Regional Science and Urban Economics 55 (2015) 28–38

Contents lists available at ScienceDirect

Regional Science and Urban Economics

journal homepage: www.elsevier.com/locate/regec

Conditionally parametric quantile regression for spatial data: An analysis of land values in early nineteenth century Chicago

Daniel McMillen

Department of Economics, 214 David Kinley Hall, 1407 W. Gregory, Urbana, IL 61801, United States

ARTICLE INFO

Article history:
Received 6 August 2014
Received in revised form 12 June 2015
Accepted 3 September 2015
Available online 14 September 2015

Keywords:
Quantile
Nonparametric
Spatial econometrics
Land values

ABSTRACT

This paper demonstrates that a conditionally parametric version of a quantile regression estimator is well suited to analyzing spatial data. The conditionally parametric quantile model accounts for local spatial effects by allowing coefficients to vary smoothly over space. The approach is illustrated using a new data set with land values for over 30,000 blocks in Chicago for 1913. Kernel density functions summarize the effects of discrete changes in the explanatory variables. The CPAR quantile results suggest that the distribution of land values shifts markedly to the right for locations near the CBD, close to Lake Michigan, near elevated train lines, and along major streets. The variance of the land value distribution is higher in locations farther from the CBD and farther from the train lines.

1. Introduction

The spatial AR model is the most commonly used alternative to OLS for spatial data analysis. The model adds spatial lags of the dependent variable to the set of explanatory variables, i.e., $Y = \theta WY + X\beta + u$, where W is a "spatial weight matrix" with rows that sum to one and zeros on the diagonals, and θ is a parameter measuring the strength of the relationship. The model can be useful when X does not fully account for the tendency for the dependent variable to be highly correlated over space, so that nearby values of Y provide significant explanatory power. The endogeneity of WY poses challenges for estimation. Most empirical applications are based on maximum likelihood estimation of the model under the assumption of normally distributed errors. Other approaches are based on instrumental variables (IV) estimation, usually with spatially lagged values of X (such as WX and WWX) as instruments for WY (e.g., Kelejian and Robinson, 1993; Kelejian and Prucha, 1999).

Several researchers have used the spatial AR model as the basis for quantile regressions in which both θ and β are allowed to vary across quantiles (e.g., Kostov, 2009; Liao and Wang, 2012; Zeitz et al., 2008; Zhang and Leonard, 2014). The estimation procedures used in these studies follow the IV approaches of either Chernozhukov and Hansen (2006) or Kim and Muller (2004). Both approaches are analogous to IV estimation of the standard spatial AR model, in which spatially lagged values of X serve as instruments for WY. What differs is that the estimated coefficients can vary across quantiles.

Typical specifications of the spatial weight matrix are based on first-order contiguity when the data are drawn from geographic units such as counties or census tracts. Though the approach is used less commonly for point data, typical specifications are similar in that the spatial weights are assumed to decline rapidly with distance between observations. Predicted values are then based on (1) the structural model, $\hat{Y} = \hat{\theta} WY + X\hat{\beta}$; (2) the reduced form, $\hat{Y} = (I - \hat{\theta}W)^{-1}X\hat{\beta}$; or (3) a decomposition into "signal" and "trend" components, $\hat{Y} = \hat{\theta}W(I - \hat{\theta}W)^{-1}X\hat{\beta} + X\hat{\beta}$. Spatial effects generally appear as noise around a spatial trend that looks much like the predicted values from an OLS regression of Y on X. I illustrate this point for a representative data set in McMillen (2012), in which I also argue that a finding of spatial autocorrelation is apt to be an indication of functional form misspecification or other forms of model misspecification that are correlated over space. In the example analyzed in this paper, I find that the signal component dominates the trend in quantile versions of the spatial AR model to such an extent that the predictions appear to be little but noise.

In McMillen (2013), I suggest an alternative to the spatial AR version of the quantile regression model that is analogous to conditionally parametric (CPAR) local linear regression. The estimation procedure involves estimating separate quantile regressions for various target points, with more weight placed on observations that are close to the target. Unlike a fully nonparametric approach, the CPAR approach produces coefficient estimates for the explanatory variables. But unlike the spatial AR version of quantile regression, the estimated coefficients vary over space. The CPAR approach is less sensitive to model misspecification than the fully parametric spatial AR approach, and it accounts for local variation in an overall spatial trend. The approach is

E-mail address: mcmillen@illinois.edu.

http://dx.doi.org/10.1016/j.regsciurbeco.2015.09.001
0166-0462/

well suited for quantile analysis in situations where the distribution of the dependent variable is, for example, highly skewed in some locations, tightly clustered in others, while all the time varying smoothly over space. Moreover, the CPAR approach does not require the specification of a large $(n \times n)$ spatial weight matrix.

In this paper, I illustrate the approach using a new, unique historical data set. The data set includes land values and geographic coordinates for more than 30,000 city blocks in Chicago for 1913. One of the central predictions of urban location theory is that land values should decline smoothly with distance from the central business district (the "CBD"). In contrast to previous studies (e.g., Mills, 1969 and McMillen, 1996), which typically have relatively small sizes, the new data set covers virtually all of Chicago with a very fine level of detail.[1] Though the large sample would seem to be desirable, it presents difficulties for the spatial AR version of the model, which requires the specification of an $(n \times n)$ spatial weight matrix. To simplify the estimation of the spatial AR model, I aggregate the micro data set to averages across city block groups as defined in the 2000 Census.

The CPAR approach produces a richer set of results than the overly restrictive spatial AR quantile model. The CPAR quantile results suggest an interesting pattern in which the distribution of land values shifts markedly to the right for locations near the CBD, close to Lake Michigan, near elevated train lines (the "EL"), and along major streets. The variance of the land value distribution is higher in locations farther from the CBD and farther from the EL.

The approach used in this paper to illustrate how changes in an explanatory variable affect the overall distribution of the dependent variable is not limited to CPAR estimation procedures. Although most researchers present results for a limited set of quantile regressions in a single large table, estimating the regressions at many different quantiles implies an entire distribution of estimated values for the dependent variable. The parametric structure of the quantile model makes it easy to construct counterfactual distributions that shift as the assumed value of an explanatory variable changes. Like any regression, the direction of the change in the counterfactual distributions shows the sign of an explanatory variable's effect on the dependent variable. However, quantile regressions also show the variable's effect on the overall distribution of the dependent variable — e.g., whether it leads to greater changes at high or low values of y, or perhaps leads simply to an increase in the variance of the distribution with no effect on the central tendency. While the counterfactual distribution approach is useful for any quantile regression analysis, it is particularly useful for spatial analysis because it allows a very large set of results to be summarized easily with a series of simple graphs.

Table 1
Descriptive statistics.

	Mean	Std. Dev.	Minimum	Maximum
Full data set (33,477 observations)				
Land value	75.0243	398.6961	0.5739	14,833.33
Log of land value	3.1078	1.3220	−0.5553	9.6046
Distance from CBD	7.1666	3.2832	0.0095	15.0848
Distance from Lake Michigan	3.7073	2.2060	0.0085	8.8107
Distance from EL Line	1.1270	1.0822	0	5.4312
Distance from major street	0.0767	0.0666	0	0.5764
Census block group averages (2206 observations)				
Land value	78.9121	385.1712	0.8609	7518.519
Log of average land value	3.2735	1.2424	−0.1498	8.9251
Distance from CBD	6.7983	3.0847	0.1209	15.5757
Distance from Lake Michigan	3.6714	2.1844	0.0695	8.8208
Distance from EL Line	1.0276	0.9919	0.0012	5.4615
Distance from Major street	0.0986	0.0532	0.0002	0.4409

2. Spatial AR quantile models

Several authors have used IV versions of quantile regression to estimate spatial AR models. Zietz et al. (2008), Liao and Wang (2012), and Zhang and Leonard (2014) use an approach proposed by Kim and Muller (2004) to account for the endogeneity of the spatially dependent variable, WY. The Kim and Muller approach is a simple two-stage procedure in which WY is replaced in the second-stage quantile regression by the predicted values from a first-stage quantile regression of WY on the original explanatory variables, X, and a set of instruments, Z. The same quantile, τ, is used for both regressions. The Kim and Muller is simply a two-stage least squares model with quantile regressions used in place of ordinary least squares (OLS).

Kostov (2009) uses an approach proposed by Chernozhukov and Hanson (2006) to construct an instrumental variable for WY. The Chernozhukov and Hanson approach differs from the Kim–Muller approach in that it does not impose that the same quantile be used in both stages of the procedure. In the version of the approach used here, the predicted values $(\widehat{WY})$ from an OLS regression of WY on X and Z are used as an instrument for WY. In the second stage, this instrumental variable is used as an explanatory variable for a series of quantile regressions of $Y - \theta WY$ on X and $(\widehat{WY})$. The same quantile, τ, is used for each of the regressions, while a grid of alternative values is used for θ. The estimated value of θ is the value that produces the coefficient on $(\widehat{WY})$ that is closest to zero. After finding $\hat{\theta}$, the estimated values of β are calculated by a quantile regression of $Y - \hat{\theta} WY$ on X. The motivation behind this estimator is a property of two-stage least squares: when instruments are chosen optimally, the coefficient on $(\widehat{WY})$ will be zero when both the actual variable, WY, and the instrumental variable are included in a regression.[2]

Parametric spatial econometric models such as the spatial AR model require the researcher to specify a spatial weight matrix that is designed to account for local departures from a broad spatial trend. Misspecification of $X\beta$ can produce statistically significant estimates of θ even if WY would add no explanatory power in a correctly specified model. For example, suppose that house prices are high in certain neighborhoods whose boundaries are not known to the researcher. Omitted or misspecified neighborhood fixed effects will tend to produce a clustering of relatively high house prices, and a weighted average of nearby house prices – of WY – will tend to add significant explanatory power to the house price regression. The spatial AR approach attempts to use a global parametric model to account for local spatial variation that may actually be caused by model misspecification due to spatially correlated missing variables and an incorrect functional form.

3. Conditionally parametric quantile regression

Conditionally parametric models can be a useful alternative to fully parametric models when the true model structure is not known with certainty. The conditionally parametric model is considered in detail by Cleveland et al. (1992) and Cleveland (1994). The idea is to use some structure to reduce the complexity of a fully parametric model. In the case where a parametric structure is appropriate for one set of variables, x, conditional on another set of variables, z, a completely nonparametric model of the form $y_i = f(x_i, z_i) + u_i$ can be replaced by the

[1] Ahlfeldt and Wendland (2011) also analyze a large data set with fine geographic detail. Their data set has approximately 11,000 observations for Berlin during the early nineteenth century.

[2] Multiple instruments can be included directly rather than using the predicted values of WY from a first stage OLS regression. Details are provided in Chernozhukov and Hanson (2006).