INTERNATIONAL REGIONAL SCIENCE REVIEW 26, 2: 208–217 (April 2003)

SPATIAL AUTOCORRELATION OR MODEL MISSPECIFICATION?

DANIEL P. MCMILLEN
Department of Economics (MC 144),
University of Illinois at Chicago, Chicago, IL,
mcmillen@uic.edu

A Monte Carlo experiment illustrates one of the pitfalls of spatial modeling. Specification tests indicate spatial autocorrelation when functional form misspecification is actually the only problem with the model. Incorrect functional forms and omitted variables that are correlated over space produce spurious spatial autocorrelation.

Keywords: *spatial autocorrelation; Lagrange multiplier (LM) tests; nonparametric; kernel regression*

A common estimation procedure for spatial data is to start with a simple specification, which then is subjected to a series of tests for spatial autocorrelation. If the tests reject the assumption of independently distributed errors, a more complicated model is specified to account for the autocorrelation. Alternatively, estimation may begin with the more complicated model, which becomes the final specification if the tests do not reject simpler models. Either estimation procedure can produce consistent and efficient parameter estimates if the spatial model is specified correctly.

Although specification tests of statistical models can reveal that something is wrong with the base model, they do not reveal what to do about it. Rejection of a simple model does not imply that the more elaborate model is correct; it simply indicates that something is wrong with the simpler model. This simple point is well known but commonly ignored in practice. A general rule is that the more complicated the alternative model, the less likely it is to be subjected to further specification testing once the base model is rejected. My proposed rule is particularly prone to being followed in spatial modeling because common alternative specifications involve maximum likelihood estimation procedures and matrix manipulations that are sufficiently computer and time intensive to make further investigation unattractive.

Tests for spatial autocorrelation also detect functional form misspecification, heteroskedasticity, and the effects of missing variables that are correlated over

DOI: 10.1177/0160017602250977

space. The problem of functional form misspecification is particularly severe in spatial modeling because theory provides little guidance and spatial relationships are often highly nonlinear. What appear to be complex spatial relationships may simply be an artifact of simple functional form specification.

In this article, I present Monte Carlo evidence of the effects of functional form misspecification on common spatial models. The context is a simple employment density function, in which log density is a linear function of distance from the city center and a secondary employment center in the suburbs. The mythical researcher either neglects the subcenter altogether or assumes an incorrect functional form for its effects on densities. Though spatial autocorrelation tests routinely reject the null hypothesis of independent errors, the rejections are in fact an indication that the assumed functional form is incorrect. Otherwise, the estimated models appear to fit the data well. These results are a warning that estimated spatial effects might be an artifact of undetected model misspecification even in a seemingly well-specified model.

THE MONTE CARLO SETUP

A stylized monocentric city serves as the basis for the Monte Carlo study. The base model is $y_i = \beta_0 + \beta_1 DCBD_i$, where y_i is the natural logarithm of employment density at location i, and $DCBD$ is distance from the central business district (CBD). There are one thousand observations. To keep estimation simple and to allow the results to be displayed in simple diagrams, all observations are along a line through the CBD. The distances, x_i, range from –50 to 50 in equal increments of 0.1001. The CBD is located at $x_i = 0$, and $DCBD_i = |x_i|$. To be consistent with a large American metropolitan area, the distances are measured in miles. The parameter values for all experiments are $\beta_0 = 1.5$ and $\beta_1 = -0.05$.

The employment subcenter is located at $x = -30$. Its effects are confined to a 6-mile radius. Employment density reaches a local peak at this distance and declines with increases in $DSUB_i = |-30 + x_i|$ over the range $-36 < x_i < -30$. In this range, the log-density function is $y_i = \beta_0 + \beta_1 DCBD_i + (\beta_2 + \beta_3 DSUB_i) + u_i$, with $\beta_2 = 0.9$ and $\beta_3 = -0.15$. After defining a dummy variable SUB_i that equals 1 when $-36 < x_i < -30$, the full model is

$$y_i = 1.5 - 0.05 DCBD_i + 0.9 SUB_i - 0.15 DSUB_i \times SUB_i + u_i. \quad (1)$$

Independent error terms are drawn from a normal distribution with constant variance of .3623, which implies that the R^2s will average .80 across the replications.[1]

I used this model to help guide my choice of procedure for identifying subcenters (McMillen 2001). Whereas the location of the CBD is known beforehand when modeling an urban area, the number and the locations of subcenters are not. Several procedures for identifying subcenters have been proposed.[2] They share an objective of finding local peaks in employment density functions of the type illustrated by the "OLS: True Model" line in Figure 1.

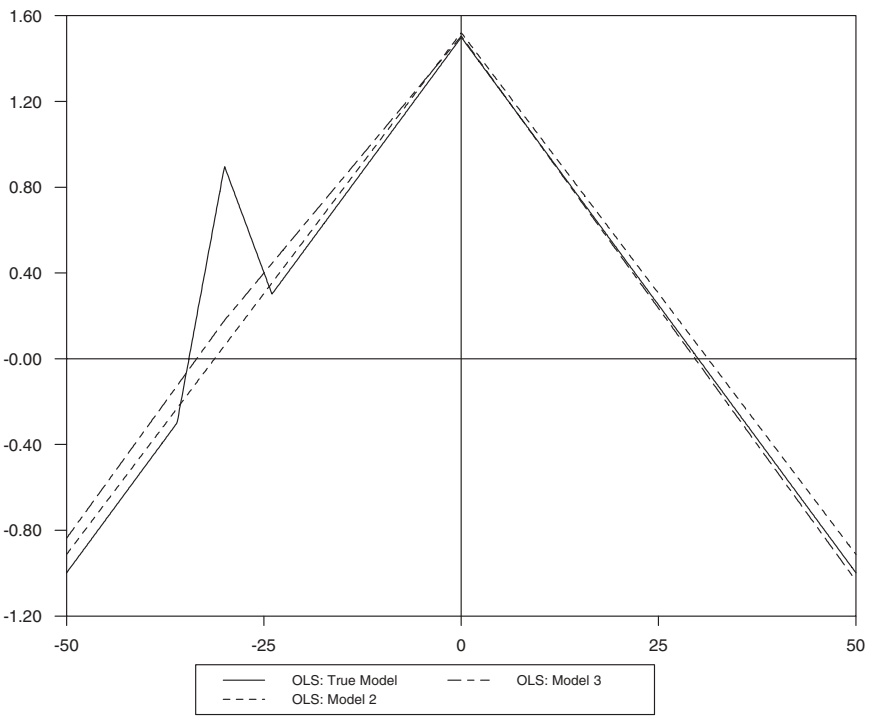

FIGURE 1. Predictions from Ordinary Least Squares (OLS) Models

If the model were known beforehand, a regression of y on *DCBD*, *SUB*, and $DSUB \times SUB$ would provide consistent and efficient estimates. *DSUB* is difficult to define in practice because the number of subcenters and their locations are unknown. Even if the locations were known, the extent of a subcenter's influence is not easy to determine. Although Figure 1 is illustrative of the type of function researchers have in mind when modeling polycentric cities, a more common regression specification has *DCBD* and *DSUB* as the explanatory variables. This specification implies that distance from the subcenter affects densities throughout the city rather than in a window of observations around $x = -30$.

Spatial autocorrelation and heteroskedasticity have been omitted from this model. Standard specification tests will not detect either problem in a correctly specified model (except by Type I error). In the Monte Carlo experiments, I use a version of the Breusch and Pagan (1979) and White (1980) Lagrange multiplier (LM) tests to detect heteroskedasticity. Squared residuals are regressed on an intercept, *DCBD*, *DSUB*, $DCBD^2$, $DSUB^2$, and $DCBD \times DSUB$, and a standard F test assesses the joint significance of the five explanatory variables.

I also use an LM test to detect spatial autocorrelation (Anselin et al. 1996; Burridge 1980). The base spatial model is the autoregressive model

$$Y = \rho WY + X\beta + u, \tag{2}$$

where W is an $n \times n$ spatial contiguity matrix ($n = 1{,}000$). I use a simple specification of W in which $W_{ij} = 1$ if $|i - j| = 1$ and $W_{ij} = 0$ otherwise. The matrix then is normalized such that each row sums to 1. With the x_i ordered from smallest to largest, this specification implies that $\{WY\}_i = (y_{i-1} + y_{i+1})/2$, except at the endpoints. The first and last values of this variable are $\{WY\}_1 = y_2$ and $\{WY\}_n = y_{n-1}$. Letting e denote the matrix of base regression residuals, the LM test statistic

$$[Tr(W'W) + Tr(WW)]^{-1}[ne'We/e'e]^2 \tag{3}$$

is distributed χ^2 with 1 degree of freedom. This expression simplifies substantially here because the structure of W implies that $Tr(W'W) + Tr(WW) = n + 1.5$.

ALTERNATIVE EMPIRICAL MODELS

I estimate five empirical models in the Monte Carlo experiments. The first is the simple regression model implied by equation 1. The second model is a simple regression of y on *DCBD*, omitting the subcenter altogether. The third model is the common empirical specification in which y is regressed on *DCBD* and *DSUB*, which implies that the subcenter affects densities throughout the urban area. Each of the first three models is estimated simply by ordinary least squares (OLS).

The fourth empirical specification is the spatial autoregressive model given by equation 2. For this model, I assume that $X = (1, DCBD)$ and estimate the model by maximizing the implied log-likelihood function (Anselin 1988). The spatial autoregressive model is used frequently in spatial data analysis. Examples of excellent applications include Anselin, Varga, and Acs (1997) and Brueckner (1998). In these studies, significantly positive values of ρ are interpreted as evidence of spatial spillovers: high values of y at one site cause high values of y at nearby sites. This interpretation assumes that the model is specified correctly. In contrast, the model here is misspecified in that *SUB* and $SUB \times DSUB$ are omitted. If the estimator adjusts to this misspecification by producing a positive value for ρ, then our omission of the subcenter leads to a mistaken finding of spillovers.

The final estimator is a version of standard kernel regression.[3] The estimated value of y at observation i is

$$\hat{y}_i = \frac{\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_j - x_i}{h}\right) y_j}{\frac{1}{nh}\sum_{j=1}^{n} K\left(\frac{x_j - x_i}{h}\right)}. \tag{4}$$

I use the bi-square kernel: $K(\psi_{ij}) = (15/16)(1 - \psi_{ij}^2)^2 I(|\psi_{ij}| < 1)$, where $\psi_{ij} = (x_j - x_i)/h$ and $I(\bullet)$ is an indicator function that equals 1 when the condition is true and 0 otherwise. I set $h = 5.005$, which implies that the nearest 100 observations are given positive weight for observations in the interior of the data set. This estimator is a special case of the locally weighted regression (LWR) estimator that has been used frequently in urban economics, real estate, and geography.[4] Like LWR, the kernel estimator places more weight on nearby observations when constructing an estimate for a target location. Despite being provided with no information on the subcenter, it is sufficiently flexible to detect the local rise in densities near $x = -30$.

I present average coefficient estimates and their standard deviations after one thousand Monte Carlo replications. I also present the average R^2 values for the three OLS models. I calculate the percentage of times that the LM tests for heteroskedasticity and spatial autocorrelation reject their null hypotheses.[5] Finally, I define the following measure of the overall bias in the estimated slopes, $\partial y_i/\partial x_i$:

$$\text{Slope bias measure} = \sqrt{\frac{\sum_{i=1}^{n}\left(\hat{\delta}_i - \delta_i\right)^2}{n-1}}, \tag{5}$$

which is analogous to the bias component of a root mean squared error.[6] The objective of equation 5 is to measure the overall bias in estimated marginal effects, which represent the primary interest of most empirical spatial studies.

Monte Carlo Results

The results of one thousand Monte Carlo replications are summarized in Table 1. The correctly specified model, labeled "OLS Model 1," is estimated precisely. It is indistinguishable from the base function, equation 1. The average coefficients are close to the true values, the standard deviations are low, and the R^2 is nearly the same as the expected value of .80. The other two OLS models also appear to fit the data well. Both have average R^2s in excess of .75, and all the estimated coefficients are highly significant. Model 3 appears on the surface to be a very attractive model: as expected, estimated densities decline significantly with distance from the city center, and distance from the subcenter has a less pronounced but still significant effect.

The OLS functions implied by the average Monte Carlo coefficients are shown in Figure 1. Model 3 detects a small rise near the subcenter, but neither of the two incorrectly specified OLS models comes close to accounting for the local rise in densities near the subcenter. For the correctly specified Model 1, the null hypotheses are rejected slightly more frequently than the nominal size of the test, which is 5 percent. In contrast, both heteroskedasticity and spatial autocorrelation are detected in nearly every case for both Models 2 and 3. Remember that neither of these problems is actually present. Both of the incorrectly specified models severely underestimate densities near the subcenter. Thus, positive residuals are

TABLE 1. Monte Carlo Results

	OLS Model 1	*OLS Model 2*	*OLS Model 3*	*Spatial AR*	*Kernel Regression*
Constant	1.4990 (0.0233)	1.5208 (0.0233)	1.6039 (0.0267)	1.2696 (0.0509)	
Distance to city center	–0.0499 (0.0008)	–0.0487 (0.0008)	–0.0475 (0.0008)	–0.0406 (0.0017)	
Subcenter dummy	0.8965 (0.0658)				
Subcenter Dummy × Distance to Subcenter	–0.1491 (0.0188)				
Distance to subcenter			–0.0033 (0.0005)		
WY				0.1651 (0.0311)	
R^2	.7999	.7553	.7647		
% rejections: Heteroskedasticity test	6.4	100.0	100.0	100.0	55.0
% rejections: Spatial autocorrelation test	5.7	100.0	99.4	0.0	7.6
Slope bias measure: All observations	.0003	.0520	.0509	.0520	.0273
Slope bias measure: Subcenter observations	.0009	.1506	.1473	.1506	.0683
Slope bias measure: Nonsubcenter observations	.0001	.0013	.0031	.0013	.0146

Note: The results are based on one thousand Monte Carlo replications. Standard deviations are in parentheses. OLS = ordinary least squares; AR = autoregressive.

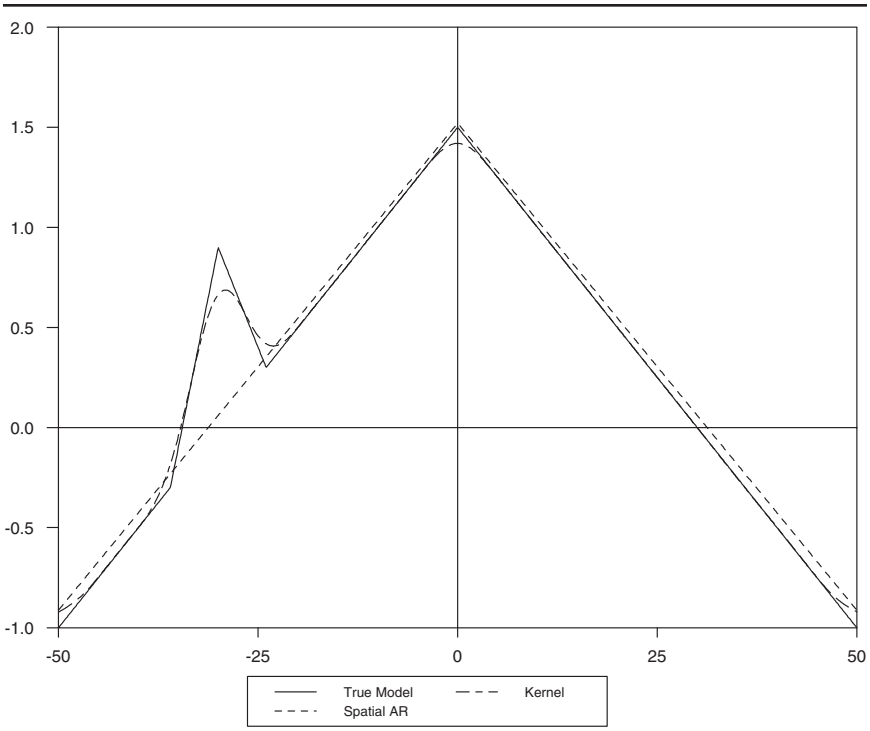

FIGURE 2. Predictions from Spatial Autoregressive (AR) and Locally Weighted Regression (LWR) Models

concentrated in this area, which produces positive autocorrelation. The subcenter has large values for the squared residuals, and the heteroskedasticity test detects this spatial pattern.

The LM tests clearly indicate that something is wrong with the incorrectly specified models. A logical next step is to try a spatial autoregressive (AR) model. These results are presented in the fourth column of results in Table 1. The model appears to work well. The estimates are highly significant, and they indicate a statistically significant value of ρ, which averages .1651 across the replications. The estimated function implied by the Monte Carlo averages is displayed in Figure 2. With only a small rise in the function near the subcenter, the spatial AR results are not much different from the results for the OLS models. Nonetheless, it would not be surprising if at this point the spatial AR model were accepted as the final model, having cured the problem of spatial autocorrelation.

Despite being given no information on the location of the subcenter, the nonparametric estimator is successful in detecting the local peak in employment density. As seen in Figure 2, the kernel estimator produces a slightly smoothed but highly accurate rise in densities near the subcenter. It also provides accurate esti-

mates at other locations, although boundary effects are evident near the endpoints of the function.[7] Although the residuals display heteroskedasticity in 55 percent of the replications, the kernel estimator effectively eliminates the spatial autocorrelation problem. The last three rows of the table show that the kernel estimator provides a much more accurate measure of the marginal effect of x than any of the misspecified models. Over all observations, the value of the slope bias measure for the kernel estimator is less than 55 percent of the value for the misspecified models. Naturally, the kernel estimator performs much better in the subcenter range than for the other observations. Overall, the kernel estimator produces a much better empirical model than the spatial AR or OLS models.

CONCLUSION

My intention in this article is not to argue that parametric spatial models should be abandoned in favor of nonparametric modeling. The simple single explanatory variable model is particularly amenable to nonparametric modeling. Estimation becomes more difficult as more explanatory variables are added to the model, and nonparametric models are subject to a "curse of dimensionality" that is akin to the loss of degrees of freedom as more explanatory variables are added to a parametric model. If prior knowledge of a model's structure is available, parametric models provide much more efficient estimates than nonparametric estimators. Furthermore, using a spatial AR model to test a simple model specification is far better than the alternative of failing to subject a model to any diagnostic testing.

My objective is to offer a precautionary warning. Autocorrelation is a common problem in spatial data, and significant advances have been made in devising parametric models that account for it. Yet autocorrelation is often produced spuriously by model misspecification. I have focused on functional form misspecification because I consider it the most likely culprit. However, what I have called a problem of incorrect functional form could just as well be considered a problem of missing variables that are correlated over space. Supplementing an incorrectly specified model with a subcenter dummy variable and an interaction term between the dummy variable and distance to the subcenter would produce an accurate model specification. Even then, matters are difficult in practice because the researcher must specify the subcenter's radius, and it is very doubtful that this information would be available. Omitted explanatory variables that are correlated over space and misspecified spatial effects will produce spatially correlated residuals, even when the true model errors are independent. Uncritical acceptance of a common alternative spatial model is not the answer to rejection of a simple null model. The answer is further investigation.

Fortunately, further investigation of modeling strategies for spatial models is under way. Using simulations, Florax, Folmer, and Rey (2002) have found that a classical forward stepwise modeling approach—subjecting a simple base model to diagnostic tests and following rejections of the base model with estimation of more

complicated models—can often detect the appropriate model. They have found that this approach outperforms a "Hendry-like" specification strategy, which starts from the more complicated model and works its way backward to simpler specifications. Graaf et al. (2001) found that they can distinguish nonlinearity from other forms of model misspecification by subjecting a model to a series of complementary specification tests. A misspecification test derived from chaos theory—the "BDS" test—is particularly powerful in detecting nonlinearity in their estimated models. Nonetheless, detecting the correct econometric model remains an intriguing and sometimes formidable challenge.

NOTES

1. Writing the model in matrix notation as $Y = X\beta + u$, one has $R^2 = \text{var}(X\beta)/[\text{var}(X\beta) + \text{var}(u)]$. With X and β set in the experiments, one is free to set a value for var(u) to achieve a desired value for R^2. With $R^2 = .8$, one has var(u) = var($X\beta$)/4, which is equal to .3623 here.

2. Notable contributions include Craig and Ng (2001), Giuliano and Small (1991), McDonald (1987), and McMillen (2001).

3. See Pagan and Ullah (1999) for an excellent discussion of kernel regression procedures.

4. Locally weighted regression was developed by Cleveland and Devlin (1988). Applications in urban economics and real estate include Fu and Somerville (2001), McMillen (1996), McMillen and McDonald (1997), Meese and Wallace (1991, 1997), and Pavlov (2000). Brunsdon, Fotheringham, and Charlton (1996) applied the model to geographical analysis and renamed it "geographically weighted regression."

5. The residuals for the spatial autoregressive (AR) model are $e_i = y_i - \beta_0 - \beta_1|x_i| - \rho\{WY\}i$, with the parameters evaluated at the estimated values. The residuals for the locally weighted regression (LWR) model are simply the difference between y_i and the expression in equation 4.

6. For the true model, the slope is .05 for $x < -36$, .20 for $-36 \le x < -30$, –.10 for $-30 \le x < -24$, .05 for $-24 \le x \le 0$, and –.05 for $x > 0$. For the spatial AR model, I calculate the slopes using the reduced form equation $Y = (I - \rho W)^{-1}X\beta$. The kernel regression slopes are simply the derivatives of equation 4.

7. Boundary effects are less pronounced in locally weighted regression, which is one of the reasons why it is used more commonly in spatial modeling than kernel regression.

REFERENCES

Anselin, L. 1988. *Spatial econometrics: Methods and models*. Boston: Kluwer Academic.

Anselin, L., A. Bera, R. Florax, and M. Yoon. 1996. Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics* 26: 77-104.

Anselin, L., A. Varga, and Z. Acs. 1997. Local geographic spillovers between university research and high technology innovations. *Journal of Urban Economics* 42: 422-48.

Breusch, T., and A. Pagan. 1979. A simple test for heteroskedasticity and random coefficient variation. *Econometrica* 47: 1287-94.

Brueckner, J. K. 1998. Testing for strategic interaction among local governments: The case of growth controls. *Journal of Urban Economics* 44: 438-67.

Brunsdon, C., A. S. Fotheringham, and M. E. Charlton. 1996. Geographically weighted regression. *Geographical Analysis* 28: 281-98.

Burridge, P. 1980. On the Cliff-Ord test for spatial autocorrelation. *Journal of the Royal Statistical Society* B42: 107-8.

Cleveland, W. S., and S. J. Devlin. 1988. Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association* 83: 596-610.

Craig, S. G., and P. T. Ng. 2001. Using quantile smoothing splines to identify employment subcenters in a multicentric urban area. *Journal of Urban Economics* 49: 100-120.

Florax, R. J. G. M., H. Folmer, and S. Rey. 2002. Specification searches in spatial econometrics: The relevance of Hendry's methodology. Manuscript, Free University, Amsterdam.

Fu, Y., and S. T. Somerville. 2001. Site density restrictions: Measurement and empirical analysis. *Journal of Urban Economics* 49: 404-23.

Giuliano, G., and K. A. Small. 1991. Subcenters in the Los Angeles region. *Regional Science and Urban Economics* 21: 163-82.

Graaf, T. de, R. J. G. M. Florax, P. Nijkamp, and A. Reggiani. 2001. A general misspecification test for spatial regression models: Dependence, heterogeneity, and nonlinearity. *Journal of Regional Science* 41: 255-76.

McDonald, J. F. 1987. The identification of urban employment subcenters. *Journal of Urban Economics* 21: 242-58.

McMillen, D. P. 1996. One hundred fifty years of land values in Chicago: A nonparametric approach. *Journal of Urban Economics* 40: 100-124.

———. 2001. Nonparametric employment subcenter identification. *Journal of Urban Economics* 50: 448-73.

McMillen, D. P., and J. F. McDonald. 1997. A nonparametric analysis of employment density in a polycentric city. *Journal of Regional Science* 37: 591-612.

Meese, R., and N. Wallace. 1991. Nonparametric estimation of dynamic hedonic price models and the construction of residential housing price indices. *Journal of the American Real Estate and Urban Economics Association* 19: 308-32.

———. 1997. The construction of residential housing price indexes: A comparison of repeat sales, hedonic regression, and hybrid approaches. *Journal of Real Estate Finance and Economics* 14: 51-73.

Pagan, A., and A. Ullah. 1999. *Nonparametric econometrics*. New York: Cambridge University Press.

Pavlov, A. D. 2000. Space-varying regression coefficients: A semi-parametric approach applied to real estate markets. *Real Estate Economics* 28: 249-83.

White, H. 1980. A heteroscedasticity-consistent covariance matrix estimator and a direct test for heteroscedasticity. *Econometrica* 48: 817-38.