SWITCHING REGRESSION MODELS WITH IMPERFECT SAMPLE SEPARATION INFORMATION - WITH AN APPLICATION ON CARTEL STABILITY

by

Lung-Fei Lee and Robert H. Porter

Discussion Paper No. 82-165, August 1982

Center for Economic Research
Department of Economics
University of Minnesota
Minneapolis, Minnesota 55455

Switching Regression Models with Imperfect Sample Separation Information

- With An Application on Cartel Stability

by

Lung-Fei Lee and Robert H. Porter

Department of Economics, University of Minnesota

July 1982

Abstract

An exogenous switching regression model with imperfect regime classification information is specified and applied to a study of cartel stability. An efficient estimation method is proposed which takes this imperfect information into account. The consequences of misclassification are analyzed. The direction of the least squares bias is derived. An optimal regime classification rule is obtained and compared theoretically and empirically with other classification rules. We then examine the Joint Executive Committee, a railroad cartel in the 1880s. The econometric evidence indicates that reversions to noncooperative behavior did occur for the firms in our sample, and these reversions involve a significant decrease in market price.

Address: Lung-Fei Lee
Department of Economics
University of Minnesota
1035 Management and Economics
271 - 19th Avenue South
Minneapolis, MN 55455

Switching Regression Models with Imperfect Sample Separation Information - with An Application on Cartel Stability

by

Lung-Fei Lee and Robert H. Porter(*)

1. Introduction

This article is concerned with the possibility of estimating a switching regression model and its application to a study on cartel stability. The switching regression model is the exogenous switching model proposed by Quandt [1972]. The regime switching problem in Quandt [1972] generalized a problem of mixture distributions (Day[1969]). The sample in this model is generated from distinct regression equations or regimes for each time period. If the investigator has a priori information on how the sample is partitioned into the underlying regimes, it is a switching regression model with known sample separation; otherwise, it is a model with unknown sample separation. The estimation of these models has been considered in Quandt [1972], Goldfeld and Quandt [1972], and Kiefer [1978, 1980], among others. The switching regression model is appropriate for the study of cartel behavior when there are price wars, as the firms will change from cooperative behavior to noncooperative behavior. This model will allow us to exploit the fact that there will be periodic and stochastic switches or reversions between cooperative and noncooperative structures.

Our econometric model, however, is different from the switching models in the econometrics literature in that there is additional imperfect sample separation (or regime classification) information available. In this article, we will generalize the switching regression model to take into account this extra information. We will analyze its potential use in the efficient estimation of the unknown parameters and the extra information that it can provide in the classification of the sample periods into different regimes. The consequences of misclassification, when this information is regarded as a perfect sample separation indicator, will be analyzed. This specification analysis should have implications on the consequences of misclassification in other models with switching; in particular, the disequilibrium market models of Fair and Jaffee [1972]. For the disequilibrium market models, many of the empirical studies assume that the direction of price changes provides perfect sample separation. This assumption is, however, unreliable and the consequences of misclassification of the sample for parameter estimation should not be neglected.

This model formulation and the estimation methods will be employed to analyze the behavior of the Joint Executive Committee railroad cartel with weekly time series data and to test the proposition that observed price ward represented a switch from collusive to noncooperative behavior.

2. A Switching Regression Model with An Imperfect Regime Classification Indicator

The classical switching regression model specified in Quandt [1972] consists of two regression equations,

$$y_t = x_{1t}\beta_1 + \varepsilon_{1t} \qquad \text{regime 1} \tag{2.1}$$

and

$$y_t = x_{2t}\beta_2 + \varepsilon_{2t}, \qquad \text{regime 2} \tag{2.2}$$

for $t = 1,\ldots,T$. The vectors x_{1t} and x_{2t} are exogenous variables. The observed dependent variable y_t in each period t is generated either from regime 1 or from regime 2, but never both. The probability that the observation y_t is generated from regime 1 is assumed to be a constant $\lambda, \lambda \in (0,1)$. In some cases, one may know exactly whether the observed sample y_t is generated from regime 1 or regime 2 for each t. These are the cases of known sample separation. When sample separation information is available, each of the equations can, of course, be estimated by standard methods such as ordinary least squares (OLS). When sample separation is unknown, maximum likelihood estimation methods have been suggested by Goldfeld and Quandt [1972], Hartley [1978], and Kiefer [1980]. Other estimation methods based on moment generating functions are investigated in Quandt and Ramsey [1978] and Schmidt [1982].

In this article, we will consider the above switching regression model when some sample separation information is available, but this information is imperfect. Specifically, we suppose that there is a dichotomous indicator w_t for each t, which provides some sample separation information. For each period t, we define a latent dichotomous variable I_t where $I_t = 1$ if the sample y_t is generated from regime 1; $I_t = 0$, otherwise. Thus w_t is a measure of I_t with measurement error. The measurement error of I_t is assumed to be independent of ε_{1t} and ε_{2t} or, equivalently, conditional on I_t, w_t is independent of ε_{1t} and ε_{2t}. The relation between w_t and I_t can best be described by a transition probability matrix

$$\begin{array}{c} \\ I=1 \\ I=0 \end{array} \begin{array}{c} \begin{array}{cc} w=1 & w=0 \end{array} \\ \begin{pmatrix} p_{11} & p_{10} \\ p_{01} & p_{00} \end{pmatrix} \end{array}$$

where $p_{11} = \text{Prob}(w_t = 1 | I_t = 1)$, $p_{01} = \text{Prob}(w_t = 1 | I_t = 0)$, $p_{10} = 1 - p_{11}$ and $p_{00} = 1 - p_{01}$. Let $p = \text{Prob}(w_t = 1)$. Since $\text{Prob}(I_t = 1) = \lambda$, it follows that $p = \lambda p_{11} + (1 - \lambda) p_{01}$.

In addition, we assume that ε_{1t} and ε_{2t} are normally distributed, $N(0, \sigma_1^2)$ and $N(0, \sigma_2^2)$, respectively, and, conditional on I_t, the triple $(\varepsilon_{1t}, \varepsilon_{2t}, w_t)$ is independently and identically distributed (i.i.d.) for different t. The i.i.d. assumption rules out the possibility of serial dependence. Serial correlation can not be handled easily in switching regression models without perfect sample separation information. Let

$$f_i(y_t) = \frac{1}{(2\Pi)^{\frac{1}{2}} \sigma_i} \exp\{-\frac{1}{2\sigma_i^2} (y_t - x_{it}\beta_i)^2\} \tag{2.3}$$

be the probability density function for y_t in regime i,

$i=1,2$. The joint density function for y_t and w_t is

$$f(y_t,w_t)=f_1(y_t)\text{Prob}(w_t,I_t=1)+f_2(y_t)\text{Prob}(w_t,I_t=0)$$

Since $\text{Prob}(w_t=1,\ I_t=1)=\lambda p_{11}$, $\text{Prob}(w_t=0,\ I_t=1)\lambda(1-p_{11})$, $\text{Prob}(w_t=1,\ I_t=0)=(1-\lambda)p_{01}$ and $\text{Prob}(w_t=0,\ I_t=0)=(1-\lambda)(1-p_{01})$, $\text{Prob}(w_t,I_t=1)=w_t\ p_{11}+(1-w_t)\lambda(1-p_{11})$ and $\text{Prob}(w_t,I_t=0)=w_t\cdot(1-\lambda)p_{01}+(1-w_t)(1-\lambda)(1-p_{01})$. Hence the joint density function for y_t and w_t is

$$\begin{aligned} f(y_t,w_t)&=f_1(y_t)[w_t\lambda p_{11}+(1-w_t)\lambda(1-p_{11})]+f_2(y_t)[w_t(1-\lambda)p_{01}\\ &\qquad+(1-w_t)(1-\lambda)(1-p_{01})]\\ &=[f_1(y_t)\lambda p_{11}+f_2(y_t)(1-\lambda)p_{01}]^{w_t}\cdot[f_1(y_t)\lambda(1-p_{11})+f_2(y_t)(1-\lambda)(1-p_{01})]^{1-w_t} \end{aligned} \tag{2.4}$$

On the other hand, the marginal density function for y_t is

$$g(y_t)=f_1(y_t)\lambda+f_2(y_t)(1-\lambda) \tag{2.5}$$

which is a mixture of the density functions $f_1(y_t)$ and $f_2(y_t)$.

The regime classification indicator w_t conveys some information on sample separation if $p_{11}\neq p_{01}$. When $p_{11}=p_{01}$, $\text{Prob}(w_t|I_t)=\text{Prob}(w_t)$ and the joint density function $f(y_t,w_t)=g(y_t)\cdot[w_tp+(1-w_t)(1-p)]$. Thus when $p_{11}=p_{01}$, the indicators w_t do not contain any information about the equations in (2.1) and (2.2). The model without any sample separation information can be regarded as a special case in our general framework. On the other hand, when $p_{11}=1$ and $p_{01}=0$, the indicator w_t provides perfect sample separation and $f(y_t,w_t)=[w_tf_1(y_t)+(1-w_t)f_2(y_t)]\cdot[w_t\lambda+(1-w_t)(1-\lambda)]$.

When the regression functions (2.1) and (2.2) are identical except for the constant terms, we have a regression equation with an unobserved dichotomous variable I_t;

$$y_t = x_t\beta + \delta I_t + \varepsilon_t \tag{2.6}$$

and w_t is a measure of I_t with measurement error. This errors-in-variables model has been studied by Aigner [1973] and Mouchart [1977]. The specification of the measurement relation between w_t and I_t by a transition matrix is exactly the specification originated in Aigner [1973]. Aigner shows that if some exogenous information (or estimate) is available for the covariance between the regressor and the error of observation, a consistent estimator can be derived by modifying the usual normal equations of ordinary least squares. Such information, however, is rarely available in practice. In a subsequent work, Mouchart [1977] provides a Bayesian approach.[1/]

3. Least Squares Bias

When the regime classification indicator w_t is not perfect, i.e., $p_{11} \neq 1$ or $p_{01} \neq 0$, but is regarded as a perfect regime classification indicator, we have a misclassification problem. Misclassification of the sample $\{y_t\}$ into different regimes may cause bias and inconsistency problems in estimation. Without loss of generality, suppose that the first T_1 observations of the sample $\{y_t\}$ are classified to regime 1 and the remaining observations to regime 2. Each of the equations (2.1) and (2.2) can then be estimated by the method of maximum likelihood or least squares. Since the distributions of ε_1 and ε_2 are normal, the least squares estimates (OLS) and the maximum likelihood estimates (MLE) are the same after a degree of freedom adjustment of the MLE of the variances σ_1^2 and σ_2^2.

Without loss of generality, we will analyze the (asymptotic) bias of the OLS of the regression equation in regime 1. The OLS of β_1 is

$$\hat{\beta}_1 = \left(\sum_{t=1}^{T_1} x_{1t}' x_{1t}\right)^{-1} \sum_{t=1}^{T_1} x_{1t}' y_t$$

$$= \left(\frac{1}{T}\sum_{t=1}^{T} w_t x_{1t}' x_{1t}\right)^{-1} \frac{1}{T}\sum_{t=1}^{T} w_t x_{1t}' y_t \qquad (3.1)$$

We note that $y_t = I_t(x_{1t}\beta_1 + \varepsilon_{1t}) + (1 - I_t)(x_{2t}\beta_2 + \varepsilon_{2t})$ and therefore

$$\hat{\beta}_1 = \left(\frac{1}{T}\sum_{t=1}^{T} w_t x_{1t}' x_{1t}\right)^{-1} \left(\frac{1}{T}\sum_{t=1}^{T} w_t I_t x_{1t}' (x_{1t}\beta_1 + \varepsilon_{1t}) + \frac{1}{T}\sum_{t=1}^{T} w_t (1 - I_t) x_{1t}' (x_{2t}\beta_2 + \varepsilon_{2t})\right)$$

Since, conditional on I_t, ε_{1t} and w_t are independent and $E(\varepsilon_{1t}|I_t)=0$, $\frac{1}{T}\sum_{t=1}^{T} w_t I_t x_{1t}' \varepsilon_{1t}$ and $\frac{1}{T}\sum_{t=1}^{T} w_t(1-I_t)x_{1t}'\varepsilon_{2t}$ converge to zero in probability under general regularity conditions on the exogenous variables, e.g., x_{1t} and x_{2t} are uniformly bounded for all t. It follows that

$$\operatorname*{plim}_{T\to\infty} \hat{\beta}_1 = \left(\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} w_t x_{1t}' x_{1t}\right)^{-1}\left(\operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} w_t I_t x_{1t}' x_{1t}\beta_1 + \operatorname*{plim}_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T} w_t(1-I_t)x_{1t}' x_{2t}\beta_2\right) \quad (3.2)$$

Suppose x_{1t} and x_{2t} are stochastic regressors, and let $\Sigma_{11}=E(x_{1t}'x_{1t})$ and $\Sigma_{12}=E(x_{1t}'x_{2t})$ denote the second order moments. Equation (3.2) becomes

$$\begin{aligned}\operatorname*{plim}_{T\to\infty} \hat{\beta}_1 &= (p\Sigma_{11})^{-1}[\lambda p_{11}\Sigma_{11}\beta_1 + (1-\lambda)p_{01}\Sigma_{12}\beta_2] \\ &= \frac{\lambda p_{11}}{p}\beta_1 + \frac{(1-\lambda)p_{01}}{p}\Sigma_{11}^{-1}\Sigma_{12}\beta_2\end{aligned}$$

and the (asymptotic) bias is

$$\operatorname*{plim}_{T\to\infty} \hat{\beta}_1 - \beta_1 = \frac{(1-\lambda)p_{01}}{p}(\Sigma_{11}^{-1}\Sigma_{12}\beta_2 - \beta_1) \quad (3.3)$$

The bias is proportional to the vector $\Sigma_{11}^{-1}\Sigma_{12}\beta_2 - \beta_1$, and the proportionality factor $\frac{(1-\lambda)p_{01}}{p}$ is the conditional probability of misclassifying $I_t=0$ into $w_t=1$ given that $w_t=1$. More analytical results can be derived for some special cases. Consider the case where $x_{1t}=x_{2t}$ for all t, i.e., equations (2.1) and (2.2) have the same regressors. The bias in (3.3) becomes

$$\operatorname*{plim}_{T\to\infty} \hat{\beta}_1 - \beta_1 = (1-\lambda)\frac{p_{01}}{p}(\beta_2 - \beta_1) \tag{3.4}$$

and

$$\operatorname*{plim}_{T\to\infty} \hat{\beta}_1 = \frac{\lambda p_{11}}{p}\beta_1 + \frac{(1-\lambda)p_{01}}{p}\beta_2 \tag{3.5}$$

The estimate $\hat{\beta}_1$ converges in probability to a weighted average of β_1 and β_2 and the weights are the conditional probabilities of classifying I_t into $w_t = 1$ given that $w_t = 1$. The larger the probability p_{01} of misclassifying regime 2 into regime 1, the larger the bias will be. Let β_{1j} and β_{2j} denote the jth component of the vectors β_1 and β_2. If the parameters β_{1j} and β_{2j} are the same, the estimate $\hat{\beta}_{1j}$ will be consistent. It will be biased upward if $\beta_{2j} > \beta_{1j}$ and biased downward if $\beta_{2j} < \beta_{1j}$. When there is no misclassification of regime 2 into regime 1, i.e., $p_{01} = 0$, there will be, of course, no bias for $\hat{\beta}_1$, independent of whether $p_{11}=1$.

The OLS of σ_1^2 is

$$\hat{\sigma}_1^2 = \left(\sum_{t=1}^{T} w_t - k\right)^{-1} \sum_{t=1}^{T} w_t (y_t - x_{1t}\hat{\beta}_1)^2 \tag{3.6}$$

where k is the dimension of the vector x_{1t}. Its (asymptotic) bias can be analyzed in a similar fashion. Since $\sum_{t=1}^{T} w_t(y_t - x_{1t}\hat{\beta}_1)^2$ $= \sum_{t=1}^{T} w_t(y_t - x_{1t}\hat{\beta}_1)y_t = \sum_{t=1}^{T} w_t(I_t y_{1t} + (1-I_t)y_{2t} - x_{1t}\hat{\beta}_1)\cdot$ $(I_t y_{1t} + (1-I_t)y_{2t})$,

$$p\cdot\operatorname*{plim}_{T\to\infty} \hat{\sigma}_1^2 = \operatorname*{plim}_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} (w_t I_t x_{1t}\beta_1 + w_t(1-I_t)x_{2t}\beta_2)^2$$

in Goldfeld and Quandt [1976], and disequilibrium market models, see, e.g., Maddala and Nelson [1974]. The bias in equation (3.4) holds for two stage least squares estimation of structural equations in cases where the explanatory variables in the corresponding structural equations are the same. The proof is similar and is omitted here. In the disequilibrium market models literature, sample separation information is usually unavailable. However, in many empirical studies, the investigators employ the sign of the first difference of the price time series to classify the sample into periods of excess demand or excess supply. This procedure may create a misclassification problem which should not be neglected.

4. Maximum Likelihood Estimation and Asymptotic Efficiency

Given the sample $\{(y_t, w_t) | t=1,\ldots,T\}$, the likelihood function is

$$L = \prod_{t=1}^{T} f(y_t, w_t)$$

$$= \prod_{t=1}^{T} [f_1(y_t)\lambda p_{11} + f_2(y_t)(1-\lambda)p_{01}]^{w_t} [f_1(y_t)\lambda(1-p_{11}) + f_2(y_t)(1-\lambda)(1-p_{01})]^{1-w_t} \tag{4.1}$$

Estimation of switching regression models without any sample separation information has been considered in Goldfeld and Quandt [1972], Quandt and Ramsey [1978], Hartley [1978], Kiefer [1980] and Schmidt [1982]. In this section, we consider the estimation of our model by maximum likelihood methods. The E-M algorithm described in Hartley [1978] and Kiefer [1980] will be extended for the estimation of our model. The switching regression model without any sample separation information corresponds to the use of only the marginal likelihood of $\{y_t\}$. It is well known that if $x_{1t} = x_{2t}$ for all t and there are no a priori restrictions on the parameter space, the equations in (2.1) and (2.2) can not be distinguished from each other if there is no sample separation information. This occurs because the names of the two regimes can be interchanged in this situation. The same problem still exists for our model with imperfect sample separation information. With the classification indicator w_t, the a priori assumption that $p_{11} > 1/2$ and $p_{00} > 1/2$ will distinguish the case

	$w=1$	$w=0$
$I=1$	p_{11}	p_{10}
$I=0$	p_{01}	p_{00}

from the following case

	$w=1$	$w=0$
$I^*=1$	p^*_{11}	p^*_{10}
$I^*=0$	p^*_{01}	p^*_{00}

where $I^*=1-I$, i.e., the names of the two regimes have been interchanged. We can distinguish between these cases because, for the first case, we have $p_{11} > 1/2$ but for the latter case, $p^*_{11}(=p_{01}) < 1/2$. However, even if we know that $p_{11} > 1/2$ and $p_{00} > 1/2$, if $w^*(=1-w)$ were to be used as the classification indicator for I, we could not discriminate between the following cases:

	$w^*=1$	$w^*=0$
$I^*=1$	p^{**}_{11}	p^{**}_{10}
$I^*=0$	p^{**}_{01}	p^{**}_{00}

	$w=1$	$w=0$
$I=1$	p_{11}	p_{10}
$I=0$	p_{01}	p_{00}

In this distance, it is apparent that $p^{**}_{11} = p_{00}$ and $p^{**}_{00} = p_{11}$.

Thus, as usual, we need to assume that there is a priori information to distinguish between the two equations in (2.1) and (2.2).

Given such a priori information, the parameters β_1, β_2, σ_1^2, σ_2^2 and λ can be identified from the marginal likelihood function of $\{y_t\}$. Conditional on $w_t=1$, the conditional density function of y_t is

$$f(y_t|w_t=1) = f_1(y_t)\frac{\lambda p_{11}}{p} + f_2(y_t)\frac{(1-\lambda)p_{01}}{p}$$

and hence $\frac{\lambda p_{11}}{p}$ and $\frac{(1-\lambda)p_{01}}{p}$ are identifiable. Since the marginal probability p can be identified from the marginal likelihood of $\{w_t\}$, both the parameters p_{11} and p_{01} are identifiable.

Let $\pounds = \ell n L$ be the log likelihood function of L in (4.1). This log likelihood function is a sum of two components:

$$\pounds = \pounds_1 + \pounds_2 \tag{4.2}$$

where

$$\pounds_1 = \sum_{t=1}^{T} \ell n\ g(y_t) \tag{4.3}$$

which is the log likelihood function of the sample $\{y_t\}$, and

$$\pounds_2 = \sum_{t=1}^{T} \ell n\ h(w_t|y_t) \tag{4.4}$$

where $h(w_t|y_t)$ is the conditional probability function of w_t given y_t. It is well known that for some values of β_1, β_2, and σ_2^2, the log likelihood function of the model without sample separation information, i.e., $\pounds_1$, tends to positive infinity as σ_1^2 goes to zero. The availability of imperfect sample separation information does not eliminate the unboundedness of the log likelihood function $\pounds$. The conditional probability function $h(w_t|y_t)$ is

$$h(w_t|y_t)=w_t\left(p_{11}\frac{f_1(y_t)\lambda}{f_1(y_t)\lambda+f_2(y_t)(1-\lambda)} + p_{01}\frac{f_2(y_t)(1-\lambda)}{f_1(y_t)\lambda+f_2(y_t)(1-\lambda)}\right)$$

$$+ (1-w_t)\left\{(1-p_{11})\frac{f_1(y_t)\lambda}{f_1(y_t)\lambda+f_2(y_t)(1-\lambda)} + (1-p_{01})\frac{f_2(y_t)(1-\lambda)}{f_1(y_t)\lambda+f_2(y_t)(1-\lambda)}\right\} \quad (4.5)$$

Evidently, the conditional probability that $I_t=1$ given y_t is

$$P(1|y_t) = \frac{f_1(y_t)\lambda}{f_1(y_t)\lambda+f_2(y_t)(1-\lambda)} \quad (4.6)$$

Hence

$$h(w_t|y_t)=w_t\left\{p_{11}P(1|y_t)+p_{01}\left[1-P(1|y_t)\right]\right\}$$
$$+(1-w_t)\left\{(1-p_{11})P(1|y_t)+(1-p_{01})\left[1-P(1|y_t)\right]\right\} \quad (4.5')$$

Therefore,

$$|\pounds_2|\le\sum_{t=1}^{T}|\ell n\ h(w_t|y_t)|$$
$$\le\sum_{t=1}^{T}\left|\ell n\left\{p_{11}P(1|y_t)+p_{01}\left[1-P(1|y_t)\right]\right\}\right|$$
$$+\sum_{t=1}^{T}\left|\ell n\left\{p_{10}P(1|y_t)+p_{00}\left[1-P(1|y_t)\right]\right\}\right|$$
$$\le\sum_{t=1}^{T}\max(|\ell np_{11}|,|\ell np_{01}|)+\sum_{t=1}^{T}\max(|\ell np_{10}|,|\ell np_{00}|).$$

The last inequality follows from the fact that the extreme values of $ax+b(1-x)$, $x\in[0,1]$, are attained at $x=0$ and $x=1$. For any p_{11}, p_{01} in $(0,1)$, the log likelihood function $\pounds_2$ is bounded. Since $\pounds_2$ is bounded from below and $\pounds_1$ is unbounded from above, $\pounds$ is unbounded from above.

Even though the likelihood function is unbounded, Kiefer [1978] has shown that, for the model without sample separation information, one of the roots of the log likelihood normal equations is consistent and asymptotically efficient. Hartley and Mallela [1977] derived the same conclusion for the disequilibrium market model. Thus to estimate our model by the ML method, we will consider the solution of the normal equations. From the density function in (2.4), the conditional probability that $I_t=1$ conditional on y_t, $w_t=1$ is

$$P(1|y_t,w_t=1) = \frac{\lambda p_{11} f_1(y_t)}{\lambda p_{11} f_1(y_t)+(1-\lambda)p_{01} f_2(y_t)} \tag{4.7}$$

and the conditional probability that $I_t=1$ conditional on y_t, $w_t=0$ is

$$P(1|y_t,w_t=0) = \frac{\lambda(1-p_{11}) f_1(y_t)}{\lambda(1-p_{11}) f_1(y_t) + (1-\lambda)(1-p_{01}) f_2(y_t)} \tag{4.8}$$

The first order derivatives of the log likelihood in (4.1) are as follows:

$$\frac{\partial \ln L}{\partial \beta_1} = \frac{1}{\sigma_1^2} \sum_{t=1}^{T} P(1|y_t,w_t) x_{1t}'(y_t - x_{1t}\beta_1) \tag{4.9}$$

$$\frac{\partial \ln L}{\partial \beta_2} = \frac{1}{\sigma_2^2} \sum_{t=1}^{T} P(0|y_t,w_t) x_{2t}'(y_t - x_{2t}\beta_2) \tag{4.10}$$

$$\frac{\partial \ln L}{\partial \sigma_1^2} = \frac{1}{2\sigma_1^4} \sum_{t=1}^{T} P(1|y_t,w_t)[(y_t - x_{1t}\beta_1)^2 - \sigma_1^2] \tag{4.11}$$

$$\frac{\partial \ln L}{\partial \sigma_2^2} = \frac{1}{2\sigma_2^4} \sum_{t=1}^{T} P(0|y_t,w_t)[(y_t - x_{2t}\beta_2)^2 - \sigma_2^2] \tag{4.12}$$

$$\frac{\partial \ln L}{\partial \lambda} = \sum_{t=1}^{T} [\tfrac{1}{\lambda} P(1|y_t,w_t) - \tfrac{1}{1-\lambda} P(0|y_t,w_t)] \tag{4.13}$$

$$\frac{\partial \ln L}{\partial p_{11}} = \sum_{t=1}^{T} \left\{\frac{w_t}{p_{11}} - \frac{1-w_t}{1-p_{11}}\right\} P(1|y_t, w_t) \tag{4.14}$$

$$\frac{\partial \ln L}{\partial p_{01}} = \sum_{t=1}^{T} \left\{\frac{w_t}{p_{01}} - \frac{1-w_t}{1-p_{01}}\right\} P(0|y_t, w_t) \tag{4.15}$$

Setting the above derivatives to zero, we have

$$\beta_1 = \left\{\sum_{t=1}^{T} P(1|y_t, w_t) x_{1t}' x_{1t}\right\}^{-1} \sum_{t=1}^{T} P(1|y_t, w_t) x_{1t}' y_t \tag{4.16}$$

$$\beta_2 = \left\{\sum_{t=1}^{T} P(0|y_t, w_t) x_{2t}' x_{2t}\right\}^{-1} \sum_{t=1}^{T} P(0|y_t, w_t) x_{2t}' y_t \tag{4.17}$$

$$\sigma_1^2 = \left\{\sum_{t=1}^{T} P(1|y_t, w_t)\right\}^{-1} \sum_{t=1}^{T} P(1|y_t, w_t)(y_t - x_{1t}\beta_1)^2 \tag{4.18}$$

$$\sigma_2^2 = \left\{\sum_{t=1}^{T} P(0|y_t, w_t)\right\}^{-1} \sum_{t=1}^{T} P(0|y_t, w_t)(y_t - x_{2t}\beta_2)^2 \tag{4.19}$$

$$\lambda = \frac{1}{T} \sum_{t=1}^{T} P(1|y_t, w_t) \tag{4.20}$$

$$p_{11} = \left\{\sum_{t=1}^{T} P(1|y_t, w_t)\right\}^{-1} \sum_{t=1}^{T} w_t P(1|y_t, w_t) \tag{4.21}$$

$$p_{01} = \left\{\sum_{t=1}^{T} P(0|y_t, w_t)\right\}^{-1} \sum_{t=1}^{T} w_t P(0|y_t, w_t) \tag{4.22}$$

To solve the normal equations $\frac{\partial \ln L}{\partial \theta} = 0$ where $\theta = (\beta_1', \beta_2', \sigma_1^2, \sigma_2^2, \lambda, p_{11}, p_{01})'$, an iterative method is required. Since the equations in (4.16)-(4.22) are equivalent to $\frac{\partial \ln L}{\partial \theta} = 0$, the root of the log likelihood normal equations is the root of the equations in (4.16)-(4.22). One can start with an initial value of θ, calculate the probabilities $P(1|y_t, w_t)$ and $P(0|y_t, w_t)$ and use these in the equations (4.16)-(4.22) to obtain a new value of θ, say θ_1. Given θ_1, new estimates of $P(1|y_t,w_t)$ and $P(0|y_t,w_t)$ can then be calculated from (4.7) and (4.8). The process can be iterated until convergence is obtained. The solution to this algorithm solves the likelihood equations. This iteration is a straightforward generalization of the iterative method for the model without sample separation information suggested by Kiefer [1980]. The iteration reduces to a set of weighted least squares estimators. The probability $P(1|y_t,w_t)$ is the conditional probability that the observation y_t is from regime 1 and $P(0|y_t,w_t)$ is the conditional probability that y_t is drawn from regime 2. The estimators of λ, p_{11} and p_{01} in (4.20), (4.21) and (4.22) are intuitively appealing. We note that $w_t P(i|y_t,w_t)$ is the conditional joint probability $\mathrm{Prob}(I_t=i,\ w_t=1|y_t,w_t)$, for $i=0,1$. As pointed out in Kiefer [1980], such iterative methods are related to the E-M algorithm (also see Hartley [1978]).

Whenever $p_{11}\neq p_{01}$, $\mathrm{Prob}(w_t|I_t)\neq \mathrm{Prob}(w_t)$ and the indicator w_t will contain information on the equations in the two regimes. This additional information can be used to improve the efficiency of the estimation of the switching regression equations in (2.1) and (2.2), relative to estimation using only the sample $\{y_t\}$. This occurs since $£ = £_1 + £_2$ and $\frac{\partial £_1}{\partial \theta}$ and $\frac{\partial £_2}{\partial \theta}$ are uncorrelated by construction. Therefore the difference between the information matrix of $£$ and the information matrix of $£_1$ is

$$E\left(-\frac{\partial^2 £}{\partial\theta\,\partial\theta'}\right) - E\left(-\frac{\partial^2 £_1}{\partial\theta\,\partial\theta'}\right) = E\left(-\frac{\partial^2 £_2}{\partial\theta\,\partial\theta'}\right)$$

which is a positive semi-definite matrix. In addition, the conditional log likelihood $£_2$ contains information about the parameters $\beta_1, \beta_2, \sigma_1^2, \sigma_2^2$ and λ. The conditional probability $P(1|y_t)$ in (4.6) can be rewritten as

$$P(1|y_t) = \left[1 + \exp(ay_t^2 + y_t\ b_t + c_t)\right]^{-1} \tag{4.23}$$

where $a = \frac{1}{2}\left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}\right)$, $b_t = x_{1t}\beta_1/\sigma_1^2 - x_{2t}\beta_2/\sigma_2^2$ and

$$c_t = \frac{1}{2}\left\{\left(\frac{x_{1t}\beta_1}{\sigma_1}\right)^2 - \left(\frac{x_{2t}\beta_2}{\sigma_2}\right)^2\right\} + \ell n\left(\frac{1-\lambda}{\lambda}\ \frac{\sigma_1}{\sigma_2}\right)$$

which is a quadratic logistic probability. When $\sigma_1^2 = \sigma_2^2$, (4.23) becomes a linear logistic probability. The log likelihood $£_2$ in (4.4) with (4.5') and (4.23) will provide estimates of some nonlinear functions of the parameters $\beta_1, \beta_2, \sigma_1, \sigma_2$ and λ in addition to the estimates of p_{11} and p_{01}. These additional pieces of information are independent of the likelihood function $£_1$ and hence they will improve the efficiency of the estimators of β_1, β_2, etc. For the model with known sample separation, the value of information has been addressed in Goldfeld and Quandt [1975] in the context of a disequilibrium model, via a Monte Carlo experiment; and Kiefer [1979] and Schmidt [1981] compared the asymptotic variances in a normal mixture model. As indicated by their numerical stimulations, the value of sample separation information is higher when the two samples are hard to disentangle and goes to zero as the two distributions become far apart.

5. Regime Classification and Probability of Misclassification

Since the true regime which generates the sample y_t for each period is unknown, it may be of some interest to classify the observations into the underlying regime for each time period. This problem is a problem in discrimation analysis. For our model, several intuitively appealing methods can be used to classify the sample y_t into the underlying regimes. One possibility is to use the indicator w_t to classify the sample. Thus, if $w_t=1$, the underlying regime in period t would be classified to regime 1; regime 2, otherwise. Another possibility is to use the conditional probability $P(1|y_t)$ in (4.6). The observation at period t belongs to regime 1 if $P(1|y_t) > 0.5$; regime 2, otherwise. On the other hand, since we have a sample consisting of both y_t and w_t, it may be useful to employ the conditional probability $P(1|y_t,w_t)$ in (4.7) and (4.8) instead of $P(1|y_t)$ to classify the regimes. The rule using the probability $P(1|y_t,w_t)$ is indeed the optimal rule in the sense of minimizing the total probability of misclassification.

Let $Q_i(w_t)$ be the probability function of w_t conditional on regime i, i=1,2. The joint density function for y_t and w_t in (2.4) can be rewritten as

$$f(y_t,w_t)=\lambda f_1(y_t)Q_1(w_t)+(1-\lambda)f_2(y_t)Q_2(w_t) \tag{5.1}$$

where

$$Q_1(w_t)=w_t p_{11} + (1-w_t)(1-p_{11}) \tag{5.2}$$

and

$$Q_2(w_t) = w_t p_{01} + (1-w_t)(1-p_{01}) \tag{5.3}$$

A classification rule D_t for each period t may be thought of as an ordered partition $D_t = \langle D_{1t}, D_{2t} \rangle$ of the sample space of (y_t, w_t), where D_t assigns period t to regime i if and only if $(y_t, w_t) \in D_{it}$. For any classification rule, misclassification may occur. Let $P_{ct}(i|j)$ be the probability of classification of the sample (y_t, w_t) from regime j to regime i. $P_{ct}(1|2)$ and $P_{ct}(2|1)$ are thus the misclassification probabilities from one regime to another;

$$P_{ct}(1|2) = \sum_{D_{1t}}\int f_2(y) Q_2(w)\, dy \tag{5.4}$$

and

$$P_{ct}(2|1) = \sum_{D_{2t}}\int f_1(y) Q_1(w)\, dy \tag{5.5}$$

Since our sample (y_t, w_t) is a mixture of continuous and discrete variables, the summation and integration operators are used in (5.4) and (5.5) with respect to w_t and y_t in D_{1t} and D_{2t}, respectively. The total probability of misclassification $C(D_t)$ for period t is

$$C(D_t) = P_{ct}(1|2)(1-\lambda) + P_{ct}(2|1)\lambda \tag{5.6}$$

The optimal partition $D_t^* = \langle D_{1t}^*, D_{2t}^* \rangle$ is the partition which minimizes the total probability of misclassification, i.e.,

$$C(D_t^*) = \min_{D_t} C(D_t)$$

It is well known (Anderson [1952]) that the optimal partition is

$$D^*_{1t} = \{(y_t, w_t) \mid \lambda f_1(y_t)Q_1(w_t) > (1-\lambda)f_2(y_t)Q_2(w_t)\}$$

and

$$D^*_{2t} = \{(y_t, w_t) \mid \lambda f_1(y_t)Q_1(w_t) < (1-\lambda)f_2(y_t)Q_2(w_t)\}.$$

Since y_t is a continuous variable, the probability that $\lambda f_1(y_t)Q_1(w_t) = (1-\lambda)f_2(y_t)Q_2(w_t)$ is zero. It is straightforward to show that the classification rule that the sample (y_t, w_t) be classified to regime 1 if and only if $P(1|y_t, w_t) > 0.5$, is exactly the above optimal rule.[2/]

The probabilities of misclassification can be evaluated by the equations (5.4), (5.5) and (5.6). For the classification rule using w_t only, the probability of misclassifying an observation from regime 2 to regime 1 is p_{01}; the probability of misclassification of regime 1 to regime 2 is p_{10}, and the total probability of misclassification is $\lambda p_{10}+(1-\lambda)p_{01}$. The computation of the other two classification rules, which are based on conditional probabilities, is relatively more complicated. Consider the case where $\sigma_1^2=\sigma_2^2=\sigma^2$. Taking the logarithmic transformation of the discrimination scores $\lambda f_1(y_t)Q_1(w_t)$ and $(1-\lambda)f_2(y_t)Q_2(w_t)$, we have $(y_t, w_t) \in D^*_{1t}$ if and only if

$$\ell n\frac{f_1(y_t)}{f_2(y_t)} > \ell n\frac{1-\lambda}{\lambda} + \ell n\frac{Q_2(w_t)}{Q_1(w_t)} \tag{5.7}$$

or, equivalently,

$$\frac{1}{\sigma^2}(x_{1t}\beta_1 - x_{2t}\beta_2)\left(y_t - \frac{1}{2}(x_{1t}\beta_1 + x_{2t}\beta_2)\right) > \ell n\frac{1-\lambda}{\lambda} + w_t \ \ell n\frac{p_{01}}{p_{11}} + (1-w_t)\,\ell n\frac{p_{00}}{p_{10}} \tag{5.8}$$

The left hand side expression is Fischer's linear discrimination function. When $\sigma_1^2 = \sigma_2^2$, the optimal classification rule is linear in y_t and w_t. Let $||d_t||^2 = \frac{1}{\sigma^2}(x_{1t}\beta_1 - x_{2t}\beta_2)^2$ be the Mahalanobis distance. If the observation in period t comes from regime 2, $y_t = x_{2t}\beta_2 + \varepsilon_{2t}$ and equation (5.8) implies that

$$\frac{1}{\sigma^2}(x_{1t}\beta_1 - x_{2t}\beta_2)\varepsilon_{2t} > w_t \ell n\frac{p_{01}}{p_{11}} + (1-w_t)\ell n\frac{p_{00}}{p_{10}} + \ell n\frac{1-\lambda}{\lambda} + \frac{1}{2}||d_t||^2$$

It follows that

$$P_{ct}(1|2,w_t) = \Phi\left\{\left(w_t \ell n\frac{p_{11}}{p_{01}} + (1-w_t)\ell n\frac{p_{10}}{p_{00}} + \ell n\frac{\lambda}{1-\lambda} - \frac{1}{2}||d_t||^2\right)/||d_t||\right\} \tag{5.9}$$

where $\Phi(\cdot)$ is the standard normal distribution function, and hence the probability of misclassification of regime 2 to regime 1 is

$$P_{ct}(1|2) = \Phi\left(\frac{\ell n\frac{\lambda p_{11}}{(1-\lambda)p_{01}} - \frac{1}{2}||d_t||^2}{||d_t||}\right)p_{01} + \Phi\left(\frac{\ell n\frac{\lambda p_{10}}{(1-\lambda)p_{00}} - \frac{1}{2}||d_t||^2}{||d_t||}\right)p_{00} \tag{5.10}$$

Similarly,

$$P_{ct}(2|1,w_t) = \Phi\left(\frac{w_t \ell n\frac{p_{01}}{p_{11}} + (1-w_t)\ell n\frac{p_{00}}{p_{10}} + \ell n\frac{1-\lambda}{\lambda} - \frac{1}{2}||d_t||^2}{||d_t||}\right) \tag{5.11}$$

and

$$P_{ct}(2|1)=\Phi\left(\frac{\ell n\frac{(1-\lambda)p_{01}}{\lambda p_{11}}-\frac{1}{2}||d_t||^2}{||d_t||}\right)p_{11}+\Phi\left(\frac{\ell n\frac{(1-\lambda)p_{00}}{\lambda p_{10}}-\frac{1}{2}||d_t||^2}{||d_t||}\right)p_{10} \quad (5.12)$$

The total probability of misclassification for D_t^* is

$$C(D_t^*)=P_{ct}(1|2)(1-\lambda)+P_{ct}(2|1)\lambda \quad (5.13)$$

The classification rule based on $P(1|y_t)$ is equivalent to classifying period t to regime 1 if and only if

$$\frac{1}{\sigma^2}(x_{1t}\beta_1-x_{2t}\beta_2)\left(y_t-\frac{1}{2}(x_{1t}\beta_1+x_{2t}\beta_2)\right)>\ell n\frac{1-\lambda}{\lambda} \quad (5.14)$$

which is the familiar discrimination rule for a mixture of normal distributions. The probability of misclassification of regime 1 to regime 2 is

$$\Phi\left(\frac{\ell n\frac{1-\lambda}{\lambda}-\frac{1}{2}||d_t||^2}{||d_t||}\right); \quad (5.15)$$

the probability of misclassification of regime 2 to regime 1 is

$$\Phi\left(\frac{\ell n\frac{\lambda}{1-\lambda}-\frac{1}{2}||d_t||^2}{||d_t||}\right) \quad (5.16)$$

and the total probability of misclassification for this rule is

$$\Phi\left(\frac{\ell n\frac{1-\lambda}{\lambda}-\frac{1}{2}||d_t||^2}{||d_t||}\right)\lambda+\Phi\left(\frac{\ell n\frac{\lambda}{1-\lambda}-\frac{1}{2}||d_t||^2}{||d_t||}\right)(1-\lambda) \quad (5.17)$$

All these probabilities become small when the distance d_t is large.[3/]

When $\sigma_1^2 \neq \sigma_2^2$, the discrimination function is quadratic in y_t and linear in w_t and so the computation of the misclassification probabilities is more complicated. In this case, monte carlo simulation methods may allow us to evaluate these probabilities. The above probabilities are functions of unknown parameters, but these parameters can be consistently estimated by the relevant consistent estimators. Discriminant analysis with mixtures of continuous and discrete variables has been considered in Chang and Afifi [1972], Krzanowski [1975] and Goldstein and Dillon [1978]. The models they considered differ from ours. The density functions of the continuous variables for each regime conditional on the discrete variables are assumed to be normally distributed and hence their marginal distributions are mixtures of normal distributions. For our model, the marginal distribution of y_t for each regime is normal and the distribution of y_t conditional on w_t is a mixture of normal distributions.

6. The Joint Executive Committee and Cartel Stability

In the following two sections, the above methodology will be applied to weekly time series data on the Joint Executive Committee (JEC) railroad cartel from 1880 to 1886. We estimate the parameters of demand and supply functions for the industry, and identify periods in which firms were behaving collusively, as opposed to noncooperatively. These different behavioral rules are reflected by differing supply functions. Thus we estimate a simultaneous equations market model, where the supply curve is drawn from one of two possible regimes.

The JEC was a cartel which controlled eastbound freight shipments from Chicago to the Atlantic seaboard in the 1880s. (Historical descriptions of the JEC are contained in MacAvoy [1965] and Ulen [1978].) Grain shipments accounted for 73 percent of all dead freight tonnage handled by the JEC. The railroads also handled eastbound shipments of flour and provisions but, with only a few exceptions, the prices changed for transporting these commodities were tied to the grain rate. While different railroads shipped grain to different port cities, most of the wheat handled by the cartel was subsequently exported overseas, and the rates charged by different firms adjusted to compensate for differences in ocean shipping rates. Finally, differences in delivery speeds were of negligible importance. Thus the assumption that a good of homogenous quality was sold seems to have been approximately satisfied. With little loss of generality, our attention will be focused on the movement of grain.

Price has typically been thought to be the strategic variable of firms in the rail-freight industry. Denote the market price in period t by p_t. Since the products provided by firms are of approximately homogeneous quality, all firms will charge the same price in equilibrium. The total quantity of grain shipped in period t, Q_t, is assumed to be a loglinear function of price,

$$\ell n Q_t = \alpha_0 + \alpha_1 \ell n\, p_t + \alpha_2 L_t + u_{2t} \tag{6.1}$$

where L_t is a dummy variable equal to one if the Great Lakes were open to navigation, and $\{u_{2t}\}$ is a sequence of independently and identically distributed normal variables with zero mean and variance σ_2^2. Here α_1 is the price elasticity of demand, and presumably negative. Also α_2 should be negative, reflecting a decrease in demand when lake steamers, the principle source of competition for the JEC, were operating. This demand equation will be invariant across regimes.

The price equation will vary across regimes to reflect collusive or Bertrand behavior on the part of firms. In the case of the JEC, the cartel agreement took the form of market share allotments. Firms set their rates individually, and the JEC office took weekly accounts so that each railroad could see the amounts transported as well as an index of listed prices. The actual market share of any particular firm would depend on both the prices charged by all the firms as well as on unpredictable stochastic forces. The internal enforcement mechanism adopted by the JEC was a variant of a trigger price strategy. According to Ulen [1978], there were several instances in which the cartel thought that cheating on the agreement had occurred, in the form of secret price cutting.

Prices were then cut in response for a time, at which point prices returned to collusive levels.

The inference problem that the firms faced in deleting cheating is quite similar to that originally posed by Stigler [1964]. As is pointed out by Green and Porter [1981], in an uncertain environment such as that faced by the JEC, a collusive price agreement can be maintained if firms occasionally revert to noncooperative behavior, i.e., engage in a price war, in response to suspected cheating. In this case, since firms are price setters, this involves reverting to Bertrand behavior, or pricing at marginal cost. Such an enforcement mechanism will provide the correct incentives for firms to adhere to the cartel agreement. A firm which considers a secret expansion of output above the collusive level must trade off immediate profit gains with the increased probability that the other firms will suspect cheating, thereby increasing the likelihood of lower future profits. In equilibrium, then, price wars are triggered solely by unpredictable disturbances, as member firms have no incentive to cheat. The predictable fluctuations in demand that resulted from the annual opening and closing of the Great Lakes to shipping, which determine the degree of outside competition, did not disrupt industry conduct but rather, rates adjsuted systematically with the lake navigation season.

Suppose I_t is a latent dichotomous variable which equals one when the industry is in a cooperative regime, and equals zero when the industry witnesses a reversionary episode. The price setting equation is then specified as

$$\ell n p_t = \beta_0 + S_t \beta_1 + \beta_2 I_t + u_{1t} \tag{6.2}$$

where S_t is a vector of exogenous variables which explain some variations in prices over time, including the Great Lakes navigation dummy variable and dummy variables which reflect entry and acquisitions in the industry. In some industries, entry has triggered price wars, which may represent a predatory response on the part of the original firms, or just the outcome of a more uncertain environment. In the case of the JEC, entry occurred twice between 1880 and 1886. It appears that the cartel passively accepted the entrants, allocating them markets shares, and thereby allowing the collusive agreement to continue. The reason for this is probably that when a firm entered the rail-freight industry in the late nineteenth century, it faced a "no-exit" constraint (Ulen [1978], pp. 70-74). Thus for the JEC, entry does not seem to have cause reversions to noncooperative behavior. The parameter β_2 should be positive, reflecting the higher prices charged in cooperative periods. This functional form thus involves only the intercept of the price equation changing across behavioral regimes. As is demonstrated in Porter [1982], this occurs because there is a constant elasticity demand curve, in conjunction with constant elasticity of individual firm marginal costs with respect to own output. (In Porter's econometric work, marginal cost was found to be approximately constant, at least locally.) If cooperative regimes involved pricing at joint-profit maximizing levels, then β_2 should be an exact function of α_1, the demand elasticity. However, in an uncertain environment, an optimal cartel trigger price strategy will not typically involve pricing at joint-profit maximizing levels. There is a trade-off between single period joint profits and the incentive to cheat, which can be reduced only by increasing the severity of punishing reactions. Thus we will not constrain β_2 to depend on α_1.

In summary, price wars should be caused by unpredictable disturbances rather than by entry or predictable fluctuations in demand. The switching of the dichotomous indicator I_t is thus assumed to be governed by the binomial distribution;

$$\begin{aligned} I_t &= 1 \quad \text{with probability } \lambda \\ &= 0 \quad \text{with probability } 1-\lambda. \end{aligned}$$

Note that if trigger price strategies were actually employed by the JEC, the $\{I_t\}$ sequence should follow a Markov process of length equal to the length of reversionary periods. Rather than attempt to estimate the simultaneous equations switching regression model of (6.1) and (6.2) subject to a constraint of this sort, which would be relatively difficult and costly, we have chosen to adopt the above formulation.

7. Data, Empirical Analysis and Results

A principle function of the JEC was information gathering and dissemination to member firms. Weekly accounts were kept in an effort to keep members abreast of developments in the industry. A list of variables used in the estimation is as follows:

- P grain rate, in dollars per 100 lbs.
- Q total quantity of grain shipped, in tons.
- L dummy variable; 1 if Great Lakes were open to navigation, 0 otherwise.
- W regime classification indicator; 1 if colluding reported by Railway Review, 0 otherwise.
- DE_1 dummy variable; 1 from week 18 in 1880 to week 10 in 1883; 0 otherwise; reflecting entry by the Grand Trunk Railway.
- DE_2 dummy variable; 1 from week 11 to week 25 in 1883, 0 otherwise; reflecting an addition to New York Central.
- DE_3 dummy variable; 1 from week 26 in 1883 to week 11 in 1886, 0 otherwise; reflecting entry by the Chicago and Atlantic.
- DE_4 dummy variable; 1 from week 12 to week 16 in 1886, 0 otherwise; reflecting acquisition of the Chicago and Atlantic.
- WK_ℓ a set of seasonal dummy variables; each year was segmented into thirteen four-week segments; ℓ refers to the ℓth segment.

The sample consists of weekly data from week 1 in 1880 to week 16 in 1886. The sample size is 328. The quantity variable, Q, is a reasonably accurate measure of the total tonnage of grain shipped by the members of the cartel. The price variable, P, is an index of prices charged by member firms and was provided by the JEC. The Great Lakes navigation dummy variable, L, documents when the JEC faced its main source of competition. It would be preferable if the prices charged by the lake steamers had also been used in the econometric work. Unfortunately, that series was not available. The classification indicator, w, equals one unless the Railway Review, a trade magazine, reported that a price war was occuring. This series concurred with the reports of the Chicago Tribune and other accounts in this period. This series was gathered by Ulen [1978]. One reason for suspecting measurement errors in this series is that it conflicts sharply with an index of cartel adherence created by MacAvoy [1965]. The various DE dummy variables proxy structural change caused by entry, acquisitions or additions to existing networks. In each case, structural change is presumed to result in a once-and-for all shift in the price equation. The following TABLE 1 gives some simple summary statistics:

TABLE 1: Summary Statistics

Variable	Mean	Standard Deviation
ℓn P	-1.4387	0.2865
ℓn Q	10.0371	0.4688
L	0.5732	0.4954
W	0.6189	0.4864
DE_1	0.4238	----
DE_2	0.0457	----
DE_3	0.4329	----
DE_4	0.0152	----

The empirical model consists of two equations corresponding to equation (6.2) and equation (6.1):

$$\ell n\, P = \beta_0 + \beta_1 L + \textstyle\sum_{j=1}^{4}\beta_{1+j}DEj + \sum_{\ell=1}^{12}\beta_{5+\ell}WK_\ell + \gamma I + u_1 \tag{7.1}$$

and

$$\ell n\, Q = \alpha_0 + \alpha_1 \ell n\, P + \alpha_2 L + \sum_{j=1}^{12} \alpha_{2+j} WK_\ell + u_2 \tag{7.2}$$

This model is formulated as a recursive simultaneous equations model. Equation (7.2) is a standard structural equation. It is overidentified. The dichotomous variable I_t in equation (7.1) is measured by w_t, possibly with error. This equation is a special case of the general switching regression model as described in section 2. Equation (7.2) can be estimated by two stage least squares (2SLS). To estimate equation (7.1), we need to modify slightly the iterative procedure described in section 4. Equations in (4.16), (4.17), (4.18), and (4.19) should be replaced by

$$\gamma = \textstyle\sum_{t=1}^{T} P(1|y_t,w_t)(y_t - x_t\beta)/\sum_{t=1}^{T} P(1|y_t,w_t) \tag{7.3}$$

$$\beta = (\textstyle\sum_{t=1}^{T} x_t'x_t)^{-1}\sum_{t=1}^{T} x_t'(y_t - P(1|y_t,w_t)\gamma) \tag{7.4}$$

and

$$\sigma^2 = \tfrac{1}{T}\textstyle\sum_{t=1}^{T}\{P(1|y_t,w_t)(y_t - \gamma - x_t\beta)^2 + P(0|y_t,w_t)(y_t - x_t\beta)^2\} \tag{7.5}$$

where y_t stands for $\ell n\, P_t$, $x_t = (1, L_t, DE_1, \ldots DE_4, WK_1, \ldots WK_{12})$ and $\beta = (\beta_0, \beta_1, \ldots \beta_{17})'$. The expressions for λ, p_{11}, p_{01} are the same as those given in equations (4.20)-(4.22) without modification.

The estimation of equation (7.1) is reported in the first column of TABLE 2.

TABLE 2
Log Price Equation - Limited Information MLE*

Variables	Estimations with samples $\{y_t, w_t\}$		Estimation with $\{y_t\}$
Constant	-1.4611 (.0480)	-1.4738 (.0424)	-1.4729 (.0421)
L	-.1599 (.0587)	-.1615 (.0159)	-.1723 (.0153)
DE_1	-.1919 (.0397)	-.1827 (.0420)	-.1835 (.0418)
DE_2	-.2114 (.0962)	-.2096 (.0868)	-.2077 (.0835)
DE_3	-.2922 (.0404)	-.2813 (.0425)	-.2769 (.0420)
DE_4	-.4002 (4.9029)	-.3903 (6.587)	-.3955 (5.9972)
I	.4725 (.0152)	.4787 (.0148)	.4830 (.0145)
WK_1	-.0543 (.0283)		
WK_2	.0316 (.0416)		
WK_3	.0052 (.0382)		
WK_4	.0026 (.0334)		
WK_5	-.0312 (.0477)		
WK_6	.0072 (.0702)		
WK_7	-.0161 (.0682)		
WK_8	-.0071 (.0647)		
WK_9	.0176 (.0692)		
WK_{10}	.000026 (.0671)		
WK_{11}	-.0224 (.0625)		
WK_{12}	.0662 (.0631)		
σ^2	.0131 (.0012)	.0132 (.0011)	.0126 (.0010)
λ	.7160 (.0284)	.7166 (.0278)	.7178 (.0279)
p_{11}	.8109 (.0278)	.8048 (.0273)	----
p_{01}	.1348 (.0391)	.1489 (.0411)	----
ℓn likelihood	-87.482	-92.859	-150.79 (combined with $\{w_t\}$)

(*)standard errors are in paretheses.

The estimates of the price equation are fairly sensible. Price was significantly higher in cooperative periods. In cooperative periods, price was 60 percent higher than the price in the noncooperative periods, ceteris parabus.[4/] Equivalently, when there were price wars, the price was cut about 37.66 percent. When the lakes were open to navigation, price was significantly lower. But it was only about 14.78 percent lower. The coefficients of the structural dummies are also reasonable. All the entry and acquisition variables have negative coefficients and, except for the variable DE_4, they are statistically significant. Both instances of entry led to a fall in market price. All the seasonal dummies are not significant at conventional levels of significance. The second column in TABLE 2 presents the estimation of the price equation without the seasonal dummy variables. A joint test using the likelihood ratio statistic confirms their insignificance. The computed value of the minus two log likelihood ratio is 10.754. The chi-square statistic with 12 degrees of freedom at a ten percent level of significance is 18.55. The estimated regime probability λ is 0.72 which implies that in seventy-two percent of our sample periods, the firms are cooperative. The estimate of p_{11} is 0.81 which implies that the Railway Review had reported colluding correctly in eighty-one percent of the sample periods. In about thirteen percent of the noncooperative periods, the Railway Review had not reported correctly. To compare the MLE of the price equation with and without the indicators $\{w_t\}$, we also estimate the price equation using the sample $\{y_t\}$ only. The estimates are reported in column three. A comparison of the second and third columns clearly reveals that there are essentially no differences between the two sets of estimates and the estimated standard errors.[5/] These indicate that, at least for this data set, the asymptotic efficiency gains in using the indicators $\{w_t\}$ in addition to the observations of $\{y_t\}$ are small. This occurs since the two distributions of $\ell n\, P_t$ are far apart. This is evident from the fact that the difference of the means is 0.4787 and the variance is only 0.013.[6/] This result supports the Monte Carlo simulation results in Schmidt [1981]. The likelihood value reported in the third column is computed from the function $\ln L$, where

$$\ln L = \sum_{t=1}^{T} \ln[\lambda f_1(y_t) + (1-\lambda) f_2(y_t)] + \sum_{t=1}^{T} \ln[w_t p + (1-w_t)(1-p)]$$

and it is evaluated at the estimates of the third column and $\hat{p} = \frac{1}{T}\sum_{t=1}^{T} w_t$. This likelihood corresponds to the constrained likelihood function of the sample $\{y_t, w_t\}$ with the constraint $p_{11} = p_{01}$. The two values of the log likelihood function in the second and third columns are directly comparable. It is apparent that the unconstrained version has a much better fit even though the estimates of the co-efficients are quite similar. Obviously, the constraint $p_{11} = p_{01}$ is rejected.

In TABLE 3, the probabilities of misclassification based on the various classification rules are computed. Since the switching regression equations have the same slope coefficients and the same variance, the probabilities of misclassification in (5.11)-(5.13) and (5.15)-(5.17) do not depend on the exogenous variables and hence the probabilities of misclassification are the same for all time periods. All these probabilities are estimated using the MLE in the second column of TABLE 2 with the exception of the probabilities in the third column of TABLE 3, which are estimated using the estimates in the last column of TABLE 2. Since these probabilities are continuous functions of the unknown parameters in the model, the estimated probabilities are consistent. Comparing the three classification rules, the recommended classification rule has the smallest total probability of misclassification C(D), as we expected theoretically. The total probability of misclassification using the indicator w only is about eighteen percent. The use of the classification rules based on the conditional probabilities $P(\cdot|y)$ and $P(\cdot|y,w)$ reduces tremendously the total misclassification probability to only 1.7 and 1.2 percent, respectively. The two estimates of the probabilities in the second and third columns of TABLE 3 are slightly different. Since the estimates in the second column are computed by using more efficient estimates of the unknown parameters, these estimated probabilities may be more reliable than the estimates in the third column. The total probability of misclassification based on either of the two conditional probabilities, $P(1|y)$ or $P(1|y,w)$, are indeed quite small. The reduction in the total probabilitity of misclassification by using the sample values of y_t and w_t is only about one half percent in absolute value, compared with using the sample values of y_t only. However, in terms of relative magnitudes, the reduction is about 28 percent. The computation of the specific probabilities of misclassification $P_c(2|1)$ and $P_c(1|2)$ also reveals that the best classification rule is to use the whole sample (y_t, w_t) and the worst classification uses only the indicator w_t. In TABLE 4, we investigate the compatibility of the three classification rules in the actual classification of the 328 sample periods. Contingency tables for pairwise comparisons are provided. Comparing the classification rules based on $P(1|y)$ and w, 66 periods out of a total of 328 have been classified differently. There are 57 periods classified differently if one compares the classification rules based on w and $P(1|y,w)$. Only 9 periods are classified differently with the rules employing $P(1|y,w)$ and $P(1|y)$. To summarize the compatibility of these rules we compute the measurement of association for the 2×2 tables based on the cross-product ratio (Fienberg [1977], pp. 16-19). The association between the two classification rules based on $P(1|y,w)$ and $P(1|y)$ is much stronger than their respective associations with the classification rule which uses the indicator w. On the other hand, the classification rule using the indicator w only has a relatively stronger association with the optimal classification rule based on $P(1|y,w)$, than with the classification rule based on $P(1|y)$.

TABLE 3

Probabilities of Misclassification

	classification rules based on			
Error probabilities	$P(\cdot\|y,w)$	$P(\cdot\|y)$		W
		MLE	Consistent Est.	
$P_c(2\|1,w=1)$	.0034	---	---	---
$P_c(2\|1,w=0)$	.0254	---	---	---
$P_c(2\|1)$	.0077	.0106	.0089	.1952
$P_c(1\|2,w=1)$	.0727	---	---	---
$P_c(1\|2,w=0)$	.0134	---	---	---
$P_c(1\|2)$	.0222	.0314	.0265	.1489
C(D)	.0118	.0165	.0139	.1821

TABLE 4

Compatibility of Sample Classification - Frequency Counts

	$P(1\|y) > .5$	$P(1\|y) < .5$	Totals
w = 1	187	16	203
w = 0	50	75	125
Totals	237	91	328

cross-product ratio = 17.53

	$P(1\|y,w) > 0.5$	$P(1\|y,w) < .5$	Totals
w = 1	190	13	203
w = 0	44	81	125
Totals	234	94	328

cross-product ratio = 26.91

	$P(1\|y,w) > .5$	$P(1\|y,w) < .5$	Totals
$P(1\|y) > .5$	231	6	237
$P(1\|y) < .5$	3	88	91
Totals	234	94	328

cross-product ratio = 1129.33

In TABLE 5, we report the OLS estimation of the log price equation with the classification indicator w as a regressor. The estimates in the first column are the OLS estimates with the seasonal dummy variables included and those in the second column are the OLS estimates without the seasonal dummy variables. Comparing the OLS estimates with the maximum likelihood estimates in Table 2, the biases of the estimate of the coefficient of I in the price equation and the variance σ^2 are obvious. As theoretically predicated, the estimated value of the coefficient of I is indeed biased downward. In absolute magnitude, this bias is about 37 percent. In statistical terms, the difference is about 12 standard deviations. This coefficient indicates that the price of shipment would increase 35 percent instead of 60 percent during cooperative periods. The OLS estimate of the variance σ^2 is three times larger than the consistent estimate. In addition, the dummy variables DE_4 and WK_3 have become significant. These comparisons indicate clearly that measurement errors in the classification indicator should not be neglected.

Finally, the demand equation in (7.2) is estimated by two stage least squares. The estimated equation is

$$\ell n\, Q = \underset{(0.184)}{9.169} - \underset{(0.1200)}{0.4367}\, L - \underset{(0.1207)}{.7420}\, \ell n\, P + \text{---(seasonal dummies)--} + \hat{u}_2$$

where the standard errors are in parentheses. In the demand equation, the predicted quantity demanded is much lower when the lakes were open. The demand elasticity is negative, as expected, and less than one in absolute value, indicating that a price increase would be expected to increase total industry revenues. This is consistent with the notion that member firms have an incentive to raise prices to increase total revenues. In the absence of cost data, we cannot identify the joint profit maximizing price.

TABLE 5

Log Price Equation - OLS Estimation

Variables	OLS Estimates	
Constant	-1.351 (.068)	-1.297 (.047)
L	-.112 (.062)	-.161 (.022)
DE_1	-.215 (.045)	-.218 (.042)
DE_2	-.168 (.064)	-.169 (.064)
DE_3	-.294 (.047)	-.298 (.044)
DE_4	-.406 (.101)	-.390 (.097)
W	.303 (.025)	.301 (.024)
WK_1	-.007 (.055)	
WK_2	.085 (.055)	
WK_3	.131 (.056)	
WK_4	.051 (.059)	
WK_5	-.018 (.080)	
WK_6	-.025 (.080)	
WK_7	-.042 (.080)	
WK_8	.019 (.081)	
WK_9	.043 (.081)	
WK_{10}	-.052 (.081)	
WK_{11}	-.020 (.080)	
WK_{12}	.073 (.080)	
σ^2	.0385	.0394
$\bar{R}^2$	.5310	.5207

8. Conclusions

This article was motivated by an empirical problem on the study of cartel stability. The empirical study employed weekly time series data on the Joint Executive Committee railroad cartel from 1880 to 1886 to test the proposition that observed price wars represented a switch from collusive to noncooperative behavior. The model is an exogenous switching regression model with unknown regime switching. However, some information was available on sample separation; a trade magazine, the Railway Review reported in each of the sample periods whether a price war was occuring or not. However, this series may not be at all accurate. We have derived the direction of the least squares bias when an imperfect regime classification indicator is regarded as perfect and used to classify the samples into the different regimes. The empirical results confirm that the biases are substantial. An efficient estimation method is suggested to take into account the imperfection of the indicator. The usefulness of this extra information in regime classification is investigated. This extra information helps to reduce regime classification errors. An optimal regime classification rule is derived and compared theoretically and empirically with others. The econometric evidence indicates that reversions to noncooperative behavior did occur in the Joint Executive Committee, with a significant decrease in market price in these periods.

Footnotes

(*) We would like to thank Professor Gary Chamberlain for valuable discussions on the problem of classification rules. Any errors are our own. Lung-Fei Lee appreciates the financial support of the National Science Foundation under Grant SES-8006481 to the University of Minnesota. Robert Porter benefited from a Sloan Foundation Grant to the University of Minnesota Economics Department.

1/ Since the likelihood function is not very simple, computation of the moments of the posterior distribution will rely on numerical techniques. The requisite computational effort is rather large for a sample size of more than fifty as indicated by Mouchart.

2/ In the context of a disequilibrium market model, both Gersovitz [1980] and Kiefer [1980a] recommend the (similar) conditional probability classification rule. In their articles, they have not provided a theoretical justification for their recommendation. Their recommended classification rule is indeed optimal according to the same argument that we provide here.

3/ The derivative of the function $\Phi\left(\frac{a - \frac{1}{2}x^2}{x}\right)$ for $x > 0$ is $-\frac{1}{2}\phi\left(\frac{a - \frac{1}{2}x^2}{x}\right)\left(1 + \frac{2a}{x^2}\right)$ and hence the function is decreasing on the range $x^2 > -2a$. It goes to zero as x tends to positive infinity.

4/ This is computed from the ratio $\frac{P_1 - P_0}{P_0} = e^{.4725} - 1 = 0.604$ where P_1 is the expected price when $I = 1$ and P_0 is the expected price when $I = 0$. The ratio $\frac{P_1 - P_0}{P_0}$ does not depend on the exogenous variables x as they are cancelled out.

5/ The standard errors in the TABLE 2 are evaluated by the cross-product of the first order sample derivatives evaluated at the corresponding estimates. The inequalities of the standard errors need not necessarily hold for finite samples.

6/ Further evidence is provided by the small classification errors reported in TABLE 3 using the classification rule based on $P(1|y)$.

References

Aigner, D.J. (1973), "Regression with a Binary Independent Variable Subject to Errors of Observation", Journal of Econometrics 1, pp. 49-60.

Anderson, T.W. (1952), An Introduction to Multivariate Statistical Methods, New York: Wiley.

Chang, P.C. and Afifi, A.A. (1972), "Classification Based on Dichotomous and Continuous Variables", Journal of American Statistical Assocation 67, pp. 336-339.

Day, N.E. (1969), "Estimating the Components of a Mixture of Normal Distributions", Biometrika 56, pp. 463-474.

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

Fienberg, S.E. (1977), The Analysis of Cross-Classified Categorical Data, MIT Press, Cambridge, MA.

Gersovitz, M. (1980), "Classification Probabilities for the Disequilibrium Model", Journal of Econometrics 14, pp. 239-246.

Goldfeld, S.M. and R.E. Quandt (1972), Nonlinear Methods in Econometrics, North-Holland.

Goldfeld, S.M. and R.E. Quandt (1975), "Estimation in a Disequilibrium Model and the Value of Information", Journal of Econometrics 3, pp. 325-348.

Goldfeld, S.M. and R.E. Quandt (1976), "Estimation of Structural Change in Simultaneous Equation Models", in Goldfeld and Quandt, eds., Studies in Nonlinear Estimations, chapter 2.

Goldstein, M. and Dillon, W.R. (1978), Discrete Discriminant Analysis, New York: John Wiley & Sons.

Green, E.J. and R.H. Porter (1981), "Noncooperative Collusion under Imperfect Price Information", C.E.R. Discussion Ppaer No. 81-142, University of Minnesota

Hartley, M.J. and P. Mallela (1977), "The Asymptotic Properties of a Maximum Likelihood Estimator for a Model of Markets in Disequilibrium", Econometrica 45, pp. 1205-1220.

Hartley, M.J. (1978), "Comment" (on Quandt and Ramsey (1978)), Journal of the American Statistical Association 73, pp. 738-741.

Kiefer, N.M. (1978), "Discrete Parameter Variation: Efficient Estimation of a Switching Regression Model", Econometrica 46, 427-434.

Kiefer, N.M. (1979), "On the Value of Sample Separation Information", Econometrica 47, pp. 997-1003.

Kiefer, N.M. (1980), "A Note on Switching Regressions and Logistic Discrimination", Econometrica 48, pp. 1065-1069.

Kiefer, N.M. (1980a), "A Note on Regime Classification in Disequilibrium Models", Review of Economic Studies 47, pp. 637-639.

Krzanowski, W.J. (1975), "Dsicrimination and Classification using Both Binary and Continuous Variables", Journal of American Statistical Association 70, pp. 782-790.

MacAvoy, P.W. (1965), The Economic Effects of Regulation, MIT Press, Cambridge.

Maddala, G.S. and Nelson, F.D. (1974), "Maximum Likelihood Methods for Models of Markets in Disequilibrium", Econometrica 42, pp. 1013-1030.

Mouchart, M. (1977), "A Regression Model with an Explanatory Variable Which is Both Binary and Subject to Errors", in D. J. Aigner and A. S. Goldberger, Eds., Latent Variables in Socio-Economic Models, chapter 4.

Porter, R.H. (1982), "A Study of Cartel Stability: The Joint Executive Committee 1880-1886", Discussion Paper No. 82-158, Center for Economic Research, Department of Economics, University of Minnesota.

Quandt, R.E. (1972), "A New Approach to Estimating Switching Regressions", Journal of the American Statistical Association 67, pp. 306-310.

Quandt, R.E., and J. Ramsey (1978), "Estimating Mixtures of Normal Distributions and Switching Regressions", Journal of the American Statistical Association 73, pp. 730-738.

Schmidt, P. (1981), "Further Results on the Value of Sample Separation Information", Econometrica 49, pp. 1339-1343.

Schmidt, P. (1982), "An Improved Version of the Quandt-Ramsey MGF Estimator for Mixture of Normal Distributions and Switching Regressions", Econometrica 50, pp. 501-516.

Stigler, G.J. (1964), "A Theory of Oligopoly", Journal of Political Economy 72, pp. 44-61.

Ulen, T.S. (1978), Cartels and Regulation, unpublished Ph.D. dissertation, Stanford University.