Peer Groups and Bias Detection in Least Squares Regression

Blankmeyer, Eric

Texas State University

15 November 2021

Peer Groups and Bias Detection in Least Squares Regression

Eric Blankmeyer

email eb01@txstate.edu

November 2021

Abstract. A correlation between regressors and disturbances presents challenging problems in linear regression. In the context of spatial econometrics LeSage and Pace (2009) show that an autoregressive model estimated by maximum likelihood may be able to detect least squares bias. I suggest that spatial neighbors can be replaced by "peer groups" as in Blankmeyer et al. (2011), thereby extending considerably the range of contexts where the autoregressive model can be utilized. The procedure is applied to two data sets and in a simulation.

Introduction

A correlation between regressors and disturbances presents challenging problems in linear regression. Measurement error, omitted variables, and simultaneity issues exemplify situations that render ordinary least squares (OLS) biased and inconsistent (Greene 2003, 74-90, 148-149, 378-381; Gujarati 2015, 131-135, 141-142; Min 2019, 47-48, 110-113). Many proposals for consistent estimation require instrumental variables (IV), but in practice valid instruments are often elusive.

In the context of spatial econometrics LeSage and Pace (2009) show that the maximum likelihood estimator (MLE) of an autoregressive model may be able to detect OLS bias. Their model "can be used to diagnose misspecification in general, and the potential existence of omitted variables" (ibid, 61 and 63-70). In this paper I suggest that spatial neighbors can be replaced by "peer groups" as in Blankmeyer et al. (2011), thereby extending considerably the range of contexts where the autoregressive model is utilized. I apply this strategy to two data sets and in a simulation.

Blankmeyer et al. (2011, 92) remark that the literature on social networks and peer groups embraces "a wide range of topics, including spillovers and strategic interaction among governments (Case et al., 1993; Brueckner, 1998), social inequality and stratification (Durlauf, 1994), choice of college roommates (Sacerdote, 2001) and the social linkages of crime (Glaeser et al., 1996)." Methodological papers dealing with social networks include Conley and Topa (2007), Handcock et al. (2007), Bramoulle et al. (2009), Soetevent (2006), DiGiorgi et al. (2010), Dahl et al. (2014), and von Hinke et al. (2019).

The administrator's salary

A report of the Texas Health and Human Services Commission (2002) provides annual data on the administrator's salary and the gross revenue in 842 nursing facilities operated for profit. The log linear model

$$\text{salary} = \alpha + \beta\text{revenue} + z \qquad (1)$$

could show to what extent the manager's pay is based on earnings. However, the facility's owners might also benchmark the salary to revenue levels in comparable nursing homes. In equation (1) this benchmarking effect is relegated to the unobservable disturbance z so OLS is vulnerable to omitted-variable bias.

Following LeSage and Pace (2009, 27-28), I implement an autoregressive model for equation 1 by forming peer groups of 20 nursing homes based on each facility's number of beds and its normal staff size. These criteria, which are predetermined (exogenous) in relation to the annual salary and revenue, provide a heuristic for the scope and complexity of the administrator's responsibilities. In the notation of spatial econometrics, the peer groups produce a vector Wsalary of n= 842 observations: its i-th element is the average log salary of the 20 facilities that most closely match the i-th

facility's combination of beds and staff size. Likewise the i-th row of the vector Wrevenue is just the average log revenue of the same 20 peers of facility i. [Blankmeyer et al. (2011) and LeSage and Pace (2009, chapter 1) have additional discussion of the W operator.]

The motivation for these peer-group variables is an assumption that the stochastic component of equation 1 is an autoregressive process with scalar parameter $\rho \in (0,1)$:

$$z = \rho W z + u \, . \qquad (2)$$

The autocorrelation as such does not bias the OLS estimates of α and β in equation 1; but in addition the model proposes that

$$u = \gamma \text{revenue} + v. \qquad (3)$$

Here v is a vector of n independent gaussians, each having mean zero and variance σ^2; but if the scalar parameter γ is not zero then equation 3 implies that the revenue regressor is correlated with the disturbance z so that OLS estimation of equation 1 is indeed biased and inconsistent.

However the following equation can be estimated consistently by maximum likelihood (LeSage and Pace 2009, chapter 3, Bivand 2019):

$$\text{salary} = \delta + \rho W\text{salary} + \text{revenue}(\beta + \gamma) + W\text{revenue}(-\rho\beta) + v. \qquad (4)$$

In equation (4) the coefficient of revenue is a biased estimate of β if $\gamma \neq 0$, but a consistent estimate of β is

$$\text{est}(\beta) = -(\text{the coefficient of } W\text{revenue})/\text{est}(\rho) \, . \qquad (5)$$

Table 1 shows that the OLS estimate of β (0.520) is smaller than its MLE from equation 5 (0.886) so the implied estimate of the bias γ is -0.366. A Hausman test (e. g., Greene 2003, 81) indicates that this difference in coefficients is indeed statistically significant.

The food expenditure budget

The data set "VietnamH" (Croissant 2015) is a 1997 survey of expenditures by 5,999 Vietnamese households. An Engel curve can project outlays for food as a function of total expenditures, household size and other factors. Omitted peer effects could again be problematic, but in addition OLS might be biased since total expenditure "and its components…are endogenous to the consumer and are determined simultaneously" (Liviatan 1961, 336). Liviatan argues that OLS will be skewed

downward when the dependent variable is a relatively stable component of expenditure like food while an upward bias should be expected for highly variable items such as major appliances. He proposes an IV regression of each expenditure component on total expenditure with income as the instrument.

Since income is not reported in this data set, I use the linear autoregressive model to estimate the "elasticity" of food expenditure with respect to total expenditure. Peer groups of 50 households are based on the years of schooling and the age of each head of household; again these are presumably exogenous variables with respect to the household's spending decisions. Table 2 shows that the elasticity coefficient is larger for the MLE than for OLS, and a Hausman test confirms that the difference is statistically significant.

Simulation

This section applies the linear autoregressive model to a simulation in which the equation of interest is

$$y = \alpha + \beta x + z \quad . \qquad (6)$$

I set $\alpha = 0$, $\beta = 0.75$, $\rho = 0.6$, and $\gamma = -0.15$. Here x is a uniform random variable ranging from 4 to 18, $v \sim N(0,0.25)$, and $z = \rho Wz + \gamma x + v$. Both grouping variables are N(0,3) and each peer group has 10 members. With 2000 observations the OLS estimate of β is 0.591 and the MLE coefficient of x is 0.602. From equation 5, however, the MLE estimate of β = -(coefficient of Wx)/est(ρ) = 0.481/0.622 = 0.773, which is close to the true value of β; and the implied estimate of the bias γ is 0.602 – 0.773 = -0.171.

Researchers may sometimes find that valid grouping variables are as elusive as valid instrumental variables. In any case simulation does not seem to support a notion that those two roles are interchangeable. If the grouping variables in this simulation are instead used as instruments for equation (6), the estimate of β is 0.180 with a standard error of 0.202. (And if the grouping variables in the Vietnam budget data are instead used as instruments, the estimated elasticity for food is 0.615 –very similar to the biased and inconsistent OLS result.)

Caveats and conclusions

There remain important issues of sensitivity analysis and model comparison. If several grouping criteria are available, which should be used in the model ? And what size should the peer groups be ? Since an exploratory essay cannot address these questions adequately, the reader is referred to Blankmeyer et al. (2011) and LeSage and Pace (2014). However it may be useful to mention two experiments involving the model of the administrator's salary. The grouping variables are beds and staff. If beds

are replaced by the facility's total area (in square feet), the Akaike Information Criterion (AIC) for equation 4 is 231.88 instead of 213.47, a strong indication that beds are preferred to total area. Variation in the size of the peer groups, on the other hand, produces these results:

Group size	15	20	25	30
AIC	224.59	213.47	210.89	211.30

It seems that the model is not very sensitive to group size.

The focus of this paper has been the detection of OLS bias in a "structural" or "behavioral" model, e. g. equation 1. This focus is of course different from another important project, prediction and forecasting. For the latter project a linear regression coefficient estimates the partial derivative of the dependent variable with respect to a regressor. However the situation is more complex for the autoregressive model in equation 4. When an observation on a regressor changes, it induces feedback effects on other observations in the same peer group; and these effects in turn ripple through the sample. In the context of spatial models LeSage and Pace (2009, chapter 2) propose estimates for the indirect and total impacts which "reflect how these changes would work through the simultaneous dependence system over time to culminate in a new steady state equilibrium" (ibid, 37). While impact computations can of course be computed for the examples in this paper, they have been omitted given my focus on structural parameters.

References

Bivand, R., 2019. *Spatial Regression Analysis* (R package '"spatialreg") available at https://cran.r-project.org/.

Blankmeyer, E., J. LeSage, J. Stutzman, K. Knox, R. Pace, 2011. Peer-group dependence in salary benchmarking: a statistical model. *Managerial and Decision Economics* 32: 91-104. DOI: 10.1002/mde.1519.

Bramoulle Y., H. Djebbari, B. Fortin, 2009. Identification of peer effects through social networks. *Journal of Econometrics* 150: 41–55.

Brueckner J., 1998. Testing for strategic interaction among local governments: the case of growth controls. *Journal of Urban Economics* 44: 438–467.

Case, A., H. Rosen, J. Hines, 1993. Budget spillovers and fiscal policy interdependence: evidence from the States. *Journal of Public Economics* 52: 285–307.

Conley T., G. Topa, 2007. Estimating dynamic local interaction models. *Journal of Econometrics* 140: 282–303.

Croissant, Y., 2015. *Ecdat: data sets for econometrics*. Available at https://cran.r-project.org/web/packages/.

Dahl, G., K. Loken, M. Mogstad, 2014. Peer effects in program participation. *American Economic Review* 104: 2049–2074.

De Giorgi, G., M. Pelizzari, S. Redaelli, 2010. Identification of social interactions through partially overlapping peer groups. *American Economic Journal: Applied Economics* 2: 241–275.

Durlauf, S., 1994. Spillovers, stratification and inequality. *European Economic Review* 38: 836–845.

Glaeser, E., B. Sacerdote, J. Scheinkman, 1996. Crime and social interactions. *Quarterly Journal of Economics* 111: 507–548.

Greene, W., 2003. *Econometric Analysis*, fifth edition. Upper Saddle River NJ: Prentice Hall.

Gujarati, D., 2015. *Econometrics by Example,* second edition. NY: Palgrave.

Handcock, M., A. Raftery, J. Tantrum, 2007. Model based clustering for social networks. *Journal of the Royal Statistical Society* A 170(Part 2): 1–22.

LeSage, J.,R. Pace, 2009. *Introduction to Spatial Econometrics*. London: CRC Press.

LeSage, J.,R. Pace, 2014. The biggest myth in spatial econometrics. *Econometrics* 2:217-249. https://doi.org/10.3390/econometrics2040217

Liviatan, N., 1961. Errors in variables and Engel curve analysis. *Econometrica* 29: 336-362.

Min, Chung-Ji. 2019. *Applied Econometrics: A Practical Guide*. NY: Routledge.

Sacerdote, B., 2001. Peer effects with random assignment: results for Dartmouth roomates. *Quarterly Journal of Economics* 116(2): 681–704.

Soetevent A., 2006. Empirics of the identification of social interactions: an evaluation of the approaches and their results. *Journal of Economic Surveys* 20: 193–228.

Texas Health and Human Services Commission, 2002. *2002 Cost Report – Texas Nursing Facility*. Austin, Texas.

von Hinke, S., G. Leckie, C. Nicoletti, 2019. The use of instrumental variables in peer effects models. *Oxford Bulletin of Economics and Statistics* 81: 1179-1191.

Table 1. Nursing facility administrator's salary
(The dependent variable is ln salary, n = 842)

	OLS	MLE
ln revenue	**0.520**	0.677
	0.015	0.021
W(ln revenue)		-0.495
		0.040
rho (ρ)		0.559
		0.067
corrected coefficient for ln revenue = -W(ln revenue)/rho		**0.886**
		0.072

Table 2. Food expenditure elasticity

(the dependent variable is ln food expenditure)
standard errors are shown under coefficients
n = 5999

	OLS	**MLE**
ln total expenditure	**0.659**	0.666
	0.005	0.005
W(ln total expenditure)		-0.212
		0.033
household size	0.043	0.041
	0.002	0.002
gender (male = 1)	0.056	0.057
	0.007	0.007
farm (yes = 1)	0.037	0.035
	0.006	0.006
rho (ρ)		0.266
		0.045
corrected coefficient		**0.797**
for ln total expenditure =		0.046
-W(ln total expenditure)/rho		