DETECTING MISSPECIFICATION OF FUNCTIONAL FORM IN HEDONIC PRICE EQUATIONS USING RECURSIVE RESIDUALS

NEIL R. SUMNER

Division of Water Resources, CSIRO, Private Bag, Wembley, W.A. 6014

Abstract

Recursive residuals are used to detect misspecification of functional form in a hedonic price function for rural properties in the Western Australian wheat belt. Normal probability plots, plots of cumulative sums of recursive residuals and cumulative sums of squares of recursive residuals and the Quandt lambda statistic revealed market stratification in the sample data when ordered on time. Significant changes in the mean price per hectare and residual variance with time confirm the presence of two distinct submarkets requiring different hedonic price functions. The analysis is repeated to verify that this functional form is correctly specified and that the model assumptions are satisfied. The use of recursive residuals, calculated on adaptively-ordered observations, helps detect model errors that could not be identified by the analysis of ordinary least squares residuals. The technique is equally useful in developing other econometric models based on multiple linear regression.

INTRODUCTION

The hedonic price function uses the relationship between the prices and attributes of a good to explain differences in the price of the product by providing estimates of marginal attribute prices. The hedonic price equation is estimated from market data and has the following form:

$$P = f(A_i) \tag{1}$$

where P is the price of the good and A_i denotes the ith attribute of the good.

Economic theory does not suggest any particular form for this function. The choice of the best functional form is empirical and unique to a particular product. Having the correct multiple regression model does however require that the functional form for each attribute is correctly specified and is linear in relation to the price of the product. These relationships are often nonlinear or conceal structural shifts. It is important that such features are identified as part of the analysis of the data and are adequately represented by the regression model. Among other assumptions the regression model assumes that the hedonic price function is correctly specified with all relevant attributes being included in the regression equation, and that the model error terms are Normally, Independently, Distributed with zero mean and constant variance σ^2 [NID $(0,\sigma^2)$].

Examination of Ordinary Least Squares (OLS) residual plots has often been used as a means of detecting various types of disagreement between data and an assumed model. However, the OLS residuals for a given regression problem are not well suited to model error diagnosis (Galpin and Hawkins, 1984): they are correlated; they may not have the same variance; and their distribution is dependent upon the observations matrix. Informative patterns in OLS residual plots can also be hidden by the general level of scatter, particularly if the data set is large.

Some of the earliest work on model error diagnosis in multiple linear regression is due to Quandt (1959). He devised a maximum likelihood procedure for estimating the change point and the model parameters when one structural shift occurs within the range of the data. Recursive residuals have been used as a tool for checking model assumptions (Brown, Durbin and Evans, 1975; Galpin and Hawkins, 1984; Freeman, 1986). These residuals have three useful properties: (1) if the model is correct, the recursive residuals are NID $(0,\sigma^2)$; (2) they are not constrained to sum to zero; and (3) the rth recursive residual is a function of the first r data points alone. Therefore, recursive residuals are well suited to the detection of model misspecification in problems where the data are ordered by the variable undergoing investigation.

This paper describes how recursive residual techniques were used to detect the misspecification of functional form in a hedonic price function for rural properties in the Western Australian wheat belt. A FORTRAN 77 program called RECRES (Bates and Sumner, 1990) was used for the analysis.

THEORY

The linear regression model is defined as

$$y_r = x_r^T \beta + e_r \qquad r = 1,\ldots,n, \tag{2}$$

where y_r is the rth observations on the dependent variable, x_r is the $p \times 1$ vector of observations on the p independent variables (including the intercept term), β is the $p \times 1$ parameter vector, and e_r is the error term.

Recursive residuals are standardized one-step-ahead prediction errors. The rth recursive residual is the error in predicting the rth observation using parameters estimated from a linear regression of the first r-1 observations. The data must be ordered on the variable being investigated. If the true functional form of this variable is non linear this will be apparent as one or more abrupt shifts in the residual variance. The recursive residuals are defined as

$$w_r = (y_r - x_r^T b_{r-1}) / (1 + x_r^T (X_{r-1}^T X_{r-1})^{-1} x_r)^{1/2} \quad r=p+1,\ldots,n \tag{3}$$

where b_r is the least squares estimate of β based on the first r observations and X_r is a $r \times p$ matrix consisting of the first r sets of observations on the independent variables.

An additional relationship is also available:

$$S_r = S_{r-1} + w_r^2 \tag{4}$$

where S_r is the residual sum of squares based on the first r observations. The cumulative sum (cusum) of recursive residuals can be used to detect misspecification of functional form (Brown, Durbin and Evans, 1975). This is revealed by a sudden downsurge or upsurge in the cusum plot (Galpin and Hawkins, 1984). The cusum of recursive residuals is calculated as follows:

$$C_1(r) = \sum_{i=p+1}^{r} w_i / s \qquad r=p+1,\ldots,n \qquad (5)$$

where $s^2 = S_n$.

The cumulative sum of squares of recursive residuals can also be used to detect misspecification of functional form (Brown, Durbin and Evans, 1975). If the true functional form of the ordered variable is nonlinear the cusum of squares will show abrupt shifts. The cusum of squares may be calculated as follows:

$$C_2(r) = S_r / S_n \qquad r=p+1,\ldots,n \qquad (6)$$

Quandt (1958) proposed a maximum likelihood method for estimating the parameters in two separate regression equations that switch at an unknown change point k $[p+1 < k < n-(p+1)]$. The location of k can be estimated using the statistic

$$\lambda_r = r \log s_1 + (n-r) \log s_2 - n \log s_3 \qquad r=p+1,\ldots,n-(p+1) \qquad (7)$$

where s_1^2, s_2^2 and s_3^2 are the ratios of the residual sum of squares to number of observations when a linear regression is fitted to the first r observations, the remaining n–r observations, and the whole set of n observations, respectively. The estimate of the change point's position is the value of r at which λ_r attains its minimum.

Computer programs such as SHAZAM (White, 1983) and RECRES are able to calculate recursive residuals. RECRES displays plots showing cusums of recursive residuals, cusums of squares of recursive residuals and the Quandt lambda statistic. In addition to these plots RECRES uses a V-mask (Galpin and Hawkins, 1984) on the cusum of recursive residuals plot to test for functional misspecification. The user is warned when a model misfit is detected. The program also informs the user of the observation where the nonlinearity was detected.

ECONOMIC ANALYSIS

The market data used contains numerous variables related to property sales for the period May 1986 to June 1989 for two districts in the Western Australian wheat belt. A hedonic price function predicting property price in dollars per hectare was required. A special consideration was the value of the public scheme water supply to these properties; 64 properties from the 131 sales in that period had access to scheme water. The marginal attribute price for access to scheme water was required for a cost benefit analysis of extending the scheme water supply.

The attributes with parameter estimates significantly different to zero at the 0.05 level using a two-tailed t test were farm size in hectares (AREA), rainfall (RAIN), value of buildings per hectare (BUILD), presence of scheme water (SCHEME), and a salt variable identifying properties with a shallow saline salt water table (SALT). The last two variables are indicator variables having a value of 0 or 1. The following regression equation was obtained:

$$\text{PPHA} = -0.037782\ \text{AREA} + 3.8712\ \text{RAIN} + 1.8008\ \text{BUILD} + 29.103\ \text{SCHEME} - 95.977\ \text{SALT} - 989.75 \qquad (8)$$

where PPHA is the sale price expressed in dollars per hectare. An examination of the correlation matrix (not shown) did not reveal the presence of multicollinearity. Figure 1 shows the plot of OLS residuals versus predicted values generated by the RECRES program. The most notable features of the plot are the three large positive residuals at high predicted values, the two negative predicted values, and the presence of slight heteroskedasticity. Heteroskedasticity often occurs in hedonic price functions because the variance in selling prices differs between the low-end and the high-end of the market. However, the heteroskedasticity in this data set was found to be not significant at the 0.05 level using a test developed by Breusch and Pagan (Kennedy, 1985, p. 108).

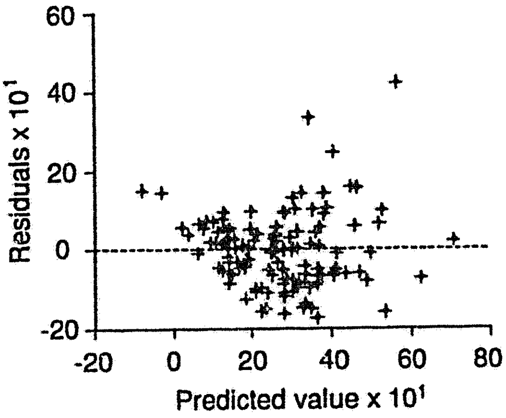

Figure 1. Ordinary least squares residuals plot for market data from the Western Australian wheat belt.

A normal probability plot of the recursive residuals, obtained by forward recursion on the data ordered by time in months (TIME), appears in Figure 2. The first six data points were used as the base for the regression. If all assumptions of the model are satisfied, the plot should show a straight line through the origin. The failure of the line to pass through the origin shows the mean of these recursive residuals is not zero, and indicates either a model misfit such as an omitted variable or the existence of outliers in the base points. Changing the base points and redrawing the plot indicated that the problem is due to a model misfit. Another feature of the plot is the existence of three outliers with large recursive residuals.

Figure 3 shows the cusum of recursive residuals plot also obtained by forward recursion. This plot is used for checking the assumption of no change of scale over the data set (Hawkins, 1981). In the case of change of scale the plot will show straight line segments. When the null hypothesis holds, we should expect the plot to take a random walk about the X axis. Figure 3 shows an abrupt change of scale at the 23rd data point. During the

Market data - W.A. wheat belt

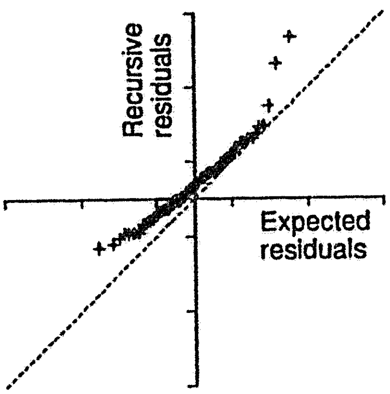

Figure 2. Normal probability plot of recursive residuals for market data.

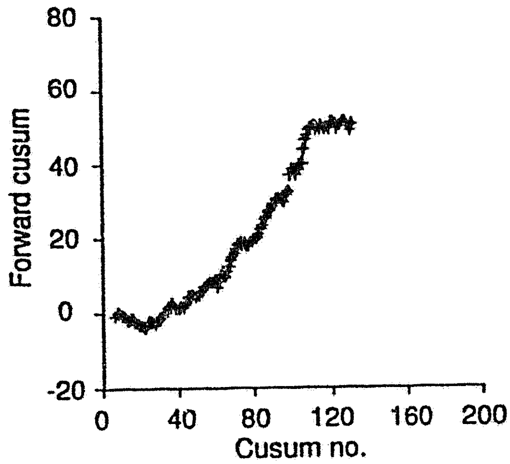

Figure 3. Cusum of recursive residuals for market data.

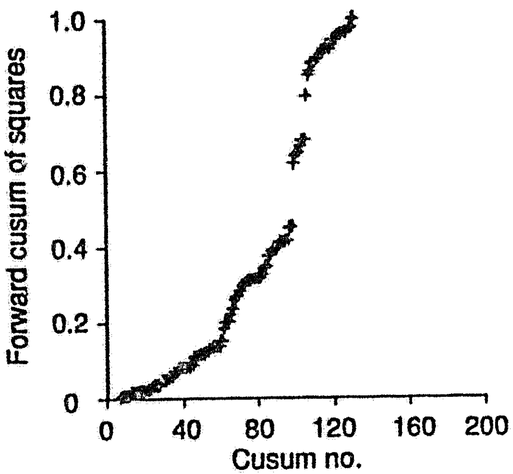

Figure 4. Cusum of squares of recursive residuals for market data.

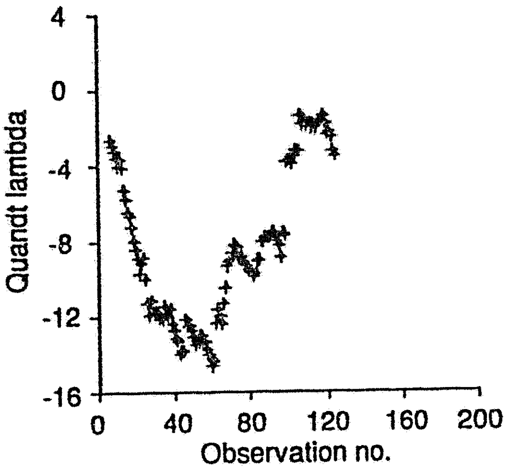

Figure 5. Quandt log-likelihood ratio for market data.

calculation of the cusum of recursive residuals the V-mask test (Galpin and Hawkins, 1984) triggered a signal at the 68th data point indicating a significant model defect.

Figure 4 shows the corresponding cusum of squares of recursive residuals plot. This plot reveals a increase in the residual variance after the 61st data point which appears as a change of slope. This instability is significant at the 0.01 level using the procedure given by Brown, Durbin and Evans (1975, p. 155). Another feature of the plot is the existence of three outliers, the 99th, 106th and 107th observations which cause jumps in the cusum of squares between consecutive data points.

The Quandt lambda plot in figure 5 suggests that the postulated relationship changes over the range of the data. This change in relationship is estimated to occur at the 61st data point where λ_r obtains its minimum.

Overall these plots suggest that there is an important variable omitted from the model and there are structural changes in the data occurring at about the 21st and 61st data points corresponding to December 87 and September 88, respectively. A plot of TIME (time in months) by PPHA shown as figure 6 reveals that PPHA increases with time for the period of the study. This plot shows noticeable jumps in PPHA for December 87 and September 88 also suggesting the existence of submarkets.

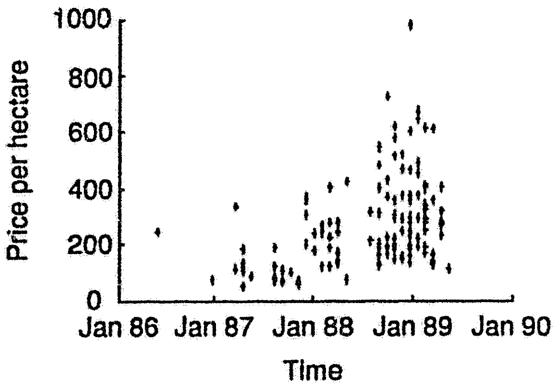

Figure 6. Changes in the price per hectare value of properties with time.

The analysis was repeated ordering the data on the AREA, RAIN and BUILD variables separately (plots not shown). Both the AREA and RAIN variables required transformations. A transformation described by Box and Cox (1964) was used to improve the functional form of the these independent variables. Although TIME is not an attribute of a hedonic price function this variable was included in the model to see if its inclusion helped to correct the previous misspecified model. Two of the outliers previously identified were removed; the 99th data point was no longer present as an outlier following the transformations. The improved model is:

$$PPHA = 36432\ AREA^{-1} + 0.0050692\ RAIN^{2} + 1.5022\ BUILD + 36.367\ SCHEME - 73.674\ SALT + 4.5446\ TIME - 512.21 \quad (9)$$

The plots shown in figures 1 to 5 were repeated for this model with the data ordered on TIME (not shown). Removing the two outliers corrected plots 1, 2, and 4 on which the outliers were present. The negative predicted values shown on figure 1 were no longer present. The inclusion of the TIME variable corrected the normal probability plot shown in figure 2 indicating that it was an omitted variable. The parameter estimate for this variable was significant at the 0.01 level indicating the existence of changing market conditions with time. The market stratification shown in figures 3, 4 and 5 was still present, occurring in December 1987 and May 1988. The improved model gave a different estimate of the second structural change, this is because the regime change involved a transition period during which the relationship changes smoothly.

Since the plots of recursive residuals have indicated the possible existence of three submarkets separate models each describing the hedonic price function for that period of time were formulated. The three hedonic price functions were for the periods May 1986 to November 1987, December 1987 to April 1988 and May 1988 to June 1989. Chow F tests (Kennedy 1985, p. 87) were used to examine whether the parameter estimates for the markets differed significantly between the submarkets at the 0.01 level. The submarket for the period September 1988 to June 1989 was significantly different from the other two submarkets, although the submarket from May 1986 to November 1987 was not significantly different from the submarket for December 1987 to April 1988. Therefore two separate submarkets were assumed to have existed from May 1986 to April 1988 and May 1988 to June 1989.

The parameter estimates obtained for each submarket are shown in table 1. The TIME variable was not significant for either submarket at the 0.05 level and was not included in the regression model. A two-tailed t test was used to compare the parameter estimates and the mean predicted PPHA for submarket one with those obtained for submarket 2. A two-tailed F test was used to compare the residual variance for the two submarkets.

Table 1. Changes in the hedonic price function between submarkets

	Attribute parameter estimates		
	Submarket 1	Submarket 2	Significance*
$AREA^{-1}$	26531.	40956.	ns
$RAIN^{2}$	0.0042424	0.0053243	ns
BUILD	1.5071	1.3113	ns
SCHEME	29.645	41.019	ns
SALT	6.1360	-87.482	ns
Constant	-327.21	-370.62	ns
Mean PPHA	175.95	308.66	s
Mean square error	2124.	5117.	s
No. observations	44	85	
R^2	79.3	81.5	

* Significance levels for difference between submarkets:
ns not significant at the 0.05 level
s significant at the 0.01 level

Submarket 2

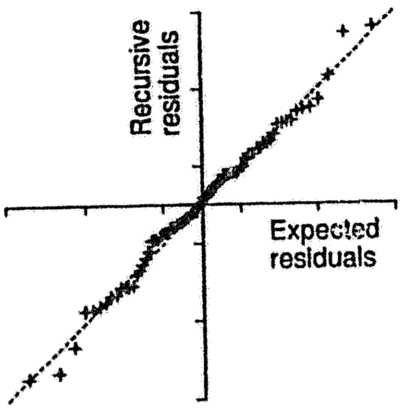

Figure 7. Normal probability plot of recursive residuals for submarket 2.

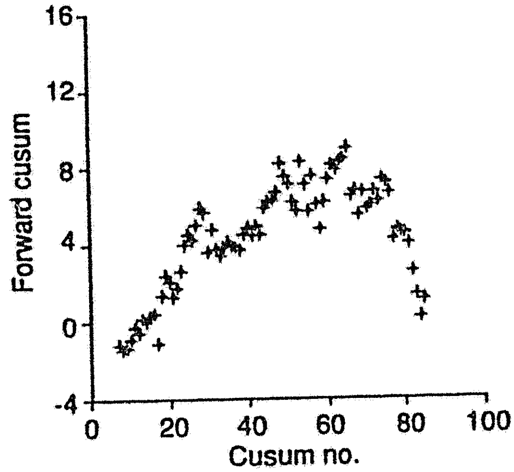

Figure 8. Cusum of recursive residuals for submarket 2.

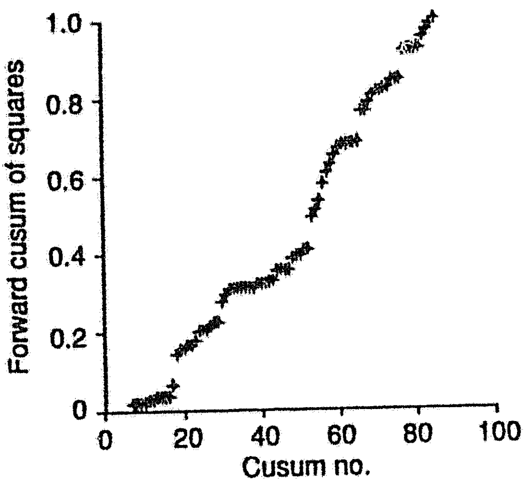

Figure 9. Cusum of squares of recursive residuals for submarket 2.

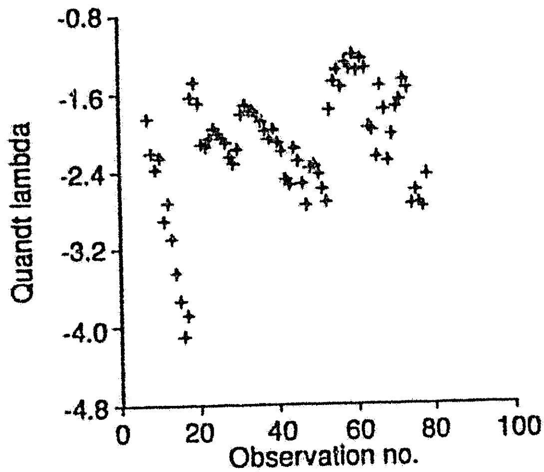

Figure 10. Quandt log-likelihood ratio for submarket 2.

The parameter estimates for SCHEME were 29.645 and 41.019 for submarkets 1 and 2 respectively as shown in table 1. There is no significant difference between the marginal price of access to scheme water between submarkets. These estimates give an indication of the value placed on access to scheme water by farmers in dollars per hectare.

The changing market conditions between the two submarkets could not be explained by changes in the parameter estimates as they were not significantly different at the 0.05 level. However, the existence of heterogeneity between submarkets was evident as the mean PPHA and residual variance were significantly different. Since the change in market conditions could not be explained by changes in the parameter estimates for the known product attributes it can be assumed that the observed differences are due to changes in the marginal prices for other unknown attributes such as the quantities of crops, livestock, and machinery present on the properties at the time of the sale.

Analysis of recursive residual plots with the data ordered on TIME for submarkets 1 (not shown) and 2 (shown as figures 6 to 10) suggests that the hedonic price function is now correctly specified and the model assumptions are satisfied. The normal probability plot for submarket 2 (figure 7) shows a straight line through the origin. The cusum of recursive residuals (figure 8) plot did not indicate any abrupt change of scale and the V-mask test was not significant. However, the appearance of this plot is not completely random with the occurrence of a slight initial upsurge caused by the transition from submarket one to submarket two. The cusum of squares of recursive residuals plot (figure 9) showed reasonably constant residual variance throughout the submarket; there was no significant instability at the 0.05 level. The Quandt lambda plot (figure 10) showed that the postulated relationship was consistent for the full range of the data.

CONCLUSIONS

In this paper a hedonic price function is developed for rural properties in the Western Australian wheat belt. Recursive residuals calculated on adaptively-ordered observations were used to look for misspecification of functional form and were useful in detecting market stratification. The use of normal probability plots, plots of cusums of recursive residuals and cusums of squares of recursive residuals and the Quandt lambda statistic helped the detection of model errors that could not be identified by the analysis of ordinary least squares residuals.

Two separate submarkets were found to exist in the sample data violating single market assumptions and estimation procedures. A hedonic price function was developed for each submarket and a two-tailed t test was used to test for heterogeneity of parameters between submarkets. The analysis revealed that the market stratification was due to significant differences in the PPHA values and residual variance between submarkets and could not be attributed to changes in the parameter estimates.

The technique is equally useful in developing other econometric models based on multiple linear regression.

A computer program with additional examples supporting the use of this methodology are available from the author on request; the program is written in FORTRAN 77 and was developed on an IBM personal computer.

ACKNOWLEDGEMENTS

Thanks are due to Tim Coelli from the Western Australian Department of Agriculture for making the data available. I am also in debt to Bryson Bates and Robin Barron, CSIRO Division of Water Resources and Tony Grassia, CSIRO Laboratory for Rural Research whose consultation on the statistical analysis is greatly appreciated.

REFERENCES

Bates, B.C. and Sumner, N.R., 1990, RECRES: A FORTRAN 77 program for detecting model misspecification in multiple linear regression: Comp. Geosc. (submitted).

Box, G.E.P. and Cox, D.R., 1964, An analysis of transformations: Jour. Royal. Statist. Soc. Ser. B, v. 26 p. 211-243.

Brown, R.L., Durbin J., and Evans, J.M., 1975, Techniques for testing the constancy of regression relationships over time: Jour. Royal Statist. Soc. Ser. B, v. 37, p. 149-163.

Freeman, J.M., 1986, A piecewise regression analysis for firm-size distributions: The Statistician, v. 35, p. 547-559

Galpin, J.S. and Hawkins, D.M., 1984, The use of recursive residuals in checking model fit in linear regression: The Am. Statistician, v. 38, no. 2, p. 94-105.

Hawkins D.M., 1981, A cusum for a scale parameter: Jour. Quality Tech., v. 13, p. 228-231.

Kennedy, P., 1985, A guide to econometrics: Basil Blackwell, Oxford, 238p.

Quandt, R.E., 1958, The estimation of the parameters of a linear regression system obeying two separate regimes: Jour. Am. Statist. Assoc., v. 53, p. 873-880.

White, K.J., 1983, SHAZAM - An econometrics computer program: University of British Columbia, Vancouver, 121p.