LEAST SQUARES AND MAXIMUM LIKELIHOOD
ESTIMATION OF SWITCHING REGRESSIONS

Stephen M. Goldfeld
Harry H. Kelejian
Richard E. Quandt

ECONOMETRIC RESEARCH PROGRAM
Research Memorandum No. 130
November 1971

The research described in this paper was partially supported by the National Science Foundation .

Econometric Research Program
PRINCETON UNIVERSITY
207 Dickinson Hall
Princeton, New Jersey

LEAST SQUARES AND MAXIMUM LIKELIHOOD ESTIMATION OF SWITCHING REGRESSIONS

Stephen M. Goldfeld
Princeton University

Harry H. Kelejian
New York University

Richard E. Quandt
Princeton University

1. Introduction

Several recent studies have analyzed the problem of estimating the parameters of regression equations subject to discontinuous shifts at unknown points in the data series.[1] In its simplest form the problem may be stated as follows: Let n_1 observations be generated by the regression equation

$$Y_1 = X_1\beta_1 + U_1 \tag{1-1}$$

and n_2 observations by

$$Y_2 = X_2\beta_2 + U_2 \tag{1-2}$$

where Y_1 and Y_2 are n_1 and n_2-element vectors of observations on the dependent variable, X_1 and X_2 are $n_1 \times k$ and $n_2 \times k$ matrices of observations on the independent variables, U_1 and U_2 are n_1 and n_2-element vectors of unobservable error terms distributed as $N(0,\sigma_1^2 I)$ and $N(0,\sigma_2^2 I)$. In general, $(\beta_1,\sigma_1^2) \neq (\beta_2,\sigma_2^2)$. The investigator does not know which particular observation was generated by which regression equation; he only observes a vector Y with n $(=n_1+n_2)$ elements and a matrix X with

---

[1]See [1], [2], [3], [5].

$n \times k$ elements. A case in point might be the estimation of an investment demand function over the business cycle where for some observations only accelerator variables might be relevant whereas for others liquidity variables might be more important. The objective is to find an appropriate partition of the rows of Y and X into

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} \text{ and } \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$$

so that the two regimes may be disentangled from one another.

As stated, the problem is quite difficult. Considerable simplification is achieved by the assumption, which appears quite realistic in an economic context, that there exists some observable variable(s) z the values of which determine whether an observation is generated by the first regression equation or by the second. A specific formulation of this type of mechanism is contained in [5] according to which Nature chooses Regimes 1 and 2 with probabilities $\lambda(z)$ and $1-\lambda(z)$ respectively. Denoting by x_i the row vector representing the ith observation on the independent variables and by y_i the ith observation on the dependent variable, the probability density function of the ith observation is

$$h(y_i|x_i) = \frac{\lambda(z_i)}{\sqrt{2\pi}\sigma_1} \exp\{-\frac{1}{2\sigma_1^2}(y_i - x_i\beta_1)^2\} +$$

$$\frac{1-\lambda(z_i)}{\sqrt{2\pi}\sigma_2} \exp\{-\frac{1}{2\sigma_2^2}(y_i - x_i\beta_2)^2\} . \qquad (1\text{-}3)$$

The corresponding log likelihood function is

$$L = \sum_i \log h(y_i | x_i) \tag{1-4}$$

which may be maximized with respect to $\beta_1, \sigma_1^2, \beta_2, \sigma_2^2$.

The purpose of the present paper is (1) to introduce a simpler, least-squares approach to estimating the parameters under weaker assumptions,[2] and (2) to investigate the maximum likelihood estimator (in contrast with the least squares estimator) on the assumption that $\lambda(z)$ is the cumulative normal integral. Section 2 is devoted to the theoretical description of the least squares model and Section 3 contains the results of some sampling experiments.

2. Theoretical Description

The ith observation is generated either by

$$y_i = x_i\beta_1 + u_{1i} \tag{2-1}$$

or by

$$y_i = x_i\beta_2 + u_{2i} \tag{2-2}$$

where $E(u_{1i}) = E(u_{2i}) = 0$, $E(u_{1i}^2) = \sigma_1^2$, $E(u_{2i}^2) = \sigma_2^2$, x is independent of the u's and the u's are not necessarily normal. Define a variable D_i that has value 1 if Nature chooses (2-1) and value 0 if it chooses (2-2). Multiplying

---

[2] It is generally more straightforward to obtain nonlinear least squares estimates than to maximize an arbitrary likelihood function since minimization algorithms can exploit the special structure inherent in sums of squares. In addition, much weaker distributional assumptions are necessary for least squares than for maximum likelihood.

(2-1) by D_i and (2-2) by $1 - D_i$ and adding, the two regimes may be combined as in

$$y_i = D_i x_i \beta_1 + (1-D_i) x_i \beta_2 + D_i u_{1i} + (1-D_i) u_{2i} \tag{2-3}$$

Let z_i be the vector representing the ith observation on p variables and let ϕ be a p-element vector of unknown parameters. Assume that Nature chooses between the regimes (i.e., sets $D_i = 1$ or 0) according to whether $z_i\phi > v$ or $z_i\phi \leqq v$, where v is distributed according to $N(0,1)$.[3] Then

$$\text{Prob}\{D_i = 1\} = \text{Prob}\{z_i\phi > v\} = \int_{-\infty}^{z_i\phi} \frac{1}{\sqrt{2\pi}} e^{-\xi^2/2} d\xi \ . \tag{2-4}$$

Denoting the integral on the right by F_i, $D_i = 1$ with probability F_i and $D_i = 0$ with probability $1-F_i$. Hence $E(D_i) = F_i$ and we can write

$$D_i = F_i + \theta_i \tag{2-5}$$

where $E(\theta_i) = 0$, $\text{Var}(\theta_i) = F_i(1-F_i)$. Substituting (2-5) in (2-3) yields

$$y_i = F_i x_i \beta_1 + (1-F_i) x_i \beta_2 + w_i \tag{2-6}$$

where the error term w_i is given by

$$w_i = (F_i+\theta_i) u_{1i} + (1-F_i-\theta_i) u_{2i} + \theta_i x_i (\beta_1-\beta_2) \ . \tag{2-7}$$

---

[3]There is no loss of generality in assuming that the mean of the normal distribution is zero and the variance is unity. If the mean were not zero we could introduce a (p+1)th z variable with values equal to unity and thus absorb the mean on the left hand side. Similarly, we can scale ϕ so that the variance is unity.

The estimation of the two regimes can be accomplished by estimating (2-6). This, in turn, can be done in two ways. The first is to disregard the heteroscedasticity of the error term and to estimate (2-6) directly by minimizing the sum of squares $\sum_{i=1}^{n} (y_i - F_i x_i \beta_1 - (1 - F_i) x_i \beta_2)^2$ with respect to β_1, β_2 and ϕ.[4] The second is to assume again a particular distribution for u_1 and u_2, say the normal, and then derive the likelihood function for (2-6) which after some manipulation may be shown to be identical with (1-4) with $\lambda(z_i)$ being replaced by F_i .

3. Some Sampling Experiments

The sampling experiments employed the equation

$$y_i = a_1 + b_1 x_i + u_{1i}$$

for Regime 1 and

$$y_i = a_2 + b_2 x_i + u_{2i}$$

for Regime 2. A single z variable was used and a sample of uniformly distributed z-values was employed over all replications of a given experiment. Similarly, the x-values used throughout the replications of a given experiment were drawn from a uniform distribution over the (0,20) interval. The true values of the parameters were $a_1 = 1.0$, $b_1 = 1.0$, $a_2 = 0.5$, $b_2 = 1.5$, $\phi = 2.0$. Other parameters varied from experiment to experiment. The variable aspects of each case are given in Table 1.

---

[4]Since θ_i is independent of x_i, the regressors are not correlated with the error term.

TABLE I. Characteristics of Sampling Experiments

	n	σ_1^2	σ_2^2	z-range
Case 1	30	2.0	2.5	-2.0 to 2.0
Case 2	60	2.0	2.5	-2.0 to 2.0
Case 3	90	2.0	2.5	-2.0 to 2.0
Case 4	30	2.0	25.0	-2.0 to 2.0
Case 5	30	2.0	2.5	-1.0 to 3.0

For each replication of a case, Nature would compute for the ith observation $(i=1,\ldots,n)$ the quantity $z_i\phi$ and compare it to a standard normal deviate v. If $z_i\phi$ was greater than v, y_i would be generated from the first regime; if $z_i\phi \leqq v$, y_i would be obtained from the second regime. In Cases 1, 2, 3 and 5 there is substantial overlap of the scatter diagrams from the two regimes. In Case 4 the overlap is nearly complete. In this respect the separation of the data into two regimes by inspection is less easy than in the sampling experiments reported in [2] and [5]. The experiments were replicated 50 times for each case. Minimization of the sum of squares and maximization of the likelihood function was accomplished by Powell's conjugate gradient algorithm [4]. In the computation of least squares estimates the algorithm failed to produce a true minimum in one instance in Case 1 and in 9 instances in Case 4. In the computation of maximum likelihood estimates a true maximum was not arrived at in 6 instances in Case 1, 2 instances in Case 4 and 12 instances in Case 5. The overall computational failure rate of 20 per cent is similar to that reported in [5].

Tables 2 and 3 contain the mean estimates and the mean square errors of the estimates. First, it is to be noted that the mean values and mean square errors for ϕ are frequently very large in absolute value. These large values are almost invariably due to one, two or three outliers. Thus, for example, the individual estimates for ϕ by least squares in Case 2 are all between 0 and 173 except for one which is 6.7×10^4. Adjusting the mean square error of the least squares estimate in Case 1 for these outliers produces an adjusted figure of 14.45 instead of 1.58×10^5; adjusting the mean square error of the maximum likelihood estimate in Case 2 for two outliers produces 86.53 rather than 1.84×10^3. The sampling variance of $\hat{\phi}$ is obviously large but rare outliers are responsible for nearly all of it. Large mean square errors are also obtained throughout Case 4 which is to be expected in view of the large residual variance of the second regime.

The maximum likelihood method exhibits a smaller mean square error than the least squares method for every case and coefficient and is thus uniformly superior. The mean square errors also decline, for both least squares and maximum likelihood, with sample size, as is to be expected, except for the coefficient ϕ . It is interesting to note that in Case 5 maximum likelihood estimation yields smaller mean square errors for Regime 1 than does Case 1 and larger ones for Regime 2 than Case 1. This is because in Case 5 the number of observations generated by Regime 1 averages 75 percent of the total, in contrast to the 50 percent share of Regime 1 in Case 1. Finally we note the measure Δ reported in Table 2 which is defined as $\sum_{i=1}^{n} |D_{i,true} - \hat{F}_i|/n$ and may be called the mean classification error. Δ is also uniformly smaller for maximum likelihood than for least squares and also declines for larger samples.

TABLE 2. Mean Estimates

	a_1		b_1		a_2		b_2		ϕ		Δ	
	LS	ML	LS	ML	LS	ML	LS	ML	LS	ML	LS	ML
Case 1	2.1185	1.1353	.6876	.9952	-.0854	.8638	1.7913	1.4737	97.3238	3.5628	.1693	.1539
Case 2	1.1088	1.0017	.9607	1.0034	.5857	.7468	1.5993	1.4847	1.3578×10^3	12.0621	.1551	.1398
Case 3	1.1389	1.1846	.9792	.9867	.5616	.5841	1.5137	1.4955	1.1275	8.0912	.1375	.1271
Case 4	20.9664	1.0115	2.5230	1.0074	-15.1818	4.9010	-.3364	1.1153	1.2868×10^3	4.0008	.2088	.1454
Case 5	1.3925	1.1629	.8665	.9932	.6361	1.1834	1.6234	1.4073	5.1070	3.3822	.1576	.1118

TABLE 3. Mean Square Errors

	a_1		b_1		a_2		b_2		ϕ	
	LS	ML	LS	ML	LS	ML	LS	ML	LS	ML
Case 1	21.0914	1.1367	2.8063	.0083	23.5198	2.2431	2.9177	.0227	1.52902×10^5	16.0732
Case 2	.8139	.5586	.0173	.0048	1.8087	1.5557	.0320	.0134	9.1006×10^7	1.8427×10^3
Case 3	.2727	.2316	.0054	.0022	.6779	.6001	.0095	.0047	6.1780×10^7	6.0403×10^2
Case 4	7.2363×10^3	1.3201	2.6698×10^2	.0089	7.0624×10^3	1.8820×10^2	2.8120×10^2	1.9891	1.5973×10^5	24.2011
Case 5	3.9244	.8794	.1895	.0058	7.2518	7.0757	.2004	.0852	2.2432×10^2	36.7120

A final statistic we report in Table 4 is the fraction of replications for each coefficient and case in which the maximum likelihood method performs better (comes closer to the true value) than the least squares method. The percentage-win figures are without exception greater than .5. Performing a

TABLE 4. Percentage Win Statistics for Maximum Likelihood

	a_1	b_1	a_2	b_2	ϕ
Case 1	.651*	.698*	.605	.512	.512
Case 2	.720*	.740*	.580	.580	.620*
Case 3	.600	.680*	.540	.520	.620*
Case 4	.744*	.795*	.667*	.615	.820*
Case 5	.526	.632	.579	.632	.737*

one-tailed test of the hypothesis that the true win statistic is .5 on the .05 level yields 12 significant entries in Table 4, (marked by an asterisk).[5] Two more (for b_1 and b_2 in Case 5) are only .001 from the critical value. It is interesting that the percentage-win figures do not improve from Case 2 to Case 3; it suggests that the two methods are asymptotically similar. The best performance of maximum likelihood is in Case 4 in which it is intrinsically most difficult to separate the two regimes.

---

[5]This disregards to obvious dependence of the percentage-win statistics in a given row of the table.

4. Conclusion

The present paper has examined a least squares and a maximum likelihood formulation of the problem of separating two regimes. In a variety of illustrative sampling experiments the maximum likelihood method appears superior to the least squares method which may be partially attributable to the fact that the maximum likelihood estimator uses some additional information. Since, in addition, the problem of testing hypotheses is solved more satisfactorily by appealing to asymptotic considerations within the maximum likelihood framework than within the nonlinear least squares framework, the maximum likelihood approach appears to be distinctly preferable.

REFERENCES

[1] Farley, J. U. and M. J. Hinich, "A Test for a Shifting Slope Coefficient in a Linear Model," Journal of the American Statistical Association, 65 (1970), 1320-1329.

[2] Goldfeld, S. M. and R. E. Quandt, Nonlinear Methods in Econometrics, North-Holland, forthcoming.

[3] McGee, V. E. and W. T. Carleton, "Piecewise Regression," Journal of the American Statistical Association, 65 (1970), 1109-1124.

[4] Powell, M. J. D., "An Efficient Method for Finding the Minimum of a Function of Several Variables without Calculating Derivatives," Computer Journal, 7 (1964), 155-162.

[5] Quandt, R. E., "A New Approach to Estimating Switching Regressions," Journal of the American Statistical Association, forthcoming.