\partial \beta_{2k}}$$

$$\frac{\partial f}{\partial \sigma_1^2} = \lambda Q_1 \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial f}{\partial \sigma_2^2} = (1 - \lambda) Q_2 \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial f}{\partial \lambda} = f_1 Q_1 - f_2 Q_2$$

$$\frac{\partial f}{\partial \gamma_{1k}} = \lambda f_1 (w - (1 - w)) \frac{\partial F(x'\gamma_1)}{\partial \gamma_{1k}}$$

$$\frac{\partial f}{\partial \gamma_{0k}} = -(1 - \lambda) f_2 (w - (1 - w)) \frac{\partial F(x'\gamma_0)}{\partial \gamma_{0k}}$$

where:

$$Q_1 = wF(x'\gamma_1) + (1 - w)(1 - F(x'\gamma_1))$$

$$Q_2 = w(1 - F(x'\gamma_0)) + (1 - w)F(x'\gamma_0)$$

$$\frac{\partial f_i}{\partial \beta_{ik}} = \frac{f_i}{\sigma_i^2} (y - x'\beta_i) x_k$$

$$\frac{\partial f_i}{\partial \sigma_i^2} = \frac{f_i}{2\sigma_i^2} \left[\frac{(y - x'\beta_i)^2}{\sigma_i^2} - 1 \right]$$

$$\frac{\partial F(x'\gamma_s)}{\partial \gamma_{sk}} = \phi(x'\gamma_s) x_k$$

$i = 1,2;\ k = 1,2,\ldots,K;\ s = 0,1$

where $\phi(x'\gamma_s)$ is a standard normal probability density function. Since θ here is of dimension $(4K + 3) \times 1$, it follows that $\mathcal{I}$ is of dimension $(4K + 3) \times (4K + 3)$.

The simulation involves a large number of trials derived from a normal random variable generator. For any set of θ values, draws were made from the switching regression model and the information matrix was obtained by averaging the expressions derived from the first derivative components of the density function over the number of replications,

given that the regime is either unknown or partly known. The asymptotic variances are the corresponding diagonal elements of the inverse of the information matrix.

3.4 The Value of Imperfect Information

We derive here two sets of ratios -- asymptotic variances with regime unknown relative to regime known; and asymptotic variances with regime partly known relative to regime known. The ratios in the former are greater than or equal to those in the latter, since the presence of information, even if imperfect, improves the efficiency of parameter estimation. In addition, all the figures we will be deriving are greater than or equal to one, and the extent to which they differ from one measures the value of information, or imperfect information, as the case may be. These ratios illustrate how close to full information efficiency our estimates will be when we are faced either with no information at all or with unreliable information.

We assume we have two exogenous variables x_1 and x_2, where x_1 is a unit vector and x_2 is defined as exp $(- x_3)$, where x_3 is a standard normal random variable, which we also derive from the normal random variable generator. We essentially conduct experiments of two types here. First, for a given set of γ values which denote some information, we vary the β parameters to find out the effects of the same amount of information on the estimation efficiencies of different regime distributions. Second, for a given set of β

values, we vary the γ parameters to find out the effects of different levels of information observability about regime classification on the estimation efficiencies of a particular sample distribution.

Given arbitrary values for $\theta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22}, \sigma_1^2, \sigma_2^2, \lambda, \gamma_{11}, \gamma_{12}, \gamma_{01}, \gamma_{02})'$ (we chose $\theta = (1, 1, 2, 2, 1, 1, .5, 1, -1, 1, 1)'$ for regime partly known), we initially compare results for the ratios of asymptotic variances when we have n = 5000 and n = 20000. Although there are differences in the absolute magnitudes of the figures which range from .1 to .7 for both regime partly known and unknown, the difference in computer costs makes us opt for the smaller sample size, since the relationship among the relative magnitudes prevails.[4/] We therefore made use of a sample size of 5000 for all our experiments.

We first need to establish the non-informative case. In the previous chapter, we had discussed an implicit "informativeness" condition in the model. When $p_{11j} = 1$ and $p_{00j} = 1$, then the indicator w_j provides perfect sample

---

[4/] For the same θ values, we also compare results under both methods of solving for the information matrix. There are differences in absolute magnitudes that become smaller as the sample size increases from 5000 to 20000. Under the first method, the differences in the ratios between the two sample sizes range from .1 to .4, while under the second method, it is from .1 to .7. It is expected that as the sample size increases some more, the absolute difference between the two methods will decline. Although the absolute magnitude differences persist, the relationships among the relative magnitudes are fairly constant. This at least partially justifies our choice of method 2 for calculating the information matrix and a sample size of 5000.

separation information. Partial sample separation information is given by w_j when p_{11j} is not equal to p_{01j}, which is equivalent to the condition that $p_{11j} \neq 1 - p_{00j}$, or that $p_{11j} + p_{00j} \neq 1$. This implies that w_j provides no sample separation information at all when $p_{11j} = 1 - p_{00j}$, or $p_{11j} + p_{00j} = 1$.

In terms of our model, where p_{11j} and p_{00j} are denoted as probit functions, then the "informativeness" condition can be expressed as a simple restriction on the parameters γ_1 and γ_0, which enter our probit models of information observability. Information is not provided when $\gamma_1 = -\gamma_0$ since then, $p_{11j} + p_{00j} = 1$; that is, $F(x_j'\gamma_1) + F(x_j'\gamma_0) = F(x_j'\gamma_1) + F(-x_j'\gamma_1) = 1$, for any x_j', where $F(\ )$ is a standard normal cumulative distribution function. Combinations of parameter values where $\gamma_1 = -\gamma_0$ can be illustrated by any number of examples. A case in point where no information is provided is when $\gamma_1 = \gamma_0 = 0$. This implies that $p_{11j} = F(x_j'\gamma_1) = F(0) = .5$ and $p_{00j} = F(x_j'\gamma_0) = F(0) = .5$. This was the non-informative case we had in the previous chapter where the probabilities of regime classification were assumed constant, i.e. $p_{11} = p_{00} = .5$.

We now proceed with the first type of experiments we have to conduct, where for given γ values, we vary our β parameters. We choose $\gamma = (1, -1, 1, 1)'$ where there is some information provided, i.e. $\gamma_1 \neq -\gamma_0$. The first case is when we allow the β parameters of regime 2 to deviate

uniformly from regime 1, such that $\beta_2 = h\beta_1$ (h = 1,2,4). We hold $\sigma_1^2 = \sigma_2^2 = 1$ and $\lambda = .5$. The results are presented in Table 5. Figures in parentheses are the ratios when the regime is unknown relative to when regime is known.

When $\beta_1 = \beta_2$ so that the regression equations are the same for both regimes, the presence of information given by γ_1 and γ_0 greatly improves the efficiency of the estimates. With no information at all, the ratios go to ∞, since the samples are impossible to disentangle while the ratios are finite with some information available. An interesting observation here is that the value of sample separation information is much greater for the slopes than for the intercepts when the regime is partly known. For the case of the estimated mixing parameter, λ, the value of information for regime partly known is ∞, and for regime unknown is 0. There is no meaning that can be attached to this parameter in this instance, since the samples are difficult to distinguish from each other anyway.

The choices for β_1 and β_2 are of course restrictive. However, note that the results are invariant with regards to location and scale, as long as $\beta_1 = \beta_2$ and $\sigma_1^2 = \sigma_2^2$. $\beta_1 = (1, 0)'$, $\beta_2 = (1, 0)'$ gives the same results as $\beta_1 = (1, 1)'$, $\beta_2 = (1, 1)'$.

When $\beta_2 = h\beta_1$ ($h \neq 1$), so that the intercept and slope of one equation move away from the intercept and slope of the other equation by the same proportion, then the value of sample separation information decreases monotonically as

Table 5. Ratios of Asymptotic Variances

Varying h ($\beta_2 = h\,\beta_1$) when $\beta_1 = (1, 1)'$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$, $\gamma = (1, -1, 1, 1)'$

Parameters				Ratios: Partly Known (Unknown)/Known						
β_{11}	β_{12}	β_{21}	β_{22}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
1	0	1	0	6.6	74.2	6.7	74.2	2.3	2.5	∞
				(∞)	(∞)	(∞)	(∞)	(∞)	(∞)	(0)
1	1	1	1	6.6	74.2	6.7	74.2	2.3	2.5	∞
				(∞)	(∞)	(∞)	(∞)	(∞)	(∞)	(0)
1	1	2	2	3.6	2.7	3.4	1.8	2.0	2.2	3.0
				(4.3)	(2.8)	(4.0)	(1.8)	(2.3)	(2.3)	(3.8)
1	1	4	4	1.8	2.5	1.7	1.6	1.1	1.2	1.0
				(1.9)	(2.5)	(1.8)	(1.6)	(1.2)	(1.2)	(1.0)

n = 5000. Figures in parentheses are the ratios of asymptotic variances when regime is unknown relative to when regime is known. This is true for the other tables in this chapter.

h increases. Note the large decline in the ratios of variances for the estimated slopes as soon as the samples become distinct from each other. When the intercepts and slopes of the two equations are sufficiently far apart, the ratios when the regime is partly known or is completely unknown tend to approach one. In addition, the ratios tend to equal each other in both cases of observability so there is very little value in obtaining sample separation information or using imperfect information (when available), when the regression equations are clearly distinguishable.

The second case we consider is when the regression equations are made distinct from each other by moving the slopes away, but keeping the intercepts constant. The results are presented in Table 6. The value of sample separation information decreases monotonically, as β_{22} increases with $\beta_{12} = 0$. Again, the decline in the ratios is very steep as soon as the samples are made distinguishable, i.e. $\beta = (0, 0, 0, 0)'$ and $\beta = (0, 0, 0, 1)'$. As the slopes move farther away, the decline in the ratios is not very great, or is rather slow. As before, there is very little value in obtaining sample separation information or using imperfect information when the samples are clearly distinct, since the ratios with partial information and with no information at all tend to equalize.

The third case is when the regression equations are made distinct by moving the intercepts farther away, but keeping the slopes constant. The results are in Table 7.

Table 6. Ratios of Asymptotic Variances

Varying β_{22} when $\beta = (0, 0, 0, \beta_{22})'$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$, $\gamma = (1, -1, 1, 1)'$

Parameters				Ratios: Partly Known (Unknown)/Known						
β_{11}	β_{12}	β_{21}	β_{22}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
0	0	0	0	6.6	74.2	6.7	74.2	2.3	2.5	∞
				(∞)	(∞)	(∞)	(∞)	(∞)	(∞)	(0)
0	0	0	1	3.8	2.9	3.7	1.8	1.9	2.0	4.9
				(5.4)	(3.3)	(5.2)	(2.0)	(2.5)	(2.6)	(5.2)
0	0	0	2	2.9	2.7	1.7	1.7	1.6	1.7	2.3
				(3.4)	(2.8)	(3.1)	(1.8)	(1.8)	(1.9)	(2.5)
0	0	0	4	2.3	2.6	2.0	1.6	1.3	1.3	1.4
				(2.5)	(2.6)	(2.2)	(1.7)	(1.4)	(1.4)	(1.5)

Table 7. Ratios of Asymptotic Variances

Varying β_{21} when $\beta = (0, 0, \beta_{21}, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$, $\gamma = (1, -1, 1, 1)'$

Parameters				Ratios: Partly Known (Unknown)/Known						
β_{11}	β_{12}	β_{21}	β_{22}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
0	0	0	0	6.6	74.2	6.7	74.2	2.3	2.5	∞
				(∞)	(∞)	(∞)	(∞)	(∞)	(∞)	(0)
0	0	1	0	14.5	4.0	25.4	3.4	6.0	9.1	70.0
				(6830.4)	(4.1)	(6863.4)	(3.4)	(408.1)	(415.3)	(54638.3)
0	0	2	0	4.8	3.2	5.5	2.8	2.6	3.8	7.1
				(51.1)	(3.2)	(51.1)	(2.8)	(15.0)	(16.3)	(99.0)
0	0	4	0	2.0	2.5	2.0	2.0	1.4	1.7	1.2
				(2.3)	(2.5)	(2.2)	(2.0)	(1.8)	(2.0)	(1.3)
0	0	8	0	1.6	2.4	1.6	1.5	1.0	1.1	1.0
				(1.6)	(2.4)	(1.6)	(1.5)	(1.0)	(1.1)	(1.0)

Here, the value of sample separation information declines but the decline is not monotonic for the estimated intercepts and variances; the decline is monotonic for the slopes though. This same observation was also found in Kiefer's study (1979) of a normal mixture model. When the intercepts are close together (in this case, they are equal), wrong classification does not seriously affect the quality of the estimates. Then, as the intercepts move farther away, the effects of misclassification become more serious and the estimates suffer. When the intercepts become still farther apart, the probability of misclassification becomes so small so that the estimates become almost as efficient as estimates based on known sample separation. As the intercepts move away from each other, the decline in the ratios is more substantial, or faster as compared to the case when the intercepts are held constant, but the slopes are moved farther apart. Again, there is very little value to obtaining information when the regression equations are clearly distinct, since the ratios with partial information and with no information at all tend to equalize.

The very large values of the variance ratios for the estimated intercepts, variances and mixing parameter, when the samples are sufficiently close and when there is no information at all, seem to suggest that the intercept is a more important component of the regression equation in determining separability of the two distributions, as compared to the slope. It is more difficult to distinguish one sample from

the other when the intercepts are close together rather than when the slopes are. Compare the cases of $\beta = (0, 0, 0, 1)'$ and $\beta = (0, 0, 0, 2)'$ in Table 6 as against the cases of $\beta = (0, 0, 1, 0)'$ and $\beta = (0, 0, 2, 0)'$ in Table 7.

Note that our values for β_1 and β_2 are restrictive, but they are invariant with regards to translation, as long as the other parameters in θ are not changed. $\beta_1 = (1, 1)'$, $\beta_2 = (2, 1)'$ gives the same results as $\beta_1 = (0, 0)'$, $\beta_2 = (1, 0)'$. A related observation is that $\beta_1 = (0, 0)'$, $\beta_2 = (2, 0)'$ gives almost the same results as $\beta_1 = (0)$, $\beta_2 = (2)$ where the latter comes from a normal mixture model. The ratios in the former are slightly bigger than the ratios in the latter, since there are more parameters to estimate in the former, even if $\beta_{12} = \beta_{22} = 0$.[5/] When we estimate a normal mixture model, the ratios corresponding to $\hat{\beta}_{11}$, $\hat{\beta}_{21}$, $\hat{\sigma}_1^2$, $\hat{\sigma}_2^2$ and $\hat{\lambda}$ are 4.1 (50.4), 4.9 (50.5), 2.6 (14.9), 3.7 (16.2) and 7.0 (98.9), respectively. When regime is unknown, the ratios Schmidt (1981) derived in an earlier paper are very similar to the above figures in parentheses. The only difference is that Schmidt's ratios are smaller (i.e.

[5/] Note that when a row and column corresponding to a certain parameter is deleted, this implies that either the model does not contain this parameter, or that the parameter is part of the model but is known a priori and need not be estimated at all. In the former case, the value of information is more important when the model is more complicated, or when θ has more parameters even if both models are presented with the same amount of information in γ_1 and γ_0. In the latter case, when some parameters are known a priori and need not be estimated, resulting ratios are lower since they understate the true value of information.

41.1, 40.4, 12.7. 12.6 and 78.8 as presented in Table 1 of the previous chapter) presumably due to a larger sample size (n = 100000) so that the results are much tighter.

We now turn to the second type of experiments we will be conducting,[6/] that of varying the γ parameters given particular β values to find out the effects of different levels of observability on the efficiency of parameter estimation with fixed regression parameters.

We had earlier established that the intercept terms are more important than the slope coefficients in determining regime classification since ratios tend to be higher (the value of sample separation information is more important) when the intercepts are moved farther away, rather than when the slopes are moved apart. For this reason, we choose a β set equal to $(0, 0, 2, 0)'$ where the slopes are equal but the intercepts are different. Note that $\beta = (0, 0, 2, 0)'$ is invariant with regards to transformation to some other β forms, i.e. $\beta = (2, 2, 4, 2)'$ and $\beta = (2, 0, 4, 0)'$.

Our first case is presented in Table 8. Given γ_1, the γ_0 combinations are arranged from highest ratios (least information so most inefficient) to lowest ratios (most information so most efficient). When $\gamma_1 = -\gamma_0$, this is

---

[6/] This is the extent of our experimentation in this chapter. We will not attempt to change the variances nor the mixing parameter, since the earlier chapter had already established the results for these cases; that is, the value of sample separation information is higher for the parameters of the regime which is observed with the lower probability, and higher for the variance parameter of the sample which has the smaller variance.

Table 8. Ratios of Asymptotic Variances

Varying γ_0 when $\beta = (0, 0, 2, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$, $\gamma_1 = (1, -1)'$

Parameters				Ratios: Partly Known/Known						
γ_{11}	γ_{12}	γ_{01}	γ_{02}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
1	-1	-1	1	51.1	3.2	51.1	2.8	14.9	16.3	99.0
1	-1	0	1	11.4	3.2	11.1	2.8	4.7	5.4	19.4
1	-1	1	0	6.0	3.2	6.7	2.8	3.3	3.9	9.6
1	-1	1	1	4.8	3.2	5.5	2.8	2.6	3.8	7.1
1	1	1	-1	5.6	3.2	4.7	2.8	3.4	3.1	7.1
1	-1	1	-1	4.4	2.8	4.1	1.8	2.2	2.4	3.7
-1	1	-1	1	4.4	2.8	4.0	1.8	2.3	2.3	3.6

the case of no information and when $\gamma_1 = \gamma_0$, this is the case of the most information.

Given γ_1, the value of information is higher when we change the slope of the probit model, γ_{02} (keeping $\gamma_{01} = 0$) rather than the intercept of the probit model, γ_{01} (keeping $\gamma_{02} = 0$). This implies that the intercept term in the probit function is more important in increasing the efficiency of the parameter estimates given the available information. Given γ_{02} and γ_1, ratios are lower when γ_{01} is higher so that there is more efficiency here, and ratios are higher when γ_{01} is lower so that there is less efficiency. The transition from least information (smaller γ_{01}) to most information (larger γ_{01}) improves efficiency when γ_0 is closer to γ_1 values. The most efficient estimates occur when $\gamma_1 = \gamma_0$. This implies that the quality of estimates is best when there is equal certainty for the sample separation to be correct for both regimes.

$\gamma_1 = (1, -1)'$, $\gamma_0 = (1, -1)'$ is invariant to $\gamma_1 = (-1, 1)'$, $\gamma_0 = (-1, 1)'$. This reflects the fact that $\gamma_1^* = -\gamma_1$ and $\gamma_0^* = -\gamma_0$ result in the same value of sample separation information as did γ_1 and γ_0. This follows from the "non-informativeness" condition on γ_1 and γ_0 when $\gamma_1 = -\gamma_0$. By the same reasoning, the information reflected in γ_1 is no different from that in γ_1^* (and likewise for γ_0 and γ_0^*) when $\gamma_1^* = -\gamma_1$ and $\gamma_0^* = -\gamma_0$, when $\gamma_1 = \gamma_0$. This follows from the relationship that: $p_{11}^* + p_{11} = 1$ and $p_{00}^* + p_{00} = 1$. On the other hand, $\gamma_1 = (1, -1)'$,

$\gamma_0 = (1, 1)'$ is not invariant to $\gamma_1 = (1, 1)'$, $\gamma_0 = (1, -1)'$. That is, $\gamma_1^* = \gamma_0$ and $\gamma_0^* = \gamma_1$ do not imply the same value of sample separation information, when $\gamma_1 \neq \gamma_0$.

We examine next the case when $\gamma_{11} = -\gamma_{01}$, so that the intercept terms imply no information, and we vary the slope terms. The results are presented in Table 9. The ratios are again arranged from highest (no information) to lowest (most information). The classification probabilities implied by the γ_1 and γ_0 parameters become higher (so that ratios become lower and efficiency improves) as γ_{12} and γ_{02} assume non-zero values. Lower ratios result when γ_{02} is non-zero (keeping $\gamma_{12} = 0$) than when γ_{12} is non-zero (keeping $\gamma_{02} = 0$), since the probabilities implied by $\gamma_1 = (1, 1)'$, $\gamma_0 = (-1, 0)'$ represent a wider divergence in probabilities p_{11} and p_{00} than that given by the combination $\gamma_1 = (1, 0)'$, $\gamma_0 = (-1, 1)'$ due to the fact that the probability associated with $\gamma_1 = (1, 1)'$ is higher than that of $\gamma_0 = (-1, 1)'$. It is to be noted that the wider the difference in probabilities p_{11} and p_{00} (particularly in the intermediate range of probability values), the less the certainty. there is on information about regime classification, and it follows that the estimates will be less efficient. The exception here is the non-informative case of $p_{11} = p_{00} = .5$, where there is no difference in the probabilities but efficiency is lowest (since it is non-informative).

The additional information provided by the non-zero γ_{12} and γ_{02} parameters improves the efficiency of the

Table 9. Ratios of Asymptotic Variances

Varying γ_{12}, γ_{02} when $\beta = (0, 0, 2, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$, $\gamma = (1, \gamma_{12}, -1, \gamma_{02})'$

Parameters				Ratios Partly Known/Known						
γ_{11}	γ_{12}	γ_{01}	γ_{02}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
1	0	-1	0	51.1	3.2	51.1	2.8	14.9	16.3	99.0
1	1	-1	0	12.6	3.2	10.9	2.7	5.4	4.5	19.1
1	0	-1	1	5.4	2.9	5.2	2.4	2.5	3.2	5.8
1	1	-1	1	4.1	2.8	3.0	1.8	2.2	1.8	2.8

estimates as opposed to the case when only γ_{11} and γ_{01} are assigned non-zero values (and $\gamma_{12} = \gamma_{02} = 0$). In effect, this implies that modelling the classification probabilities in such a way that they are not constant for every observation increases the quality of the estimates, compared to the case in which these classification probabilities are fixed for all observations ($\gamma_{12} = \gamma_{02} = 0$).

We have earlier shown that when $\gamma_1 = \gamma_0$, so that the classification probabilities are equal, the ratios of asymptotic variances are lowest. This is the next case we consider, the results of which are shown in Table 10. Again, we start with the non-informative case, where $\gamma_1 = \gamma_0 = 0$. As the γ_1 and γ_0 values increase in magnitude, the classification probabilities associated with them increase too, and there is more information as the implied probabilities get higher (i.e. $p_{11} = p_{00}$ approach one). The ratios decline monotonically as the implied probabilities rise, and when these probabilities are sufficiently high, the quality of the estimates approximates that when there is perfect information, and the corresponding regimes are fully known.

The last case we evaluate is when we try to approximate the γ_1 and γ_0 values that will duplicate our results in the previous chapter, where classification probabilities were fixed. We test our model with non-constant p_{11} and p_{00} against the alternative of constant p_{11} and p_{00}, which is actually a special case of our specification. The results are in Table 11. In particular, we have $F(.8416) = p_{11} = p_{00}$

Table 10. Ratios of Asymptotic Variances

Varying γ ($\gamma_1 = \gamma_0$) when $\beta = (0, 0, 2, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$

Parameters				Ratios Partly Known/Known						
γ_{11}	γ_{12}	γ_{01}	γ_{02}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
0	0	0	0	51.1	3.2	51.1	2.8	14.9	16.3	99.0
.5	.5	.5	.5	3.8	2.7	3.3	1.8	2.2	2.2	2.9
1	1	1	1	2.4	2.6	2.2	1.6	1.5	1.5	1.5
2	2	2	2	1.8	2.4	1.9	1.6	1.1	1.2	1.1

Table 11. Ratios of Asymptotic Variances

Comparison of $F(x'\gamma_1) = F(x'\gamma_0) = .8$ and $p_{11} = p_{00} = .8$
when $\beta = (0, 0, 2, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\lambda = .5$

Parameters				Ratios Partly Known/Known						
γ_{11}	γ_{12}	γ_{01}	γ_{02}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\lambda}$
n = 5000										
0	0	0	0	51.1	3.2	51.1	2.8	14.9	16.3	99.0
.8416	0	.8416	0	5.0	3.2	4.6	2.5	2.9	3.1	5.8
.8416	n.a.	.8416	n.a.	4.9	3.0	4.6	2.3	2.9	3.1	5.8
n = 100000										
0	n.a.	0	n.a.	41.1	n.a.	40.4	n.a.	13.0	12.4	79.4
.8416	n.a.	.8416	n.a.	3.7	n.a.	3.7	n.a.	2.8	2.9	5.4

n.a. = not applicable

= .8; in terms of our probit models, $\gamma_{12} = \gamma_{02} = 0$, so that p_{11} and p_{00} are now constant for all observations.

We have two basic experiments here -- when we delete and do not delete γ_{12} and γ_{02} from the model. When they are not deleted, they are set equal to zero, but implicitly still estimated. Both results as well as the non-informative case are reported here, and we compare these figures to our earlier results patterned after the Lee and Porter model where p_{11} and p_{00} are fixed at .8 using a sample size of n = 100000.

As Table 11 shows, when γ_{12} and γ_{02} are not deleted, the resulting figures are slightly larger due most probably to the fact that we estimate more parameters in the model so that efficiencies may suffer. When we compare our model with the deleted γ parameters to our fixed probabilities specification of the earlier chapter, we observe that the ratios we derive now are larger than those we derived before. This could be due to a number of reasons. First, we now have more parameters to estimate in β, i.e. $\beta = (\beta_{11}, \beta_{12}, \beta_{21}, \beta_{22})'$ as against $\beta = (\beta_{11}, \beta_{21})'$ in the earlier chapter. Second, we now use a smaller sample size so that the resulting figures may be less tight. Lastly, we employed different methods of evaluating the information matrix in both cases. All these reasons could account for the differences in the absolute magnitudes of our ratios, although the relationships among the relative magnitudes are quite similar.

This second set of experiments we have just conducted

on varying γ for a given set of β has highlighted two main observations. First, invariance in the ratios occurs when $\gamma_1^* = -\gamma_1$ and $\gamma_0^* = -\gamma_0$ for $\gamma_1 = \gamma_0$; and when $\gamma_1^* = -\gamma_0$ and $\gamma_0^* = -\gamma_1$ for $\gamma_1 \neq \gamma_0$. There does not seem to exist any form of multiplicative or additive transformation for γ where invariance may result in the derived ratios, since any other change introduced to the γ_1 and γ_0 parameters will lead to probability changes reflected in $F(x'\gamma_1)$ and $F(x'\gamma_0)$. Second, when evaluating the γ_1 and γ_0 parameters, it is to be remembered that when γ_1 and γ_0 are closer to each other, it follows that the probabilities $F(x'\gamma_1)$ and $F(x'\gamma_0)$ are also closer. This means that there is almost equal certainty of proper sample separation into the two regimes, so that the information is quite reliable and efficiency improves. At the extreme, $\gamma_1 = \gamma_0$ and efficiency gains are highest, particularly when the probabilities implied by these parameters belong to the extreme range. At the other extreme, when $\gamma_1 = -\gamma_0$ there is no information at all in the regime classification information.

3.5 Summary

We have improved our earlier model on the value of imperfect sample separation information by allowing more exogenous variables in the switching regression model and by postulating that the classification probabilities are non-constant. As in the earlier model, all the parameters have to be estimated. The latter extension where the classification

probabilities can be modelled as probit functions is aimed at providing more information and flexibility to the model since the probabilities of regime classification are dependent on the exogenous variables at each observation.

Two basic types of experiments were conducted using simulation techniques applied on a large sample size. First, we vary the β parameters for a given information level (denoted by the γ parameters) to find out the effects on efficiency of estimation of varying the degree to which the samples are separate. Second, we vary the γ parameters for a given set of β parameters to evaluate the effects of different levels of information observability, given a particular sample distribution.

Among our findings, the following two are most important. (1) The use of information, even if imperfect, still presents large gains relative to when there is no information at all. Naturally, the more reliable the imperfect sample separation information, the greater the gains in efficiency, where the reliability of the information can be evaluated by the γ_1 and γ_0 parameters. (2) The value of imperfect sample separation information largely depends on how much alike the two samples are. When the samples are hard to distinguish from one another, then the value of any information is highest. If we consider the β parameters as denoting sample separability, the intercept parameters are more important than the slope parameters in determining how distinct the samples are from each other.

CHAPTER FOUR

THE CASE OF NON-CONSTANT REGIME CLASSIFICATION PROBABILITIES AND NON-CONSTANT SWITCHING PROBABILITIES

4.1 Introduction

In the preceding chapter, we evaluated the value of imperfect sample separation information in a switching regression model with two exogenous variables, where the probabilities of regime classification are non-constant. We argued that such a specification has its merits in the fact that more reliable information on sample separation is provided at each observation. This implies that the values of the exogenous variables do affect the chances of proper regime membership given the actual regime, so that the observed imperfect indicator of sample separation is a more accurate measure of the latent perfect indicator at each observation, when the regime classification probabilities are non-constant.

However, we assumed then that the switching probabilities were constant for all observations. That is, the probability that each observation is generated by a particular regime is fixed. We now re-formulate this assumption to take into account that the switching probabilities are non-constant, and can also be modelled as probit functions of the exogenous variables. The rationale behind this is fairly intuitive -- certain values of the exogenous variables have higher chances of being associated with observations which

are generated by a particular regime, while other values of the exogenous variables are better associated with observations generated by another regime. Therefore, the values of the explanatory variables affect the probabilities of actual regime classification (λ), and not just the probabilities of presumed regime classification given the actual regimes (p_{11} and p_{00}). While the preceding chapter explored the latter approach, we now deal with the former possibility as well as the latter.

In terms of the Lee and Porter railroad cartel stability model, the explanatory variables include: (1) a Great Lakes dummy variable which represents when the Great Lakes were made open to navigation so that the cartel faced its chief source of competition; and (2) several structural changes dummy variables which represent the entry, acquisitions and additions to existing networks in the railroad industry. When the cartel faced its main source of competition or when there were significant structural changes in the industry, we expect these events to affect the occurence of either collusive or non-collusive behavior within the cartel. This implies that these explanatory variables affect not only the probabilities of proper regime classification given the true regime (i.e. whether price wars were probably occuring or not), but also the probabilities of actual regime classification (i.e. whether price wars were really occuring or not).

We postulate here that switching probabilities or

probabilities of actual regime membership assume non-constant values for all observations, which introduces more flexibility to the model and improves the model's ability to classify observations based on the values of the explanatory variables. Our model here on non-constant switching probabilities can also be extended to consider our past models with a constant mixing parameter as a special case, so we can compare the performance of those models against the alternative of our present model.

4.2 The Model

We still maintain the basic switching regression model of the previous chapter but we now designate the switching probabilities as non-constant. Therefore, our model can be expressed as:

$$y_j = x_j' \beta_1 + u_{1j} \quad \text{with probability } \lambda_j \text{ for observation j (regime 1)} \tag{4.1}$$

$$y_j = x_j' \beta_2 + u_{2j} \quad \text{with probability } (1 - \lambda_j) \text{ for observation j (regime 2)}$$

β_1 and β_2 are (K x 1) vectors of parameters corresponding to the explanatory variables of the (K x n) matrix x. The error terms u_{1j} and u_{2j} are assumed to be independently and normally distributed with means 0 and variances σ_1^2 and σ_2^2, respectively. The non-constant switching probabilities

can be modelled as probit functions of the exogenous variables. That is,

$$\lambda_j = F(x_j'\xi)$$

$$1 - \lambda_j = 1 - F(x_j'\xi) = F(-x_j'\xi)$$

where $F(\ )$ is a standard normal cumulative distribution function, and ξ is a (K x 1) vector of parameters. This contains the constant switching probabilities model as a special case where all the elements of ξ are zero, except for that corresponding to the constant term. In the present model, we still retain the assumption of the previous chapter regarding the treatment of the regime classification probabilities as non-constant. Therefore, we have the following probit models on probability classification:

$$p_{11j} = F(x_j'\gamma_1) \quad \text{where } p_{11j} = \text{Prob}(w_j = 1/I_j = 1) \text{ for each observation } j$$

$$p_{00j} = F(x_j'\gamma_0) \quad \text{where } p_{00j} = \text{Prob}(w_j = 0/I_j = 0) \text{ for each observation } j$$

where $F(\ )$ is again the standard normal cumulative distribution function, and γ_1 and γ_0 are (K x 1) vectors of parameters. w_j is the observed dichotomous indicator which provides sample separation information, while I_j is the unobserved dichotomous indicator of actual regime classification.

When there is imperfect sample separation information, the joint density function for y_j and w_j can be re-written from (3.2) as:

$$f_j = f(y_j, w_j; \theta) \tag{4.2}$$

$$= \lambda_j f_1(y_j)(w_j p_{11j} + (1 - w_j)(1 - p_{11j})) + (1 - \lambda_j) f_2(y_j)(w_j(1 - p_{00j}) + (1 - w_j)p_{00j})$$

$$= F(x_j'\phi) f_1(y_j)(w_j F(x_j'\gamma_1) + (1 - w_j)F(-x_j'\gamma_1)) + F(-x_j'\phi) f_2(y_j)(w_j F(-x_j'\gamma_0) + (1 - w_j)F(x_j'\gamma_0))$$

where:

$$f_i(y_j) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[\frac{-(y_j - x_j'\beta_i)^2}{2\sigma_i^2}\right]$$

$$\lambda_j = \int_{-\infty}^{x_j'\phi} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v_j^2}{2}\right] dv_j$$

$$p_{11j} = \int_{-\infty}^{x_j'\gamma_1} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v_j^2}{2}\right] dv_j$$

$$p_{00j} = \int_{-\infty}^{x_j'\gamma_0} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v_j^2}{2}\right] dv_j$$

$i = 1,2;\ j = 1,\ldots,n$

$f_1(y_j)$ and $f_2(y_j)$ are normal probability density functions with means and variances given by $N(x_j'\beta_1, \sigma_1^2)$ and $N(x_j'\beta_2, \sigma_2^2)$, respectively. λ_j is represented by a probit model of the actual switching probabilities; and p_{11j} and p_{00j} are represented by probit models of the presumed classification probabilities given the actual regimes.

4.3 Derivation of Asymptotic Variances

When the regime is known, the asymptotic variances of $\sqrt{n}(\hat{\beta}_1 - \beta_1)$, $\sqrt{n}(\hat{\beta}_2 - \beta_2)$, $\sqrt{n}(\hat{\sigma}_1^2 - \sigma_1^2)$, $\sqrt{n}(\hat{\sigma}_2^2 - \sigma_2^2)$, and $\sqrt{n}(\hat{\xi} - \xi)$ are, respectively:

$$\sigma_1^2 \left(\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \lambda_j x_j x_j' \right)^{-1} ;$$

$$\sigma_2^2 \left(\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} (1 - \lambda_j) x_j x_j' \right)^{-1} ;$$

$$\frac{2\sigma_1^4}{\frac{\sum_{j=1}^{n} \lambda_j}{n}} ;$$

$$\frac{2\sigma_2^4}{\frac{\sum_{j=1}^{n} (1 - \lambda_j)}{n}} ; \text{ and}$$

$$\left(\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \frac{(\phi(x_j'\xi))^2 (x_j x_j')}{F(x_j'\xi)(1 - F(x_j'\xi))} \right)^{-1} .$$

The asymptotic variances for $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$ will reduce to the corresponding asymptotic variances given in the previous chapter if λ were constant. However, our switching probabilities are no longer constant in our present specification so that we have different values of λ_j for each observation. Since $\lambda_j = F(x_j'\xi)$ which is a probit model, then the asymptotic variance of $\hat{\xi}$ corresponds to the asymptotic variance of the parameters in a standard probit model. The above expression was derived from Judge et. al. (1980), and Ashford and Sowden (1970), where $\phi(\ )$ is a standard normal probability density function and $F(\ )$ is a standard

normal cumulative distribution function. There are no asymptotic variances for the estimated parameters $\hat{\gamma}_1$ and $\hat{\gamma}_0$ since these parameters are not relevant at all when the regimes are fully known. As in the previous chapter, the above expressions for the asymptotic variances of $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$ are derived from the inverse of the information matrix, where this information matrix corresponds to the likelihood function for the case of known regimes.

For the models where the regime is either completely unknown or is partly known, the asymptotic variances of $\sqrt{n}(\hat{\theta} - \theta)$ are derived in the same manner -- from the diagonal elements of the inverse of the information matrix. It follows that $\sqrt{n}(\hat{\theta} - \theta)$ approaches the distribution designated by the expression $N\left(0, \lim_{n\to\infty}\left(\frac{1}{n}\mathcal{I}\right)^{-1}\right)$.

The information matrix is defined by the following expression:

$$\mathcal{I} = -E\ \frac{\partial^2 \ln L}{\partial\theta\,\partial\theta'}$$

where:

$$L = \prod_{j=1}^{n} f_j$$

$$\ln L = \sum_{j=1}^{n} \ln f_j$$

Therefore, when the regime is completely unknown, f corresponds to the density function of (3.3) given as:

$$f(y_j;\ \theta) = \lambda_j f_1(y_j) + (1 - \lambda_j) f_2(y_j) \qquad (4.3)$$

θ is a $(3K + 2) \times 1$ vector of parameters, where K is the

number of explanatory variables. Therefore, $\theta = (\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \xi')'$ where $\beta_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1K})'$, $\beta_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2K})'$ and $\xi = (\xi_1, \xi_2, \ldots, \xi_K)'$. When the regime is partly known, f corresponds to the density function in (4.2) given as:

$$\begin{aligned} f(y_j, w_j; \theta) = {} & F(x_j'\xi) f_1(y_j)(w_j F(x_j'\gamma_1) + \\ & (1 - w_j)(1 - F(x_j'\gamma_1))) + \\ & (1 - F(x_j'\xi)) f_2(y_j)(w_j(1 - F(x_j'\gamma_0)) + \\ & (1 - w_j)F(x_j'\gamma_0)) \end{aligned}$$

θ is a $(5K + 2) \times 1$ vector of parameters given by $(\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \xi', \gamma_1', \gamma_0')'$ where the vectors β_1, β_2, and ξ have been defined as previously; and $\gamma_1 = (\gamma_{11}, \gamma_{12}, \ldots, \gamma_{1K})'$ and $\gamma_0 = (\gamma_{01}, \gamma_{02}, \ldots, \gamma_{0K})'$.

To facilitate comparison of the results here with those of the preceding chapter, we calculate the information matrix in the same manner using similar simulation techniques. The information matrix will be evaluated in the following way:

$$\begin{aligned} \mathcal{I} &= -E \sum_{j=1}^{n} \left[\frac{\partial^2 \ln f_j}{\partial\theta\, \partial\theta'} \right] \\ &= E \sum_{j=1}^{n} \left[\frac{1}{f_j^2} \left(\frac{\partial f_j}{\partial \theta} \right) \left(\frac{\partial f_j}{\partial \theta} \right)' \right] \end{aligned}$$

We therefore need to derive the first derivative expressions of f when regime is either unknown or partly known. For the case when regime classification is completely unknown (there is no sample separation information at all), the first derivatives of $f(y; \theta)$ (we omit the subscript j

for simplicity) with respect to θ are:

$$\frac{\partial f}{\partial \beta_{1k}} = F(x'\xi) \frac{\partial f_1}{\partial \beta_{1k}}$$

$$\frac{\partial f}{\partial \beta_{2k}} = (1 - F(x'\xi)) \frac{\partial f_2}{\partial \beta_{2k}}$$

$$\frac{\partial f}{\partial \sigma_1^2} = F(x'\xi) \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial f}{\partial \sigma_2^2} = (1 - F(x'\xi)) \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial f}{\partial \xi_k} = (f_1 - f_2) \frac{\partial F(x'\xi)}{\partial \xi_k}$$

where:

$$\frac{\partial f_i}{\partial \beta_{ik}} = \frac{f_i}{\sigma_i^2} (y - x'\beta_i) x_k$$

$$\frac{\partial f_i}{\partial \sigma_i^2} = \frac{f_i}{2\sigma_i^2} \left[\frac{(y - x'\beta_i)^2}{\sigma_i^2} - 1 \right]$$

$$\frac{\partial F(x'\xi)}{\partial \xi_k} = \phi(x'\xi) x_k$$

$i = 1,2;\ k = 1,\ldots,K$

where $\phi(\ )$ is a standard normal probability density function. Since θ is of dimension $(3K + 2) \times 1$, then $\mathcal{I}$ is of dimension $(3K + 2) \times (3K + 2)$.

For the case when regime classification is partly known due to the presence of imperfect sample separation information, the first derivatives of $f(y, w; \theta)$ with respect

to θ are:

$$\frac{\partial f}{\partial \beta_{1k}} = F(x'\xi)Q_1 \frac{\partial f_1}{\partial \beta_{1k}}$$

$$\frac{\partial f}{\partial \beta_{2k}} = (1 - F(x'\xi))Q_2 \frac{\partial f_2}{\partial \beta_{2k}}$$

$$\frac{\partial f}{\partial \sigma_1^2} = F(x'\xi)Q_1 \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial f}{\partial \sigma_2^2} = (1 - F(x'\xi))Q_2 \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial f}{\partial \xi_k} = (f_1Q_1 - f_2Q_2) \frac{\partial F(x'\xi)}{\partial \xi_k}$$

$$\frac{\partial f}{\partial \gamma_{1k}} = F(x'\xi)f_1(w - (1 - w)) \frac{\partial F(x'\gamma_1)}{\partial \gamma_{1k}}$$

$$\frac{\partial f}{\partial \gamma_{0k}} = -(1 - F(x'\xi))f_2(w - (1 - w)) \frac{\partial F(x'\gamma_0)}{\partial \gamma_{0k}}$$

where:

$$Q_1 = wF(x'\gamma_1) + (1 - w)(1 - F(x'\gamma_1))$$

$$Q_2 = w(1 - F(x'\gamma_0)) + (1 - w)F(x'\gamma_0)$$

$$\frac{\partial f_i}{\partial \beta_{ik}} = \frac{f_i}{\sigma_i^2}(y - x'\beta_i)x_k$$

$$\frac{\partial f_i}{\partial \sigma_i^2} = \frac{f_i}{2\sigma_i^2}\left[\frac{(y - x'\beta_i)^2}{\sigma_i^2} - 1\right]$$

$$\frac{\partial F(x'\xi)}{\partial \xi_k} = \phi(x'\xi)x_k$$

$$\frac{\partial F(x'\gamma_s)}{\partial \gamma_{sk}} = \phi(x'\gamma_s)\, x_k$$

$$i = 1,2;\ k = 1,\ldots,K;\ s = 0,1$$

where $\phi(\ )$ denotes a standard normal probability density function. θ is of dimension $(5K + 2) \times 1$, so that it follows that $\mathcal{I}$ is of dimension $(5K + 2) \times (5K + 2)$.

We follow the simulation techniques of the preceding chapter.[7/] Using a sample size of $n = 5000$, and faced with specific parameter values, we draw observations from the switching regression model using a normal random variable generator. We evaluate the information matrix by averaging the expressions derived from the first derivative components of the density functions, when regime is either unknown or is partly known. The asymptotic variances are the corresponding diagonal elements of the inverse of the information matrix.

4.4 The Value of Imperfect Information

We again derive here two sets of asymptotic variance ratios for each experiment -- one, with regime partly known relative to regime known; and two, with regime unknown

---

7/ In the preceding chapter, we showed that the information matrix can be evaluated in two ways. We evaluated it by the second method, using first derivative components of the appropriate density functions. In our present model, we followed the same method in order to facilitate a comparison of the simulation results. However, we can also evaluate the information matrix using the second derivative components (although we did not do this) as we did in Chapter 2. For the reader's interest, the expressions are shown in Appendix B.

relative to regime known. A comparison of these two ratios will show how much more efficient our estimates will be when we use partial information as compared to no information at all in determining sample separability. Understandably, the ratios in the former case will all be less than or equal to those in the latter case (they are equal when the partial information is not informative at all). All the ratios will, however, be greater than or equal to one (they are equal when the estimates derived are as efficient as full information estimates), and the extent to which they differ from one indicates the value of information or imperfect information, as the case may be.

We maintain the use of two exogenous variables x_1 and x_2, where x_1 is a unit vector and x_2 is equal to $\exp(-x_3)$, where components of x_3 are distributed as N(0, 1). x_3, like our dependent variable y, is derived from the normal random variable generator. The sample size is set at n = 5000. We retain the experimental conditions of the preceding chapter, in order to make comparisons later with the resulting ratios.

We conduct three sets of experiments. In the first set, for given σ_1^2, σ_2^2, and γ parameters which denote some information, we vary our β parameters in order to make sample separation more distinct. We do this twice -- first, using ξ values which imply constant switching probabilities (i.e. $\xi_2 = 0$ but estimated) and second, using ξ values which indicate non-constant switching probabilities (i.e.

$\xi_2 \neq 0$ and also estimated). In the second set of experiments, for given σ_1^2, σ_2^2, and β parameters which denote a distinct sample distribution, we vary our γ parameters to show different levels of information observability. Again, we do this twice -- where our chosen ξ values imply both constant and non-constant switching probabilities. In the last set of experiments, we vary our ξ parameters given fixed β, σ_1^2, σ_2^2, and γ values. The purpose of this last set of experiments is to find out the effects of different ξ values (all of which imply non-constant switching probabilities) on parameter estimation efficiencies.

For the first experiment, we vary the β values given $\sigma_1^2 = \sigma_2^2 = 1$, and $\gamma = (1, -1, 1, 1)'$ which is informative. Since we had earlier established that the intercept term is more important in determining sample separability than the slope term, we vary our intercept term β_{21}, holding the other intercept term fixed; therefore, we have $\beta = (0, 0, \beta_{21}, 0)'$ where the two distributions are made increasingly distinct from each other as β_{21} increases. We choose $\xi = (0, 0)'$ which essentially implies constant switching probabilities of .5, even if λ is modelled as a probit. This particular choice of ξ values enables us to test our model with non-constant switching probabilities, $\lambda = F(x'\xi) = .5$ where $\xi = (0, 0)'$ but estimated, against the alternative of constant λ (implicitly, $\lambda = F(x'\xi) = .5$, where $\xi = (0, 0)'$ but not estimated), which is actually a special case of our present specification. The results

are presented in Table 12, and the ratios here can be compared with the ratios in Table 7 of the previous chapter, where, for fixed λ, we perform the same experiments. We call this Case 1.

γ has some information, i.e. $\gamma_1 \neq -\gamma_0$, so that the variance ratios when regime is partly known are less than or equal to the variance ratios when regime is unknown. They become more equal as the two regimes become distinctly separate, meaning that there is little value in obtaining more information on sample separation when the two distributions are clearly far apart.

Compared to fixed λ, where ϵ is not estimated (as in Table 7), the ratios we derive now are slightly larger (particularly for $\hat{\beta}_{12}$, $\hat{\beta}_{22}$, and $\hat{\epsilon}$) probably due to the fact that more parameters are estimated here, or maybe simply due to randomness. But as the regimes become clearly separate, i.e. $\beta = (0, 0, 4, 0)'$, the ratios now are almost equal to those derived when λ was fixed.

The $\hat{\epsilon}_1$ and $\hat{\epsilon}_2$ variance ratios are higher than the $\hat{\lambda}$ ratios even when both imply that $\epsilon = 0$. The reason behind this is that parameter values for ϵ now have to be estimated, thereby introducing more randomness in the process, compared to the case when $\epsilon = 0$, but not estimated.

When regimes are quite close to each other, i.e. $\beta = (0, 0, 1, 0)'$, $\hat{\beta}_{12}$ and $\hat{\beta}_{22}$ variance ratios are larger than the $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$ variance ratios, a pattern very unlike that when λ was fixed. However, this observation only holds

Table 12. Ratios of Asymptotic Variances

Varying β_{21} when $\beta = (0, 0, \beta_{21}, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\xi = (0, 0)'$, $\gamma = (1, -1, 1, 1)'$

Parameters				Ratios: Partly Known (Unknown)/Known							
β_{11}	β_{12}	β_{21}	β_{22}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\xi}_1$	$\hat{\xi}_2$
0	0	1	0	17.5	33.4	25.5	27.7	6.0	9.9	77.4	213.8
				(6831.1)	(45.6)	(6864.1)	(36.5)	(408.7)	(415.3)	(54643.4)	(294.3)
0	0	2	0	5.2	7.4	5.5	4.6	2.7	3.8	7.8	8.1
				(52.3)	(7.8)	(51.7)	(4.8)	(15.0)	(16.3)	(101.3)	(8.5)
0	0	4	0	2.0	2.4	2.0	2.0	1.4	1.7	1.7	1.6
				(2.3)	(2.5)	(2.2)	(2.0)	(1.8)	(2.0)	(1.8)	(1.6)

n = 5000. Figures in parentheses are the ratios of asymptotic variances when regime is unknown relative to when regime is known. This is true for the other tables in this chapter.

when regime is partly known. When regimes are completely unknown, $\hat{\beta}_{12}$ and $\hat{\beta}_{22}$ ratios are much smaller than the $\hat{\beta}_{11}$ and $\hat{\beta}_{21}$ ratios, a pattern evident when λ was fixed. The same pattern holds, but on a smaller scale when $\beta = (0, 0, 2, 0)'$. This implies that when regimes are very close, and there is partly known information on regime classification, then there is a larger value of sample separation information when λ is not fixed, as compared to when λ is fixed.

We now repeat the previous experiment, this time choosing non-zero ζ parameters -- call this Case 2. The results are presented in Table 13. We set $\zeta = (1, -1)'$ so that the implied probability for each observation is no longer constant at .5. The mean value of the (different) λ_j is .438. This implies that there is a slightly larger probability that an observation is generated from regime 2 rather than regime 1. Consistent with the findings of past experiments, the ratios which reflect the value of sample separation information are larger for the regime which is sampled with the lower probability. Since regime 1 is sampled with the lower probability on the average, then the variance ratios of $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, and $\hat{\sigma}_1^2$ are all larger than the corresponding estimated parameter ratios of regime 2.

The only difference between Case 1 and Case 2 is in the value of the ζ parameters. In Case 1, the choice of the ζ values assure that for each observation, $\lambda = .5$; in Case 2, the choice of the ζ values assure that for each

Table 13. Ratios of Asymptotic Variances

Varying β_{21} when $\beta = (0, 0, \beta_{21}, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\ell_4 = (1, -1)'$, $\gamma = (1, -1, 1, 1)'$

Parameters				Ratios: Partly Known (Unknown)/Known							
β_{11}	β_{12}	β_{21}	β_{22}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\ell}_1$	$\hat{\ell}_2$
0	0	1	0	14.8	20.5	8.7	3.6	4.9	2.2	73.5	83.7
				(25.6)	(32.3)	(12.6)	(4.4)	(9.6)	(2.6)	(198.3)	(116.3)
0	0	2	0	6.0	8.6	4.6	2.7	2.4	2.1	8.7	11.8
				(7.7)	(11.4)	(7.7)	(3.4)	(4.0)	(2.3)	(14.6)	(14.3)
0	0	4	0	3.5	3.5	2.2	2.2	1.4	1.4	3.5	3.8
				(3.7)	(4.0)	(2.4)	(2.3)	(1.6)	(1.5)	(3.6)	(3.9)

observation, λ assumes different values, depending on the magnitude of the independent variables x that determine the value of λ, i.e. $\lambda = F(x'\zeta)$. A comparison of the ratios between Case 1 and Case 2 shows that, in the latter, the value of sample separation information varies less. That is, when the two regimes are fairly close and the value of sample separation information is important (or the ratios are high) in Case 1, the value of sample separation information is less important (or the ratios are lower) in Case 2. On the other hand, when the regimes become farther apart, the value of sample separation information in Case 1 becomes lower or the ratios of asymptotic variances tend to approach one as they should. For Case 2, the decline in the ratios is slower, so that ratios in Case 2 are higher than those obtained in Case 1, when regimes are distinctly separate. To illustrate, take the ratios for $\hat{\beta}_{11}$. In Case 1, they range from 17.5 to 2.0 (as the distributions become farther apart) when the regimes are partly known, and from 6831.4 to 2.3 when the regimes are unknown. In Case 2, they range from 14.8 to 3.5 when the regimes are partly known, and from 25.6 to 3.7 when the regimes are unknown. The same pattern holds for all the other variance ratios of the estimated parameters.

There does seem to be an advantage in postulating that the switching probabilities be non-constant rather than constant (even if the ζ parameters have to be estimated in both cases), so that the probability that an observation is generated by a particular distribution depends on the values of

the exogenous variables. However, this advantage only holds when $\zeta_2 \neq 0$.[8/] This is supported by the observation that the efficiency of the estimates does not suffer as much (variance ratios are lower) when λ is non-constant ($\zeta \neq 0$), as compared to when λ is constant ($\zeta = 0$), when the regimes are very close to each other and are hardly distinct, i.e. $\beta = (0, 0, 1, 0)'$. It is evident when regime is either partly known or unknown. As the regimes become separate, the decline in the ratios is quite slow, so that the variance ratios are actually lower when λ is constant.

Among all the ratios, the highest values belong to the estimated ζ parameters, just as in Case 1. This means that among all the parameters to be estimated, the largest efficiency losses originate from the parameters that determine the switching probabilities. This is fairly intuitive, since the efficiency of the estimates for the parameters in the two regimes are affected by the initial probability of switching regimes or of correctly matching the observations with the proper regimes; therefore, the greater burden of efficiency losses correspond to the ζ parameters, which enter the switching probability probit function. These are applicable only when regimes are difficult to distinguish from one

---

[8/] $\zeta_2 \neq 0$ basically implies that $\lambda = F(x'\zeta)$ is non-constant for all observations, while $\zeta_2 = 0$ implies that $\lambda = F(x'\zeta)$ is constant, since the effects of the variable x_2 are wiped out and are not reflected in the resulting values of λ. In our experiments, we adopted the special case of $\lambda = F(x'\zeta) = .5$, where $\zeta = (0, 0)'$ but estimated.

another. When the regimes are sufficiently apart, then the ratios of the ξ parameters are comparable in magnitude to the other ratios. Consistent with the observation in Case 1, the decline in the variance ratios is monotonic for all estimated parameters as the distributions become far apart.

In our second set of experiments, we fix $\sigma_1^2 = \sigma_2^2 = 1$ and also choose a particular sample mix, i.e. $\beta = (0, 0, 2, 0)'$. We vary our γ parameters to reflect different observability levels. We do this set of experiments twice -- Case 1, where $\xi = (0, 0)'$ and Case 2, where $\xi = (1, -1)'$. The results are presented in Tables 14 and 15, respectively.

Let us start with Case 1. In the first experiment, $\gamma_1 = -\gamma_0$, that is, the γ parameters imply that no information is provided at all, and the ratios derived here are very similar to those derived when λ is fixed for all observations, but $\xi = 0$ is not estimated (as seen in Table 8 of the previous chapter). Ratios when regime is partly known are exactly equal to those derived when regime is unknown. The only difference between the ratios derived here and those derived when $\xi = 0$ but not estimated is that the $\hat{\beta}_{12}$, $\hat{\beta}_{22}$, and $\hat{\xi}$ ratios are much higher when the regimes are close together, i.e. $\beta = (0, 0, 2, 0)'$.

When information is now introduced into the γ parameters ($\gamma_1 \neq -\gamma_0$), as in $\gamma = (1, -1, 1, 1)'$ and $\gamma = (1, 1, -1, 1)'$, then the ratios when regime is partly known are less than the ratios when regime is completely unknown. That is, the presence of sample separation information

Table 14. Ratios of Asymptotic Variances

Varying γ ($\gamma_1 \neq \gamma_0$) when $\beta = (0, 0, 2, 0)'$,

$\sigma_1 = \sigma_2 = 1$, $\xi = (0, 0)'$

Parameters				Ratios Partly Known/Known							
γ_{11}	γ_{12}	γ_{01}	γ_{02}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\xi}_1$	$\hat{\xi}_2$
1	-1	-1	1	52.4	7.8	51.1	4.8	15.0	16.3	101.4	8.5
1	-1	1	1	5.2	7.4	5.5	4.6	2.7	3.8	7.8	8.1
1	1	-1	1	4.6	2.9	3.2	1.7	2.3	1.8	4.8	1.9

presents efficiency gains as reflected in the decline of the ratios as compared to when there is no information at all. The results here can be compared to the ratios in Table 8 and Table 9 for the same parameter values of σ_1^2, σ_2^2, β, γ, and $\lambda = .5$ where $\xi = 0$ but not estimated.

$\gamma = (1, -1, 1, 1)'$ presents a wider divergence in regime classification probabilities p_{11} and p_{00}, as compared to $\gamma = (1, 1, -1, 1)'$. That is why, ratios are lower or efficiency gains are higher when p_{11} is close to p_{00} as in $\gamma = (1, 1, -1, 1)'$. The observation of the previous experiment also applies here. That is, the ratios derived when λ is fixed and $\xi = 0$ but not estimated are close, but slightly less than the ratios derived here where λ is also fixed and $\xi = 0$, but estimated. Again, the difference may be due to randomness or to the fact that more parameters have to be estimated this time.

We now explore Case 2, which is shown in Table 15. The ξ values are set differently, where $\xi = (1, -1)'$ so that the switching probabilities vary for all observations. This results in an average value of $\lambda = .438$, meaning that there is a slightly larger probability on the average that an observation is generated from regime 2 rather than regime 1. Consequently, the ratios are larger for the estimated parameters of the regime which is observed with the lower probability.

The first experiment illustrates a non-informative case, where $\gamma = (1, -1, -1, 1)'$. Therefore, ratios when

Table 15. Ratios of Asymptotic Variances

Varying γ ($\gamma_1 \neq \gamma_0$) when $\beta = (0, 0, 2, 0)'$,

$\sigma_1 = \sigma_2 = 1$, $\ell = (1, -1)'$

Parameters				Ratios Partly Known/Known							
γ_{11}	γ_{12}	γ_{01}	γ_{02}	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\ell}_1$	$\hat{\ell}_2$
1	-1	-1	1	7.7	11.4	7.7	3.4	4.0	2.3	14.6	14.3
1	-1	1	1	6.0	8.6	4.6	2.7	2.4	2.1	8.7	11.8
1	1	-1	1	6.8	6.2	3.3	2.6	3.1	1.4	11.5	7.3

regime is partly known are equal to the ratios when regime is completely unknown. The next two experiments provide informative γ choices, which essentially duplicate those in Case 1, so that the variance ratios decline when information is not denied from the model.

When the ζ values ensure that the switching probabilities are non-constant for all observations, the ratios in Case 2 vary less than those in Case 1. Even when the γ parameters are non-informative, variance ratios in Case 2 are lower than the corresponding variance ratios of Case 1. This re-enforces our earlier findings in the first set of experiments that there are efficiency advantages when we postulate that the switching probabilities be modelled as non-constant ($\zeta_2 \neq 0$). However, as information is provided on sample separation, the decline in the variance ratios is very slow or is quite minimal in Case 2. To illustrate this point, consider the $\hat{\beta}_{11}$ ratios -- in Case 1, the decline in the values ranges from 52.4 to 4.6 when information is provided, while in Case 2, the decline in the values ranges from 7.7 to 6.8 when the same γ information is provided. A similar pattern is evident for the ratios of the other parameters. Therefore, the advantages of improved efficiency associated with non-constant switching probabilities seems to occur only within that range of parameter values where information is very valuable in determining sample separability -- in this instance, when the γ parameters are non-informative.

Since $\gamma = (1, 1, -1, 1)'$ provides less divergence

in the p_{11} and p_{00} regime classification probabilities as compared to $\gamma = (1, -1, 1, 1)'$, we would expect that the ratios in the latter should be consistently higher than the ratios in the former. However, this does not hold, particularly in the case of the $\hat{\beta}_{11}$, $\hat{\sigma}_1^2$, and $\hat{\xi}_1$ variance ratios, where the decline in the ratios is not monotonic as we vary the γ values from the least informative to the more informative.

Another effect of the non-constant λ values is seen in the fact that among all the derived ratios of asymptotic variances, it is the $\hat{\xi}$ ratios which are always the highest. This implies that as information is provided on regime classification, efficiency losses associated with the ξ parameters remain quite substantial when λ is not constant for all observations. When λ is constant for all observations, but ξ parameters still have to be estimated (as in Case 1), then the ratios are much lower (when information is provided to the model) and the decline in the values of the asymptotic variance ratios is monotonic as more information is provided on sample separation.

The last set of experiments we conduct involves varying the values assumed by the ξ parameters given fixed values for σ_1^2, σ_2^2, β, and γ. The results are presented in Table 16. For these experiments, $\gamma_1 \neq -\gamma_0$, so we have informative cases. Ratios when information is partly available are less than ratios derived when there is no information available at all.

Table 16. Ratios of Asymptotic Variances

Varying ξ when $\beta = (0, 0, 2, 0)'$, $\sigma_1 = \sigma_2 = 1$, $\gamma = (1, -1, 1, 1)'$

Mean Values		Parameters		Ratios							
				Partly Known (Unknown)/Known							
λ	$1 - \lambda$	ξ_1	ξ_2	$\hat{\beta}_{11}$	$\hat{\beta}_{12}$	$\hat{\beta}_{21}$	$\hat{\beta}_{22}$	$\hat{\sigma}_1^2$	$\hat{\sigma}_2^2$	$\hat{\xi}_1$	$\hat{\xi}_2$
.5	.5	0	0	5.2	7.4	5.5	4.6	2.7	3.8	7.8	8.1
				(52.3)	(7.8)	(51.7)	(4.8)	(15.0)	(16.3)	(101.3)	(8.5)
.438	.562	1	-1	6.0	8.6	4.6	2.7	2.4	2.1	8.7	11.8
				(7.7)	(11.4)	(7.7)	(3.4)	(4.0)	(2.3)	(14.6)	(14.3)
.562	.438	-1	1	4.3	2.7	5.0	7.7	1.7	2.5	9.0	11.2
				(7.8)	(3.6)	(7.5)	(10.5)	(2.2)	(3.8)	(15.6)	(15.4)
.852	.148	.5	.5	3.5	2.1	9.6	13.4	1.8	4.6	10.2	15.9
				(12.0)	(3.3)	(24.4)	(20.5)	(3.6)	(10.8)	(31.4)	(25.2)
.962	.038	1	1	2.4	2.0	16.8	25.8	1.4	6.1	15.9	27.6
				(4.7)	(2.5)	(35.6)	(44.3)	(1.9)	(11.1)	(38.6)	(58.1)

The different ξ values suggest different average values for λ and $(1 - \lambda)$. The resulting variance ratios are consistent with the expectations that the ratios associated with the parmaters of the regime observed with the lower probability assume higher values. Therefore, as the average value of λ goes up, the ratios associated with regime 1, i.e. $\hat{\beta}_{11}$, $\hat{\beta}_{12}$, and $\hat{\sigma}_1^2$ all go down.

Regarding the ratios corresponding to the ξ parameters, the lowest ratios occur when $\lambda = 1 - \lambda = .5$; this means that the efficiency of the estimates on the parameters of the λ model is highest when there are equal probabilities for an observation to be generated by either regime. As the switching probabilities increase for any one regime, i.e. as the ξ parameter values increase absolutely, then $\hat{\xi}$ ratios also increase monotonically, implying that the efficiency of the estimates declines substantially when the switching probabilities become biased in favor of any one regime.

When $\xi = (1, -1)'$ and $\xi^* = (-1, 1)'$, then $\xi^* = -\xi$, so that $F(x'\xi) = 1 - F(-x'\xi) = 1 - F(x'\xi^*)$. Therefore, $\lambda = 1 - \lambda^*$. This transformation is a similar action to simply interchanging the names of the regimes. Consequently, $\hat{\beta}_1$ ratios for λ are simply equal to $\hat{\beta}_2$ ratios for $1 - \lambda^*$; $\hat{\sigma}_1^2$ ratios are similar to $\hat{\sigma}_2^2$ ratios when $-\xi$ is equal to ξ^*; and so on. Any difference in the values of the variance ratios may be attributed to randomness, and to differences in the information provided by the γ parameters to both regimes.

4.5 Summary

This chapter has focused on the possibility of modelling the switching probabilities as probit functions of the exogenous variables in a switching regression model. It has, however, retained the other features of the preceding chapter -- two exogenous variables, and modelling the regime classification probabilities given the true regime also as probit functions of the explanatory variables. In addition, all the parameters will have to be estimated. This expanded model is aimed at improving on the previous specification since using all the available observations on the dependent and independent variables may increase the chances of correct switching between regimes. It also serves as a better indication of the model's ability to classify observations based on the values of the exogenous variables.

Different types of experiments were conducted here. In the first two sets -- vary β given γ, and vary γ given β -- we apply both constant ($\xi = 0$) and non-constant ($\xi \neq 0$) switching probabilities, where the ξ parameters are estimated in both instances. When $\xi = 0$, we have the special case of our former model with fixed λ (implicitly, $\xi = 0$ but not estimated) and the ratios we derived previously can be compared with our present results. When $\xi \neq 0$, we can evaluate the merits of our probit model when the resulting switching probabilities are either constant ($\xi = 0$) or non-constant ($\xi \neq 0$). In the last set of experiments, we vary our ξ parameters, all $\xi \neq 0$, to find out the effects

of such an action on the resulting parameter efficiencies.

We come up with the following important findings. First, there are advantages when the switching probabilities are modelled as non-constant ($\zeta_2 \neq 0$) as compared to constant switching probabilities ($\zeta_2 = 0$ but still estimated). These advantages are in terms of greatly improved efficiency of the estimates of the parameters. However, these gains only occur during instances where information is most valuable -- when samples are hardly distinct from each other, and when the information provided by the γ parameters is not informative at all. Under these circumstances, we get smaller variance ratios when the switching probabilities are not fixed for all observations. Second, since there are more parameters to estimate in this model, a lot of randomness and variability is introduced. This may account for the fact that the ratios we derive here are slightly larger than those derived when λ was fixed ($\zeta = 0$ but not estimated). In addition, the slope variance ratios in the regression model are now larger than the intercept variance ratios in instances when the value of information is most important (as mentioned above) and for the sample which is observed with the lower probability on the average. This was not evident at all when we had a constant mixing parameter λ . Third, when we vary the ζ parameters to yield various average levels of switching probabilities, the variance ratios of the estimated parameters which correspond to the sample observed with the lower average probability are generally higher. The $\hat{\zeta}$

variance ratios also increase as the probability of observing a particular regime diverges from .5. Last, the value of imperfect sample separation information is still largely dependent on the natural separation of the two samples. Variance ratios are higher when the samples are more difficult to distinguish from each other, and they are lower when samples are far apart. Also, the use of imperfect information improves parameter estimates as compared to when no information is used at all. Naturally, the more reliable the imperfect information (as evident from the γ parameters), the better our estimates will be.

CHAPTER FIVE

CONCLUSIONS

We set out in this study with the purpose of assessing the value or importance of imperfect sample separation information in a switching regression model, where all the parameters have to be estimated, so as not to understate the true value of such information. We accomplished this by evaluating information matrices using simulation experiments over a large sample size (i.e. 100000 and 5000) in order to derive the asymptotic variances of the estimated parameters when regime is either unknown (no available sample separation information) or partly known (the available information is imperfect). These asymptotic variances are simply the corresponding diagonal elements of the inverse of the information matrix. We then solved for asymptotic variance ratios when regime is either partly known or completely unknown, relative to when regime is completely known (full sample separation information). A comparison of these two sets of ratios shows the advantage of using imperfect regime classification information relative to no information at all.

All these ratios are greater than or equal to one, and the extent to which they differ from one measures the value of information, or imperfect information, as the case may be. The higher these variance ratios, the greater is the value of regime classification information. On the other hand, variance ratios which approach the lower bound of 1.0 imply

that information is not very valuable to the model.

In the past three chapters, where we evaluated the value of imperfect sample separation information, we made variations on the basic switching regression model by postulating different assumptions about the parameter values. In Chapter 2, we examined a normal mixture model with imperfect regime classification information, where the probabilities of correct regime classification (given actual regime classification) are constant over observations. This is a straightforward extension of Schmidt's work to the Lee and Porter model with constant regime classification probabilities, p_{11} and p_{00}. Our experiments consisted of varying the regime classification probabilities, the difference between the means of the two samples, the difference between the variances of the two samples, and the mixing parameter -- each time holding the other parameter values fixed.

In Chapter 3, we added another explanatory variable into our switching regression model and further assumed that the presumed regime classification probabilities (p_{11} and p_{00}) are non-constant over observations, but are in fact, probit functions of the exogenous variables. This extension is aimed at improving the flexibility of the model and is plausible since it is highly likely that the imperfect regime classification probabilities vary from one observation to another, and that their values are affected by the exogenous variables. Our experiments consisted of varying the probit parameters of the regime classification probabilities

for a particular sample mix, and varying the sample mixes for a particular set of imperfect indicators -- each time holding the other parameter values constant.

In Chapter 4, we maintained the features of Chapter 3 but added another assumption, namely, that the switching probabilities, formerly assumed to be a constant mixing parameter for all observations, are now non-constant and can be modelled as a probit function of the explanatory variables. This extension is aimed at providing the model with a better ability to classify observations into the two regimes, by using as much information as possible at each observation. Therefore, actual regime classification probabilities as well as imperfect regime classification probabilities are modelled here as probit functions of the explanatory variables. There are three sets of experiments here: varying the probit parameters of the imperfect regime classification probabilities for a particular sample mix, and varying the sample mixes for a particular set of imperfect regime classification probabilities, each time using non-constant switching probabilities; and varying the parameters in the switching probabilities probit model given fixed values of the other parameters.

We have discussed the results of our experiments in detail already, so here we will discuss only a few of the more important findings. First, there are advantages in terms of efficiency gains when using imperfect sample separation information, as compared to no information at

all. These efficiency gains can be substantial in some cases. This is especially so when the two samples are not very distinct, so that there is not much sample separation information in the sample itself.

Our second important finding follows from the first one. There are two cases in which imperfect sample separation information does not improve efficiency of estimation:

(1) The imperfect sample separation information is not informative. This occurs when the probability of a particular observed regime classification does not depend on the true regime classification; that is, when $p_{11} = 1 - p_{00}$. In terms of the model of Chapter 3, where these probabilities are modelled as probit functions, this occurs when $\gamma_1 = -\gamma_0$.

(2) The samples are very distinct. The two distributions are sufficiently far apart so that there is a very small probability of misclassification for any observation. Therefore, there is hardly any need for information (imperfect or otherwise) in determining sample separability. This occurs when the means of the two distributions in the sample are clearly separate (μ_1 distinct from μ_2; β_1 distinct from β_2).

Our third conclusion again follows from the first. The value of imperfect sample separation information is highest, or the gains in efficiency in using unreliable information are greatest, under the following circumstances:

(1) The imperfect sample separation information is

highly informative. In the extreme case, $p_{11} = p_{00} = 1$ so that the imperfect indicators are perfect indicators, and the regime is fully identifiable based on the available information. The more reliable the imperfect indicators, the more efficient the estimates are. This occurs when $p_{11} = p_{00}$ or p_{11} near p_{00} in the extreme ranges of probability, where there is great certainty and confidence that both regime classifications are right.

(2) The samples are not very distinct. It is here where information (even if imperfect) is most helpful in determining sample separability and improving the efficiency of the estimates. This agrees with the findings of previous studies (Kiefer, 1979; Schmidt, 1981; Lee and Porter, 1984) that the value of sample separation information is largely dependent on the natural separation of the two samples. The closer the distributions in the sample (μ_1 close to μ_2; β_1 close to β_2) and the closer the variances are, the more important is information in assigning regime membership.

Fourth, it is the intercept term rather than the slope term in a switching regression model which mostly determines sample separability. It is more difficult to distinguish one sample from the other when the intercepts are close together rather than when the slopes are. Therefore, the efficiency losses in using no information or using partial information are far greater for the parameter estimates when the intercepts are hardly distinct from each other as

compared to when the slopes are hardly distinct from one another.

Fifth, the value of sample separation information is highest for the estimates corresponding to the mixing parameter or to the parameters of the switching probabilities.

Sixth, there are definite efficiency gains when we model our switching probabilities as non-constant probit functions of the explanatory variables. These gains occur in circumstances where information is most valuable; that is, when samples are hardly distinct from each other and when the imperfect regime classification information is not very informative.

Seventh, as we continually expand on our basic switching regression model, we find that regime classification information becomes more valuable. The value of sample separation information is more important for complicated models, as Kiefer (1979) suggested. This is due to the fact that as we try to estimate more parameters, more variability is introduced to the estimates, which is naturally reflected in larger variances. This notion of more variability in the model is also evident in other situations -- when the δ parameters are not very informative, when samples are difficult to disentangle, and when a particular regime is observed with a lower probability.

Eighth, in accordance with the findings of Schmidt (1981), the value of information, imperfect or otherwise, is higher for the regime which is observed with the lower

probability.

In light of these findings, a final word is warranted. Sample separation information, even if imperfect or unreliable, can be used to improve the efficiency of parameter estimates in switching regression models. Its use is most valuable when the samples are hard to disentangle from each other, and when the imperfect information is informative and fairly reliable. Under these conditions, it may also be advisable to model the switching probabilities as non-constant, since this action can further increase the efficiency of the estimates, particularly when the samples are difficult to distinguish from one another. Presumed regime classification probabilities given the actual regimes may also be modelled as non-constant to further improve the model's flexibility. However, when the imperfect information is highly unreliable or when the samples are clearly separate, there is little point in using imperfect information, since only small efficiency gains are possible. In addition, one must consider the trade-off implied when adding more parameters to the model (like imperfect regime indicator functions with probit parameters) since such an action gives the model more variability and tends to increase the variance estimates. Therefore, gains achieved by improving the model's plausibility may be lost or at least partially offset by introducing more variability into the model when additional parameters have to be estimated.

APPENDICES

APPENDIX A

THE SECOND DERIVATIVE COMPONENTS OF THE INFORMATION MATRIX IN THE CASE OF NON-CONSTANT CLASSIFICATION PROBABILITIES

The density function (we drop the subscript j for simplicity) when the regime is unknown is:

$$f(y;\theta) = \lambda f_1(y) + (1 - \lambda) f_2(y)$$

where:

$$\theta = (\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \lambda)'$$

$$\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK})'$$

$$f_i(y) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[\frac{-(y - x'\beta_i)^2}{2\sigma_i^2}\right]$$

$$i = 1,2$$

The information matrix is given by:

$$\vartheta = -E \sum_{j=1}^{n} \left[\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'}\right]$$

where:

$$\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'} = \frac{1}{f_j}\frac{\partial^2 f_j}{\partial\theta\,\partial\theta'} - \frac{1}{f_j^2}\left(\frac{\partial f_j}{\partial\theta}\right)\left(\frac{\partial f_j}{\partial\theta}\right)'$$

The first derivatives of f with respect to θ are given in the text of Chapter 3. The non-zero second derivatives of f with respect to θ are:

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\beta_{1m}} = \lambda \frac{\partial^2 f_1}{\partial\beta_{1k}\,\partial\beta_{1m}}$$

$$\frac{\partial^2 f}{\partial \beta_{1k}\, \partial \sigma_1^2} = \lambda \frac{\partial^2 f_1}{\partial \beta_{1k}\, \partial \sigma_1^2}$$

$$\frac{\partial^2 f}{\partial \beta_{1k} \partial \lambda} = \frac{\partial f_1}{\partial \beta_{1k}}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\, \partial \beta_{2m}} = (1 - \lambda) \frac{\partial^2 f_2}{\partial \beta_{2k}\, \partial \beta_{2m}}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\, \partial \sigma_2^2} = (1 - \lambda) \frac{\partial^2 f_2}{\partial \beta_{2k}\, \partial \sigma_2^2}$$

$$\frac{\partial^2 f}{\partial \beta_{2k} \partial \lambda} = - \frac{\partial f_2}{\partial \beta_{2k}}$$

$$\frac{\partial^2 f}{\partial (\sigma_1^2)^2} = \lambda \frac{\partial^2 f_1}{\partial (\sigma_1^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2 \partial \lambda} = \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial^2 f}{\partial (\sigma_2^2)^2} = (1 - \lambda) \frac{\partial^2 f_2}{\partial (\sigma_2^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2 \partial \lambda} = - \frac{\partial f_2}{\partial \sigma_2^2}$$

where:

$$\frac{\partial^2 f_i}{\partial \beta_{ik}\, \partial \beta_{im}} = \frac{f_i}{\sigma_i^2} (- x_k x_m) + \frac{\partial f_i}{\partial \beta_{im}} \frac{x_k}{\sigma_i^2} (y - x' \beta_i)$$

$$\frac{\partial^2 f_i}{\partial \beta_{ik}\, \partial \sigma_i^2} = (y - x' \beta_i)\, x_k \left[\frac{\partial f_i}{\partial \sigma_i^2} \frac{1}{\sigma_i^2} - \frac{f_i}{\sigma_i^4} \right]$$

$$\frac{\partial^2 f_i}{\partial(\sigma_i^2)^2} = \frac{\partial f_i}{\partial \sigma_i^2}\left[\frac{(y - x'\beta_i)^2}{2\sigma_i^4} - \frac{1}{2\sigma_i^2}\right] + f_i\left[\frac{1}{2\sigma_i^4} - \frac{(y - x'\beta_i)^2}{\sigma_i^6}\right]$$

$i = 1,2;\ k,m = 1,2,\ldots,K$

The joint density function (we omit the subscript j for simplicity) when the regime is partly known is:

$$f(y, w; \theta) = \lambda f_1(y)(wp_{11} + (1 - w)(1 - p_{11})) + (1 - \lambda)f_2(y)(w(1 - p_{00}) + (1 - w)p_{00})$$

where:

$$\theta = (\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \lambda, \gamma_1', \gamma_0')'$$

$$\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK})'$$

$$\gamma_s = (\gamma_{s1}, \gamma_{s2}, \ldots, \gamma_{sK})'$$

$$f_i(y) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[\frac{-(y - x'\beta_i)^2}{2\sigma_i^2}\right]$$

$$p_{11} = F(x'\gamma_1) = \int_{-\infty}^{x'\gamma_1} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v^2}{2}\right] dv$$

$$p_{00} = F(x'\gamma_0) = \int_{-\infty}^{x'\gamma_0} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v^2}{2}\right] dv$$

$i = 1,2;\ s = 0,1$

$F(\ ) =$ standard normal cumulative distribution function

The information matrix is given by:

$$\mathcal{I} = -E\sum_{j=1}^{n}\left[\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'}\right]$$

where:

$$\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'} = \frac{1}{f_j}\frac{\partial^2 f_j}{\partial\theta\,\partial\theta'} - \frac{1}{f_j^2}\left(\frac{\partial f_j}{\partial\theta}\right)\left(\frac{\partial f_j}{\partial\theta}\right)'$$

The first derivatives of f with respect to θ are given in the text of Chapter 3. The non-zero second derivatives of f with respect to θ are:

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\beta_{1m}} = \lambda Q_1 \frac{\partial^2 f_1}{\partial\beta_{1k}\,\partial\beta_{1m}}$$

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\sigma_1^2} = \lambda Q_1 \frac{\partial^2 f_1}{\partial\beta_{1k}\,\partial\sigma_1^2}$$

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\lambda} = Q_1 \frac{\partial f_1}{\partial\beta_{1k}}$$

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\gamma_{1m}} = \lambda(w - (1 - w)) \frac{\partial f_1}{\partial\beta_{1k}} \frac{\partial F(x'\gamma_1)}{\partial\gamma_{1m}}$$

$$\frac{\partial^2 f}{\partial\beta_{2k}\,\partial\beta_{2m}} = (1 - \lambda)Q_2 \frac{\partial^2 f_2}{\partial\beta_{2k}\,\partial\beta_{2m}}$$

$$\frac{\partial^2 f}{\partial\beta_{2k}\,\partial\sigma_2^2} = (1 - \lambda)Q_2 \frac{\partial^2 f_2}{\partial\beta_{2k}\,\partial\sigma_2^2}$$

$$\frac{\partial^2 f}{\partial\beta_{2k}\,\partial\lambda} = - Q_2 \frac{\partial f_2}{\partial\beta_{2k}}$$

$$\frac{\partial^2 f}{\partial\beta_{2k}\,\partial\gamma_{0m}} = - (1 - \lambda)(w - (1 - w)) \frac{\partial f_2}{\partial\beta_{2k}} \frac{\partial F(x'\gamma_0)}{\partial\gamma_{0m}}$$

$$\frac{\partial^2 f}{\partial(\sigma_1^2)^2} = \lambda Q_1 \frac{\partial^2 f_1}{\partial(\sigma_1^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2 \partial \lambda} = Q_1 \frac{\partial f_1}{\partial \sigma_1^2}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2 \partial \gamma_{1k}} = \lambda (w - (1 - w)) \frac{\partial f_1}{\partial \sigma_1^2} \frac{\partial F(x'\gamma_1)}{\partial \gamma_{1k}}$$

$$\frac{\partial^2 f}{\partial (\sigma_2^2)^2} = (1 - \lambda) Q_2 \frac{\partial^2 f_2}{\partial (\sigma_2^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2 \partial \lambda} = - Q_2 \frac{\partial f_2}{\partial \sigma_2^2}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2 \partial \gamma_{0k}} = - (1 - \lambda)(w - (1 - w)) \frac{\partial f_2}{\partial \sigma_2^2} \frac{\partial F(x'\gamma_0)}{\partial \gamma_{0k}}$$

$$\frac{\partial^2 f}{\partial \lambda \partial \gamma_{1k}} = f_1 (w - (1 - w)) \frac{\partial F(x'\gamma_1)}{\partial \gamma_{1k}}$$

$$\frac{\partial^2 f}{\partial \lambda \partial \gamma_{0k}} = f_2 (w - (1 - w)) \frac{\partial F(x'\gamma_0)}{\partial \gamma_{0k}}$$

$$\frac{\partial^2 f}{\partial \gamma_{1k} \partial \gamma_{1m}} = \lambda f_1 (w - (1 - w)) \frac{\partial^2 F(x'\gamma_1)}{\partial \gamma_{1k} \partial \gamma_{1m}}$$

$$\frac{\partial^2 f}{\partial \gamma_{0k} \partial \gamma_{0m}} = - (1 - \lambda) f_2 (w - (1 - w)) \frac{\partial^2 F(x'\gamma_0)}{\partial \gamma_{0k} \partial \gamma_{0m}}$$

where:

$$Q_1 = wF(x'\gamma_1) + (1 - w)(1 - F(x'\gamma_1))$$

$$Q_2 = w(1 - F(x'\gamma_0)) + (1 - w)F(x'\gamma_0)$$

$$\frac{\partial^2 f_i}{\partial \beta_{ik} \partial \beta_{im}} = \frac{f_i}{\sigma_i^2} (- x_k x_m) + (y - x'\beta_i) \frac{x_k}{\sigma_i^2} \frac{\partial f_i}{\partial \beta_{im}}$$

$$\frac{\partial^2 f_i}{\partial \beta_{ik}\, \partial \sigma_i^2} = (y - x'\beta_i)\, x_k \left[\frac{\partial f_i}{\partial \sigma_i^2} \frac{1}{\sigma_i^2} - \frac{f_i}{\sigma_i^4} \right]$$

$$\frac{\partial^2 f_i}{\partial (\sigma_i^2)^2} = \frac{\partial f_i}{\partial \sigma_i^2} \left[\frac{(y - x'\beta_i)^2}{2\sigma_i^4} - \frac{1}{2\sigma_i^2} \right] + f_i \left[\frac{1}{2\sigma_i^4} - \frac{(y - x'\beta_i)^2}{\sigma_i^6} \right]$$

$$\frac{\partial^2 F(x'\gamma_s)}{\partial \gamma_{sk}\, \partial \gamma_{sm}} = \phi(x'\gamma_s)(-x'\gamma_s)\, x_k x_m$$

$i = 1,2$; $s = 0,1$; $k,m = 1,2,\ldots,K$

$F(\ )$ = standard normal cumulative distribution function

$\phi(\ )$ = standard normal probability density function

APPENDIX B

THE SECOND DERIVATIVE COMPONENTS OF THE INFORMATION MATRIX IN THE CASE OF NON-CONSTANT CLASSIFICATION PROBABILITIES AND NON-CONSTANT SWITCHING PROBABILITIES

The density function (we drop the subscript j for simplicity) when the regime is unknown is:

$$f(y;\ \theta) = \lambda f_1(y) + (1 - \lambda) f_2(y)$$

where:

$$\theta = (\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \xi')'$$

$$\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK})'$$

$$\xi = (\xi_1, \xi_2, \ldots, \xi_K)'$$

$$f_i(y) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[\frac{-(y - x'\beta_i)^2}{2\sigma_i^2}\right]$$

$$\lambda = F(x'\xi) = \int_{-\infty}^{x'\xi} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v^2}{2}\right] dv$$

$i = 1,2$

$F(\ )$ = standard normal cumulative distribution function

The information matrix is given by:

$$\mathcal{J} = -E \sum_{j=1}^{n} \left[\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'}\right]$$

where:

$$\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'} = \frac{1}{f_j}\frac{\partial^2 f_j}{\partial\theta\,\partial\theta'} - \frac{1}{f_j^2}\left(\frac{\partial f_j}{\partial\theta}\right)\left(\frac{\partial f_j}{\partial\theta}\right)'$$

The first derivatives of f with respect to θ are given in the text of Chapter 4. The non-zero second derivatives of f with respect to θ are:

$$\frac{\partial^2 f}{\partial \beta_{1k}\, \partial \beta_{1m}} = F(x'\xi)\, \frac{\partial^2 f_1}{\partial \beta_{1k}\, \partial \beta_{1m}}$$

$$\frac{\partial^2 f}{\partial \beta_{1k}\, \partial \sigma_1^2} = F(x'\xi)\, \frac{\partial^2 f_1}{\partial \beta_{1k}\, \partial \sigma_1^2}$$

$$\frac{\partial^2 f}{\partial \beta_{1k}\, \partial \xi_m} = \frac{\partial f_1}{\partial \beta_{1k}}\, \frac{\partial F(x'\xi)}{\partial \xi_m}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\, \partial \beta_{2m}} = (1 - F(x'\xi))\, \frac{\partial^2 f_2}{\partial \beta_{2k}\, \partial \beta_{2m}}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\, \partial \sigma_2^2} = (1 - F(x'\xi))\, \frac{\partial^2 f_2}{\partial \beta_{2k}\, \partial \sigma_2^2}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\, \partial \xi_m} = -\frac{\partial f_2}{\partial \beta_{2k}}\, \frac{\partial F(x'\xi)}{\partial \xi_m}$$

$$\frac{\partial^2 f}{\partial (\sigma_1^2)^2} = F(x'\xi)\, \frac{\partial^2 f_1}{\partial (\sigma_1^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2\, \partial \xi_k} = \frac{\partial f_1}{\partial \sigma_1^2}\, \frac{\partial F(x'\xi)}{\partial \xi_k}$$

$$\frac{\partial^2 f}{\partial (\sigma_2^2)^2} = (1 - F(x'\xi))\, \frac{\partial^2 f_2}{\partial (\sigma_2^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2\, \partial \xi_k} = -\frac{\partial f_2}{\partial \sigma_2^2}\, \frac{\partial F(x'\xi)}{\partial \xi_k}$$

$$\frac{\partial^2 f}{\partial \xi_k\, \partial \xi_m} = (f_1 - f_2)\, \frac{\partial^2 F(x'\xi)}{\partial \xi_k\, \partial \xi_m}$$

where:

$$\frac{\partial^2 f_i}{\partial \beta_{ik}\, \partial \beta_{im}} = \frac{f_i}{\sigma_i^2}(-x_k x_m) + \frac{\partial f_i}{\partial \beta_{im}} \frac{x_k}{\sigma_i^2}(y - x'\beta_i)$$

$$\frac{\partial^2 f_i}{\partial \beta_{ik}\, \partial \sigma_i^2} = (y - x'\beta_i)x_k \left[\frac{\partial f_i}{\partial \sigma_i^2}\frac{1}{\sigma_i^2} - \frac{f_i}{\sigma_i^4}\right]$$

$$\frac{\partial^2 f_i}{\partial (\sigma_i^2)^2} = \frac{\partial f_i}{\partial \sigma_i^2}\left[\frac{(y - x'\beta_i)^2}{2\sigma_i^4} - \frac{1}{2\sigma_i^2}\right] + f_i\left[\frac{1}{2\sigma_i^4} - \frac{(y - x'\beta_i)^2}{\sigma_i^6}\right]$$

$$\frac{\partial^2 F(x'\xi)}{\partial \xi_k\, \partial \xi_m} = \phi(x'\xi)(-x'\xi)x_k x_m$$

$i = 1,2$; $k,m = 1,2,\ldots,K$

$F(\ )$ = standard normal cumulative distribution function

$\phi(\ )$ = standard normal probability density function

The joint density function (we omit the subscript j for simplicity) when the regime is partly known is:

$$f(y, w; \theta) = \lambda f_1(y)(wp_{11} + (1 - w)(1 - p_{11})) + (1 - \lambda)f_2(y)(w(1 - p_{00}) + (1 - w)p_{00})$$

where:

$$\theta = (\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \xi', \gamma_1', \gamma_0')'$$

$$\beta_i = (\beta_{i1}, \beta_{i2}, \ldots, \beta_{iK})'$$

$$\gamma_s = (\gamma_{s1}, \gamma_{s2}, \ldots, \gamma_{sK})'$$

$$\xi = (\xi_1, \xi_2, \ldots, \xi_K)'$$

$$f_i(y) = \frac{1}{\sqrt{2\pi}\,\sigma_i} \exp\left[\frac{-(y - x'\beta_i)^2}{2\sigma_i^2}\right]$$

$$\lambda = F(x'\xi) = \int_{-\infty}^{x'\xi} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v^2}{2}\right] dv$$

$$p_{11} = F(x'\gamma_1) = \int_{-\infty}^{x'\gamma_1} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v^2}{2}\right] dv$$

$$p_{00} = F(x'\gamma_0) = \int_{-\infty}^{x'\gamma_0} \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-v^2}{2}\right] dv$$

$i = 1,2$; $s = 0,1$

$F(\ )$ = standard normal cumulative distribution function

The information matrix is given by:

$$\mathcal{I} = -E \sum_{j=1}^{n} \left[\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'}\right]$$

where:

$$\frac{\partial^2 \ln f_j}{\partial\theta\,\partial\theta'} = \frac{1}{f_j}\frac{\partial^2 f_j}{\partial\theta\,\partial\theta'} - \frac{1}{f_j^2}\left(\frac{\partial f_j}{\partial\theta}\right)\left(\frac{\partial f_j}{\partial\theta}\right)'$$

The first derivatives of f with respect to θ are given in the text of Chapter 4. The non-zero second derivatives of f with respect to θ are:

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\beta_{1m}} = F(x'\xi)Q_1 \frac{\partial^2 f_1}{\partial\beta_{1k}\,\partial\beta_{1m}}$$

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\sigma_1^2} = F(x'\xi)Q_1 \frac{\partial^2 f_1}{\partial\beta_{1k}\,\partial\sigma_1^2}$$

$$\frac{\partial^2 f}{\partial\beta_{1k}\,\partial\xi_m} = Q_1 \frac{\partial f_1}{\partial\beta_{1k}} \frac{\partial F(x'\xi)}{\partial\xi_m}$$

$$\frac{\partial^2 f}{\partial \beta_{1k}\,\partial \gamma_{1m}} = F(x'\xi)(w - (1 - w))\,\frac{\partial f_1}{\partial \beta_{1k}}\,\frac{\partial F(x'\gamma_1)}{\partial \gamma_{1m}}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\,\partial \beta_{2m}} = (1 - F(x'\xi))Q_2\,\frac{\partial^2 f_2}{\partial \beta_{2k}\,\partial \beta_{2m}}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\,\partial \sigma_2^2} = (1 - F(x'\xi))Q_2\,\frac{\partial^2 f_2}{\partial \beta_{2k}\,\partial \sigma_2^2}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\,\partial \xi_m} = - Q_2\,\frac{\partial f_2}{\partial \beta_{2k}}\,\frac{\partial F(x'\xi)}{\partial \xi_m}$$

$$\frac{\partial^2 f}{\partial \beta_{2k}\,\partial \gamma_{0m}} = - (1 - F(x'\xi))(w - (1 - w))\,\frac{\partial f_2}{\partial \beta_{2k}}\,\frac{\partial F(x'\gamma_0)}{\partial \gamma_{0m}}$$

$$\frac{\partial^2 f}{\partial (\sigma_1^2)^2} = F(x'\xi)Q_1\,\frac{\partial^2 f_1}{\partial (\sigma_1^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2\,\partial \xi_k} = Q_1\,\frac{\partial f_1}{\partial \sigma_1^2}\,\frac{\partial F(x'\xi)}{\partial \xi_k}$$

$$\frac{\partial^2 f}{\partial \sigma_1^2\,\partial \gamma_{1k}} = F(x'\xi)(w - (1 - w))\,\frac{\partial f_1}{\partial \sigma_1^2}\,\frac{\partial F(x'\gamma_1)}{\partial \gamma_{1k}}$$

$$\frac{\partial^2 f}{\partial (\sigma_2^2)^2} = (1 - F(x'\xi))Q_2\,\frac{\partial^2 f_2}{\partial (\sigma_2^2)^2}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2\,\partial \xi_k} = - Q_2\,\frac{\partial f_2}{\partial \sigma_2^2}\,\frac{\partial F(x'\xi)}{\partial \xi_k}$$

$$\frac{\partial^2 f}{\partial \sigma_2^2\,\partial \gamma_{0k}} = - (1 - F(x'\xi))(w - (1 - w))\,\frac{\partial f_2}{\partial \sigma_2^2}\,\frac{\partial F(x'\gamma_0)}{\partial \gamma_{0k}}$$

$$\frac{\partial^2 f}{\partial \xi_k\,\partial \xi_m} = (f_1 Q_1 - f_2 Q_2)\,\frac{\partial^2 F(x'\xi)}{\partial \xi_k\,\partial \xi_m}$$

$$\frac{\partial^2 f}{\partial \xi_k \partial \gamma_{1m}} = f_1(w - (1 - w)) \frac{\partial F(x'\xi)}{\partial \xi_k} \frac{\partial F(x'\gamma_1)}{\partial \gamma_{1m}}$$

$$\frac{\partial^2 f}{\partial \xi_k \partial \gamma_{0m}} = f_2(w - (1 - w)) \frac{\partial F(x'\xi)}{\partial \xi_k} \frac{\partial F(x'\gamma_0)}{\partial \gamma_{0m}}$$

$$\frac{\partial^2 f}{\partial \gamma_{1k} \partial \gamma_{1m}} = F(x'\xi) f_1(w - (1 - w)) \frac{\partial^2 F(x'\gamma_1)}{\partial \gamma_{1k} \partial \gamma_{1m}}$$

$$\frac{\partial^2 f}{\partial \gamma_{0k} \partial \gamma_{0m}} = -(1 - F(x'\xi)) f_2(w - (1 - w)) \frac{\partial^2 F(x'\gamma_0)}{\partial \gamma_{0k} \partial \gamma_{0m}}$$

where:

$$Q_1 = wF(x'\gamma_1) + (1 - w)(1 - F(x'\gamma_1))$$

$$Q_2 = w(1 - F(x'\gamma_0)) + (1 - w)F(x'\gamma_0)$$

$$\frac{\partial^2 f_i}{\partial \beta_{ik} \partial \beta_{im}} = \frac{f_i}{\sigma_i^2}(-x_k x_m) + \frac{\partial f_i}{\partial \beta_{im}} \frac{x_k}{\sigma_i^2}(y - x'\beta_i)$$

$$\frac{\partial^2 f_i}{\partial \beta_{ik} \partial \sigma_i^2} = (y - x'\beta_i)x_k \left[\frac{\partial f_i}{\partial \sigma_i^2} \frac{1}{\sigma_i^2} - \frac{f_i}{\sigma_i^4} \right]$$

$$\frac{\partial^2 f_i}{\partial (\sigma_i^2)^2} = \frac{\partial f_i}{\partial \sigma_i^2} \left[\frac{(y - x'\beta_i)^2}{2\sigma_i^4} - \frac{1}{2\sigma_i^2} \right] + f_i \left[\frac{1}{2\sigma_i^4} - \frac{(y - x'\beta_i)^2}{\sigma_i^6} \right]$$

$$\frac{\partial^2 F(x'\xi)}{\partial \xi_k \partial \xi_m} = \phi(x'\xi)(-x'\xi)x_k x_m$$

$$\frac{\partial^2 F(x'\gamma_s)}{\partial \gamma_{sk} \partial \gamma_{sm}} = \phi(x'\gamma_s)(-x'\gamma_s)x_k x_m$$

$i = 1,2;\ s = 0,1;\ k,m = 1,2,\ldots,K$

$F(\ )$ = standard normal cumulative distribution function

$\phi(\ )$ = standard normal probability density function

BIBLIOGRAPHY

BIBLIOGRAPHY

Ashford, J.R. and Sowden, R.R. "Multivariate Probit Analysis." Biometrics, 1970, Volume 26, 535-546.

Eaton, Jonathan and Gersovitz, Mark. "LDC Participation in International Financial Markets: Debt and Reserves." Journal of Development Economics, 1980, Volume 7, 3-21.

Fair, R.C. and Jaffee, D.M. "Methods of Estimation for Markets in Disequilibrium." Econometrica, 1972, Volume 40, 497-514.

Gersovitz, Mark. "Classification Probabilities for the Disequilibrium Model." Journal of Econometrics, 1980, Volume 14, 239-246.

Goldfeld, Stephen and Quandt, Richard. "Estimation in a Disequilibrium Model and the Value of Information." Journal of Econometrics, 1975, Volume 3, 325-348.

Hamermesh, Daniel. "Wage Bargains, Threshold Effects, and the Phillips Curve." Quarterly Journal of Economics, 1970, Volume 84, 501-517.

Hartley, Michael. "Comment on "Estimating Mixtures of Normal Distributions and Switching Regressions"." Journal of the American Statistical Association, 1978, Volume 73, 738-741.

Judge, George G., Griffiths, William E., Carter Hill, R., and Lee, Tsoung-Chao. The Theory and Practice of Econometrics. 1980, New York: John Wiley and Sons, Inc.

Kendall, Maurice and Stuart, Alan. The Advanced Theory of Statistics, Volume 1. 1963, New York: Hafner.

Kiefer, Nicholas. "Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model." Econometrica, 1978, Volume 46, 427-434.

_____. "On the Value of Sample Separation Information." Econometrica, 1979, Volume 47, 997-1003.

_____. "A Note on Regime Classification in Disequilibrium Models." Review of Economic Studies, 1980, Volume 47, 637-639.

Laffont, Jean-Jacques and Garcia, Rene. "Disequilibrium Econometrics for Business Loans." Econometrica, 1977, Volume 45, 1187-1204.

Lee, Lung-Fei and Porter, Richard. "Switching Regression Models with Imperfect Sample Separation Information -- with an Application on Cartel Stability." Econometrica, 1984, Volume 52, 391-418.

Quandt, Richard. "A New Approach to Estimating Switching Regression Models." Journal of the American Statistical Association, 1972, Volume 67, 306-310.

_____ and Ramsey, James. "Estimating Mixtures of Normal Distributions and Switching Regressions." Journal of the American Statistical Association, 1978, Volume 73, 730-738.

Rosen, Harvey and Quandt, Richard. "Estimation of a Disequilibrium Aggregate Labor Market." Review of Economic Statistics, 1978, Volume 60, 371-379.

Schmidt, Peter. "Further Results on the Value of Sample Separation Information." Econometrica, 1981, Volume 49, 1339-1343.

_____. "An Improved Version of the Quandt-Ramsey MGF Estimator for Mixtures of Normal Distributions and Switching Regressions." Econometrica, 1982, Volume 50, 501-516.

Suits, Daniel. "An Econometric Model of the Watermelon Market." Journal of Farm Economics, 1955, Volume 37, 237-251.