Instrumental Variable Methods in Real Estate Data Analysis

Omar Vazquez
Advisor: Professor Ery Arias-Castro

June 5, 2020

Abstract

Although linear models enjoy widespread use in economics, due in part to their simple but flexible form, real estate data generally violates many of their standard assumptions. This paper presents some methods that have allowed applied researchers to conduct more principled analyses with real estate data, while still using linear models. We begin with a review of the motivation behind instrumental variables and the two-stage least squares estimator. Then, we explore some instrumental variable methods and, briefly, an application, under four settings. The first is spatial dependence, which remains relevant within the two topics that follow: quantile regression and simultaneous equation models. Finally, we give a brief overview of instrumental variable estimation for probit regression, where the response is explicitly modeled as binary. This paper may be of interest to those looking to survey how the familiar two-stage least squares estimator has been adapted and generalized to less restrictive settings over the past few decades, with many of the methods discussed remaining in active use.

Contents

1 Instrumental Variables Review

First, we review the use of instrumental variables in the context of linear regression models, with a focus on explaining the motivation behind the use of two-stage least squares estimation. This forms the basis of many econometric methods for causal inference.

1.1 Linear Causal Models

Consider the general model explaining the observed random variable y as a function of P observed predictors $\mathbf{x} = [x_1, \ldots, x_P]'$ and all unobserved random variables u

$$y = g(\mathbf{x}, u)$$

Linear models make the following two assumptions

1. *Additive error:* $g(\mathbf{x}, u) = f(\mathbf{x}) + u$
2. *Linearity:* $f(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$ where $\boldsymbol{\beta}$ is deterministic

This simplifies to

$$y = \mathbf{x}'\boldsymbol{\beta} + u$$

Here β_j is the *causal* effect of a unit increase in x_j on y, holding all else constant. To make this notion more precise, consider a simple experiment to find the causal effect of a drug on an individual's blood pressure. Let the treatment random variable x be binary, so

$$x = \begin{cases} 1 & \text{if the individual receives the drug} \\ 0 & \text{otherwise} \end{cases}$$

and y is the individual's blood pressure. We have the causal model[1] $y = x\beta + u$ with two *potential outcomes* to our experiment, holding all other variables constant:

$$y = \begin{cases} y^{(1)} = \beta + u & \text{if we set } x = 1 \\ y^{(0)} = u & \text{if we set } x = 0 \end{cases}$$

The causal effect of the drug is simply the difference $y^{(1)} - y^{(0)} = \beta$. The issue, of course, is that we can only observe one of the two cases.

Suppose we then decide to run the experiment on a group, with independent and identically distributed (iid) results (x_i, y_i), $i = 1, \ldots, n$, from the model

$$y_i = x_i\beta + u_i$$

Under random treatment assignment, we avoid two basic sources of bias:

[1] Our well-known framework for causation is often attributed to Neyman and Rubin.

1. *Reverse causality*: each x_i will not depend on $y_i^{(0)}, y_i^{(1)}$
2. *Omitted variables*: each x_i will not depend on u_i

and so we can use the average causal effect estimate

$$\hat{\beta} = \frac{1}{n_1}\sum_{x_i=1} y_i - \frac{1}{n_0}\sum_{x_i=0} y_i = \beta + \frac{1}{n_1}\sum_{x_i=1} u_i - \frac{1}{n_0}\sum_{x_i=0} u_i$$

which is unbiased, since

$$E[u_i|x_i = 1] = E[u_i|x_i = 0] \text{ so } E[\hat{\beta}] = \beta$$

It is, in fact, the ordinary least squares estimate, whose assumptions and optimality will be reviewed shortly.

The essential consequence of random assignment is that $E[u_i|x_i]$ is constant for all possible values of x_i. In practice, the researcher may not be able to randomly assign the variables of interest. This is certainly true in real estate data, which will be our focus. Even in observational studies, however, we can often reasonably assume the **exogeneity** condition

$$E[u_i|\mathbf{x}_i] = 0$$

on the linear model

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

where some predictors may be included to control for confounding factors.

1.2 Ordinary Least Squares (OLS)

The standard linear regression model is: given iid data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \text{ or equivalently } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \text{ in matrix form,}$$

where we assume

1. *Exogeneity*: $E[u_i|\mathbf{x}_i] = 0$
2. $\mathbf{X}$ has full column rank so $\boldsymbol{\beta}$ is *uniquely specified*
3. $(\mathbf{x}_i, y_i)$ have finite fourth moments (large outliers are unlikely)

Then the ordinary least squares (OLS) estimate

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

is unbiased, consistent, and asymptotically normal. Furthermore, under

4. *Homoskedasticity*: $Var(u_i|\mathbf{x}_i) = \sigma^2$

the OLS estimator is efficient[2] among the class of linear unbiased estimators. It is also worth noting that OLS corresponds to maximum likelihood estimation (MLE) under normal iid errors.

Since the OLS assumptions are restrictive, they often do not hold in practice, so the estimator $\hat{\boldsymbol{\beta}}_{OLS}$ can often be inconsistent for $\boldsymbol{\beta}$. One problematic assumption in observational studies is exogeneity. The condition that $E[u_i|\mathbf{x}_i] = 0$ implies that all the predictors $\mathbf{x}_i$ are *uncorrelated* with the error term u_i. The latter is sufficient for the consistency of the OLS estimator.[3]

1.3 Endogeneity and Instrumental Variables

When $Cov(x_j, u) \neq 0$, the predictor variable x_j is called **endogenous**. In this case there is an unobserved confounding factor that will generally bias the OLS estimator. We may suspect endogeneity based on domain knowledge, or more formally through, for example, the well-known Durbin-Wu-Hausman test.

We can consider finding a variable z that can only affect the response variable y through the endogenous predictor x_j. Thus, conditional on all other predictors, we require:

1. *Relevance*:[4] $Cov(z, x_j) \neq 0$
2. *Exogeneity*: $Cov(z, u) = 0$

The variable z is called an **instrument**.

The instrumental variable regression model with K endogenous predictors and $P - K$ exogenous predictors is: given iid data $(\mathbf{x}_i, \mathbf{z}_i, y_i)$, $i = 1, \ldots, n$,

$$y_i = \mathbf{x}_i'\boldsymbol{\beta} + u_i, \text{ or equivalently } \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \text{ in matrix form,}$$

[2]Since homoskedasticity is rarely justified, in economics it is more common to use heteroskedasticity-robust standard errors based on White (1980).

[3]This alone does, however, result in a biased estimator.

[4]Weak correlation will result in poor finite-sample performance. See Bound et al. (1995).

Partition $\mathbf{x}_i = \begin{bmatrix} \mathbf{w}_i \\ \mathbf{x}_i^e \end{bmatrix}$ where $\mathbf{w}_i$ contain the exogenous entries of $\mathbf{x}_i$ and $\mathbf{x}_i^e$ contain the endogenous ones. We assume:

1. $E[u_i|\mathbf{w}_i] = 0$
2. $\mathbf{X}$ has full column rank
3. $(\mathbf{x}_i, \mathbf{z}_i, y_i)$ have finite fourth moments
4. The L entries of $\mathbf{z}_i$ are valid instruments

Let the matrix $\mathbf{Z}$ contain the instruments $\mathbf{z}_i$ *and* the exogenous predictors $\mathbf{w}_i$. Denote its projection matrix by

$$\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}$$

To extend the instrument conditions to multiple endogenous predictors, we also assume

4a. $\mathbf{P}_Z\mathbf{X}$ has full column rank

which implies

4b. $L \geq K$: there are not less instruments than endogenous regressors

Here the OLS estimator is generally not consistent due to the endogeneity.

1.3.1 The Two-Stage Least Squares (2SLS) Estimator

The most common estimator used in instrumental variable regression is two-stage least squares:

1. Regress each endogenous variable on $\mathbf{Z}$ by OLS to get the predicted values
$$\hat{\mathbf{X}} = \mathbf{P}_Z\mathbf{X}$$
Note that the exogenous variables are not changed by the projection.
2. Replace $\mathbf{X}$ in the original regression with the exogenous $\hat{\mathbf{X}}$ and estimate $\boldsymbol{\beta}$ by OLS in
$$\mathbf{y} = \hat{\mathbf{X}}\boldsymbol{\beta} + \mathbf{u}$$
This gives the 2SLS estimator, which simplifies to
$$\hat{\boldsymbol{\beta}}_{2SLS} = (\mathbf{X}'\mathbf{P}_Z\mathbf{X})^{-1}\mathbf{X}'\mathbf{P}_Z\mathbf{y}$$

The standard errors are calculated from the closed-form solution, usually to be robust to heteroskedasticity. Although the 2SLS estimator is consistent and asymptotically normal, it is generally biased in finite samples.

The intuitive formulation of the 2SLS procedure is one of its major strengths[5] in applied research, contributing to its prevalence over other, potentially more efficient estimators, such as IV-GMM. It can often be computed much faster than methods that require iterative optimization, as is often the case with MLE, for example. When discussing the applications, we will note that many of the estimation procedures we cover are essentially generalizations of the familiar 2SLS procedure, in the presence of many available alternatives.

1.3.2 Tests for the Instrument Conditions

Testing for relevant instruments given the jth endogenous predictor with observations $\mathbf{x}^{(j)}$ can be accomplished via the first-stage regression

$$\mathbf{x}^{(j)} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\epsilon}_j$$

where if the instruments are relevant then their coefficients $\boldsymbol{\delta}$ should be nonzero. Consequently, in applied research with only one endogenous predictor, the standard F-test for

$$H_0 : \boldsymbol{\delta} = 0 \text{ vs } H_1 : \boldsymbol{\delta} \neq 0$$

is often reported. Rejection is evidence that the instruments are significantly correlated with the endogenous variable. When there are many endogenous variables, the full column rank condition on $\hat{\mathbf{X}} = \mathbf{P}_Z\mathbf{X}$ may be more convenient to verify.

Recall L is the number of instruments and K is the number of endogenous variables. Testing for exogenous instruments can only be done when $L > K$. Consider the regression

$$\hat{\mathbf{u}}_{2SLS} = \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

where $\hat{\mathbf{u}}_{2SLS} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{2SLS}$ are the residuals from the 2SLS estimate. Sargan's J-test statistic is like the F-test statistic for

$$H_0 : \boldsymbol{\gamma} = 0 \text{ vs } H_1 : \boldsymbol{\gamma} \neq 0$$

[5]Note that the 2SLS estimate can have a causal interpretation under milder conditions than linearity. See Angrist and Imbens (1995).

In the GMM framework, it tests the null hypothesis that all instruments and exogenous predictors are indeed exogenous.

1.3.3 The Generalized Method of Moments (GMM)

To simplify our notation here, we denote the rows of $\mathbf{Z}$ as $\mathbf{z}_i$ *for only this section.* The exogeneity of these variables can be summarized as the *vector-valued* moment conditions

$$E[g_i(\boldsymbol{\beta})] = 0$$
$$\text{for } g_i(\boldsymbol{\beta}) = \mathbf{z}_i' u_i = \mathbf{z}_i'(y_i - \mathbf{x}_i'\boldsymbol{\beta}), \quad i = 1 \cdots n$$

Since including additional valid instruments could[6] increase efficiency, we may have more exogenous variables (moment restrictions) than predictors (parameters to optimize). The generalized method of moments procedure seeks to minimize a norm of the sample mean

$$\bar{g}(\hat{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n} g_i(\hat{\boldsymbol{\beta}})$$

of the moments, given by

$$J(\hat{\boldsymbol{\beta}}) = \bar{g}(\hat{\boldsymbol{\beta}})'\mathbf{W}\bar{g}(\hat{\boldsymbol{\beta}})$$

for some positive definite weight matrix $\mathbf{W}$. Setting the gradient of this convex function to zero, the optimum is found to be

$$\hat{\boldsymbol{\beta}}_{GMM} = (\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z}\mathbf{W}\mathbf{Z}'\mathbf{y}$$

which is consistent and asymptotically normal.

Denote $\boldsymbol{\Omega} = \mathbf{Z}'E[\mathbf{u}\mathbf{u}']\mathbf{Z}$. The choice[7] of $\mathbf{W}$ that maximizes the estimator's asymptotic efficiency has been shown to be

$$\mathbf{W} \propto \boldsymbol{\Omega}^{-1}$$

This choice, where the variance estimate can be based on 2SLS residuals, is commonly referred to as IV-GMM. Clearly the choice of $\mathbf{W} = (\mathbf{Z}'\mathbf{Z})^{-1}$ corresponds to the 2SLS estimator, so we see that 2SLS is asymptotically efficient among GMM estimators only under homoskedasticity, i.e. when $E[\mathbf{u}\mathbf{u}'] = \sigma^2\mathbf{I}$.

[6] This may also affect the interpretation of the resulting estimates. In general, the choice of instruments to include requires specific domain knowledge and is often a major focus of research efforts.

[7] Note that any positive scalar multiple of $\mathbf{W}$ gives the same estimator.

2 Extensions for Real Estate Data Analysis

Real estate data has many common features that can result in endogeneity. The first two that we will cover, spatial dependence and quantile effects, are in part consequences of having heterogeneous observations. With these, we can model the endogeneity as an omitted variable problem. Simultaneity, on the other hand, is a consequence of multiple variables being determined at an equilibrium and requires an extension of the single-equation linear model.

These conditions, and some corresponding adaptations of the 2SLS estimator, are defined and discussed in the following sections. Additionally, we discuss instrumental variable estimation when the response variable is explicitly modeled as binary.

2.1 Spatial Dependence

Consider the linear model explaining housing price as a function of some exogenous covariates: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$. It is well-known that the price of a house at one location is positively correlated with the prices of nearby houses. This contradicts the assumption that observed data $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, is iid.

The spatial lag model is a simple *stationary* spatial dependence model, in which $Cov(y_i, y_j)$ is a function of only their distance d_{ij}. The response variable is assumed to follow a spatial autoregressive process[8]

$$\mathbf{y} = \rho\mathbf{W}\mathbf{y} + \mathbf{X}\boldsymbol{\beta} + \mathbf{u} \text{ with } |\rho| < 1$$

where each row is

$$y_i = \rho \sum_{j=1}^{n} w_{ij} y_j + \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

The parameter ρ is the degree of autocorrelation and $\mathbf{W}$ is a matrix of deterministic weights w_{ij} that are decreasing in the distance d_{ij} and zero for non-neighbors. Note that this model is equivalent to

$$\mathbf{y} = (\mathbf{I} - \rho\mathbf{W})^{-1}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u})$$

where $(\mathbf{I} - \rho\mathbf{W})$ should be invertible so that $\boldsymbol{\beta}$ is uniquely specified. There are a few other conditions on $\mathbf{W}$, including that it should be row stochastic,[9] to allow the power series expansion $(\mathbf{I} - \rho\mathbf{W})^{-1} = \mathbf{I} + \rho\mathbf{W} + \rho^2\mathbf{W}^2 + \ldots$

[8] See Su (2012) for estimation of a spatial autoregressive model allowing the exogenous variables to enter nonparametrically.

[9] All its row sums are equal to one.

Thus, we can also write the model as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \rho\mathbf{W}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) + \rho^2\mathbf{W}^2(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) + \dots$$

where the $\rho\mathbf{W}$ term captures the feedback effect of the neighbors, the $\rho^2\mathbf{W}^2$ term captures the effect of the neighbors-of-neighbors, and so on. Some factors that have contributed to its widespread use are its similarity to autoregressive models in the time series context as well as readily available software for estimation, giving economists without specialized knowledge in geospatial data analysis a reasonable model option.

One major complication is that the matrix $\mathbf{W}$ has $\sim n^2$ entries, generally far too many to estimate with n observations. For this reason, the weights are often specified by the researcher as

$$w_{ij} = \begin{cases} 0 & \text{if } i = j \text{ or } d_{ij} \geq c \\ \frac{f(d_{ij})}{\sum_{j=1}^{n} f(d_{ij})} & \text{otherwise} \end{cases}$$

for some threshold $c > 0$ and function f. For example, Liao and Wang (2012) choose $f(d_{ij}) = \exp(-d_{ij})$, over $f(d_{ij}) = 1$, since it allows the estimation results to be less sensitive to the specific choice of c. Another common choice is the inverse distance $f(d_{ij}) = d_{ij}^{-1}$.

Since the spatially-lagged house prices $\mathbf{Wy}$ are endogenous, there are two basic approaches for estimating $\boldsymbol{\beta}$ and ρ. While Ord (1975) outlined a maximum likelihood estimation procedure, we will focus on a popular[10] instrumental variable approach.

2.1.1 A Spatial Two-Stage Least Squares (S2SLS) Estimator

Here we describe the model assumptions and 3-step estimation procedure proposed by Kelejian and Prucha (1998), which allows for spatial autoregressive disturbances[11] *and* response. We will refer to it as the S2SLS estimator, which is sometimes reserved for the case in which only the response variable is autoregressive (this roughly corresponds to only the first of three steps described shortly). The general model is then:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \rho\mathbf{W}\mathbf{y} + \mathbf{u} \text{ with } |\rho| < 1$$

[10] It is used in the Stata spatial regression commands `spregress`, for exogenous predictors, and `spivregress`, for endogenous predictors, with MLE as an option in the former.

[11] This dependence structure is intended to act as a proxy for omitted variables with spatial dependence.

$$\mathbf{u} = \lambda \mathbf{M}\mathbf{u} + \boldsymbol{\varepsilon} \text{ with } |\lambda| < 1$$

where the weight matrices $\mathbf{W}$ and $\mathbf{M}$ are chosen, $\boldsymbol{\varepsilon}$ contains iid[12] innovations with finite fourth moments, and the parameters $\boldsymbol{\beta}$, λ, and ρ will be estimated.

As discussed previously, we require that $(\mathbf{I} - \rho\mathbf{W})$ and $(\mathbf{I} - \lambda\mathbf{M})$ be invertible so that the model uniquely specifies our parameter of interest $\boldsymbol{\beta}$. There are also some standard regularity conditions, omitted here for brevity.

Define the matrix $\mathbf{H} = [\mathbf{X}\ \mathbf{W}\mathbf{y}]$ to contain all the regressors in the full model, with coefficients $\boldsymbol{\delta} = \begin{bmatrix} \boldsymbol{\beta} \\ \rho \end{bmatrix}$. The model is then summarized as:

$$\mathbf{y} = \mathbf{H}\boldsymbol{\delta} + \mathbf{u}$$

with spatial autoregressive disturbances. The instrument matrix[13] $\mathbf{Z}$ for $\mathbf{H}$ is augmented with the columns of $\begin{bmatrix} \mathbf{X} & \mathbf{W}\mathbf{X} & \mathbf{W}^2\mathbf{X} & \dots \end{bmatrix}$ corresponding to exogenous predictors. As (possibly linear combinations of) variables not correlated with the error term, they satisfy the exogeneity condition. The intuition is that these instruments also satisfy the relevance condition with respect to $\mathbf{W}\mathbf{y}$ since researchers only include a variable as a predictor when it is correlated with the response. We now outline the three-step procedure:

1. Using the instrument matrix $\mathbf{Z}$, proceed by 2SLS to obtain an estimate $\hat{\boldsymbol{\delta}}_{2SLS}$. Although this estimator is consistent, despite the autocorrelated disturbances, it is inefficient.

2. Use the estimated residuals $\hat{\mathbf{u}}_{2SLS} = \mathbf{y} - \mathbf{H}\hat{\boldsymbol{\delta}}_{2SLS}$ from the previous step to obtain consistent estimates of λ and the innovation variance. The details of the relatively simple procedure are omitted here.

3. Subtract[14] $\lambda\mathbf{M}\mathbf{y} = \lambda\mathbf{M}\mathbf{H}\boldsymbol{\delta} + \lambda\mathbf{M}\mathbf{u}$ from both sides of the model equation to get:

   $$\mathbf{y} - \lambda\mathbf{M}\mathbf{y} = (\mathbf{H} - \lambda\mathbf{M}\mathbf{H})\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

   where the transformed model is now in terms of the same parameters $\boldsymbol{\delta}$ but with iid errors. Substituting[15] our consistent estimator $\hat{\lambda}$ into the equation above, we can again estimate $\boldsymbol{\delta}$ by 2SLS. This is the S2SLS

[12] This was later extended to allow for heteroskedasticity in Kelejian and Prucha (2010).

[13] See Lee (2003) for the asymptotically optimal choice.

[14] This differencing approach is known as a Cochrane-Orcutt type transformation.

[15] Kelejian and Prucha (1998) show that the asymptotic distribution of the S2SLS estimator is the same as it would be if the true value of λ had been used instead.

estimator and it is consistent and asymptotically normal, as well as more efficient than the estimator from the first step.

When compared to the MLE, S2SLS is less efficient[16] but much simpler computationally.

2.1.2 Application: Measuring the Benefit of Air Quality Improvement

The S2SLS estimator is used by Kim et al. (2003) to study the effect of changes in pollution on housing values in Seoul. Their control variables include housing characteristics, such as number of rooms and time to the nearest school, as well as the relative neighborhood income level. Two pollutants are measured: SO2, which is primarily generated by heating and industrial sources, and NOx, which is primarily generated by transportation.

Variable	OLS	ML	S2SLS	S2SLS robust
ρ		0.469***	0.588***	0.549 ***
		(0.070)	(0.096)	(0.082)
House (binary)	0.127***	0.138***	0.141***	0.123***
	(0.041)	(0.039)	(0.040)	(0.042)
NhdIncome (binary)	0.223***	0.162***	0.147***	0.156***
	(0.042)	(0.040)	(0.042)	(0.041)
HouseFuel (binary)	0.184***	0.186***	0.187***	0.176***
	(0.053)	(0.051)	(0.051)	(0.056)
Bathrooms	0.0774**	0.0813***	0.0823***	0.0704**
	(0.0317)	(0.0304)	(0.0307)	(0.0340)
HouseAge	-0.00462**	-0.00564**	-0.00590***	-0.00658***
	(0.00232)	(0.00223)	(0.00225)	(0.00236)
NearestHospital	-0.00359	-0.00423**	-0.00439**	-0.00417**
	(0.00223)	(0.00214)	(0.00216)	(0.00189)
SO2	-0.0151***	-0.00795**	-0.00612*	-0.00651**
	(0.00300)	(0.00311)	(0.00326)	(0.00303)
NOx	0.00252	0.00137	0.00108	0.00135
	(0.00230)	(0.00223)	(0.00224)	(0.00221)
R^2	0.612	0.644	0.644	0.644

Table 1: Results summarized from Kim et al. (2003)

[16] See Lee (2007) for a GMM-based alternative that may improve upon this.

They choose the spatial lag model, without spatial autoregressive errors, using Lagrange multiplier tests.[17] Four estimation procedures are compared: OLS, assuming no spatial autocorrelation, maximum likelihood (ML), incorporating the spatial lag model and assuming the errors have a normal distribution, and S2SLS under homoskedasticity as well as under heteroskedasticity. Their results for a subset of their predictors are summarized in the table above, where the number of asterisks indicate increasing levels of significance.

Compared to the final S2SLS estimates, using maximum likelihood underestimates the degree of spatial autocorrelation. There are substantial variations in other parameters' magnitudes that Kim et al. (2003) attribute in part to the normality assumption of ML not being appropriate for this dataset.

2.2 Quantile Effects

Recall that, with exogenous predictors, OLS estimates the conditional mean $E[y_i|\mathbf{x}_i] = \mathbf{x}_i'\boldsymbol{\beta}$ by minimizing $\sum_{i=1}^{n}(y_i - \mathbf{x}_i'\mathbf{b})^2$. Another popular regression approach is to estimate the conditional α-quantile

$$Q_\alpha(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}_\alpha, \quad \alpha \in (0,1)$$

by minimizing the weighted sum

$$\sum_{i=1}^{n} \psi(y_i - \mathbf{x}_i'\mathbf{b}), \quad \psi(k) = |k|\left((2\alpha - 1)sign(k) + 1\right)$$

which attains its minimum value when $\alpha * 100\%$ of the residuals are negative. Although there is no general closed-form solution, the estimate can be computed relatively efficiently since the objective function is convex.

It is well-known that estimating the conditional median (i.e. $\alpha = 0.5$, the least absolute deviations estimator) has the advantage of being more robust to outliers in the response than OLS, and general quantile regression shares this feature through its minimization of absolute instead of squared residuals.

Another advantage is under the possibility that quantile effects are present. For example, individuals who buy expensive houses may have different preferences than those who buy cheaper houses. If our coefficients of interest

[17] See Anselin (1988).

$\boldsymbol{\beta}_\alpha$ vary significantly along the conditional distribution of y, then these potentially meaningful trends are "averaged-out" by the conditional mean estimate, which may provide misleading results.[18]

Since 2SLS is also a conditional mean estimator, instrumental variable methods have been developed specifically for quantile regression under endogeneity.[19]

2.2.1 Two-Stage Quantile Regression (2SQR)

Kim and Muller (2004) proposed an intuitive generalization of the 2SLS procedure to quantile regression, commonly referred to as 2SQR (or DSQR). Although it has a few predecessors, 2SQR has excelled by having a form most familiar to applied researchers in economics. Since the model assumptions are analogous to the general instrumental variable regression model, we simply state the procedure.

Given a fixed $\alpha \in (0,1)$ and iid data $(y_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, where the variables $\mathbf{z}_i$ are valid instruments for the subset of the predictors $\mathbf{x}_i$ that are endogenous, to estimate the parameters in the αth conditional quantile:

$$Q_\alpha(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}_\alpha$$

1. For the jth endogenous predictor with observations $\mathbf{x}^{(j)}$, use quantile regression to obtain estimates $\hat{\boldsymbol{\delta}}_{j,\alpha}$ of the parameters in the linear model

   $$Q_\alpha(\mathbf{x}^{(j)}|\mathbf{Z}) = \mathbf{Z}\boldsymbol{\delta}_{j,\alpha}$$

   As before, the matrix $\mathbf{Z}$ contains the observations of *all* of the exogenous variables, including the instruments. Repeat this for every endogenous predictor and construct the estimated conditional quantiles

   $$\hat{\mathbf{X}}_j = \mathbf{Z}\hat{\boldsymbol{\delta}}_{j,\alpha}$$

2. Replace the endogenous variables with their estimated conditional quantiles from stage one (the exogenous predictors are unchanged) and use quantile regression to obtain the 2SQR estimator $\hat{\boldsymbol{\beta}}_{2SQR}$ from

   $$Q_\alpha(\mathbf{y}|\mathbf{X}) = \hat{\mathbf{X}}\boldsymbol{\beta}_\alpha$$

[18] Zietz et al. (2008) investigate this issue and its relation to spatial autocorrelation, both of which when unaccounted for have led to inconsistent findings among many housing studies.

[19] See Chernozhukov and Hansen (2006) for a discussion of their causal interpretation within the potential outcomes framework.

The 2SQR estimator is consistent and asymptotically normal, although its variance estimate is usually obtained by bootstrap. Kim and Muller (2004) also show through both theoretical results and simulations that it is more robust to outliers than 2SLS and offers better performance than some alternatives, such as obtaining the stage-one predicted values from OLS.

2.2.2 Instrumental Variable Quantile Regression (IVQR)

The leading alternative to 2SQR is a GMM-based estimation procedure first attributed to Chernozhukov and Hansen (2006) and commonly referred to as IVQR. As with 2SQR, we omit the familiar model assumptions and simply state the procedure.

Given a fixed $\alpha \in (0,1)$ and iid data $(y_i, \mathbf{x}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, where the variables $\mathbf{z}_i$ are valid instruments for the subset of the predictors $\mathbf{x}_i$ that are endogenous, to estimate the parameters in the αth conditional quantile

$$Q_\alpha(y_i|\mathbf{x}_i) = \mathbf{x}_i'\boldsymbol{\beta}_\alpha$$

first partition the matrix $\mathbf{X} = \begin{bmatrix}\mathbf{X}^e & \mathbf{W}\end{bmatrix}$, where $\mathbf{X}^e$ contains the endogenous predictors and $\mathbf{W}$ contains the exogenous ones. Similarly, partition $\boldsymbol{\beta}_\alpha = \begin{bmatrix}\boldsymbol{\beta}_{\alpha,1} \\ \boldsymbol{\beta}_{\alpha,2}\end{bmatrix}$. As usual, the instrument matrix $\mathbf{Z}$ contains the instruments $\mathbf{z}_i$ *and* the exogenous predictors $\mathbf{W}$. Rewrite the conditional quantile as

$$Q_\alpha(\mathbf{y}|\mathbf{X},\mathbf{Z}) = \mathbf{X}^e\boldsymbol{\beta}_{\alpha,1} + \mathbf{Z}\boldsymbol{\delta}_\alpha$$

and partition $\boldsymbol{\delta}_\alpha = \begin{bmatrix}\boldsymbol{\gamma}_\alpha \\ \boldsymbol{\beta}_{\alpha,2}\end{bmatrix}$, where the coefficients $\boldsymbol{\gamma}_\alpha$ correspond to the instruments $\mathbf{z}_i$. Note that $\boldsymbol{\gamma}_\alpha = 0$ because of the instrument conditions. This motivates the GMM objective function $\|\hat{\boldsymbol{\gamma}}_\alpha\|_{\mathbf{W}} = \hat{\boldsymbol{\gamma}}_\alpha'\mathbf{W}\hat{\boldsymbol{\gamma}}_\alpha$ for a positive definite matrix $\mathbf{W}$. Chernozhukov and Hansen (2006) suggest minimization via grid search:

1. For a grid of values $\{\hat{\boldsymbol{\beta}}_{\alpha,1}^{(j)}, j = 1, \ldots, J\}$, use quantile regression to obtain estimates $\hat{\boldsymbol{\delta}}_\alpha^{(j)}$ of the parameters in the linear model

$$Q_\alpha(\mathbf{y} - \mathbf{X}^e\hat{\boldsymbol{\beta}}_{\alpha,1}^{(j)}|\mathbf{Z}) = \mathbf{Z}\boldsymbol{\delta}_\alpha^{(j)}$$

2. Choose the values $\hat{\boldsymbol{\beta}}_{IVQR} = \begin{bmatrix}\hat{\boldsymbol{\beta}}_{\alpha,1}^{(k)} \\ \hat{\boldsymbol{\beta}}_{\alpha,2}^{(k)}\end{bmatrix}$ where $k = \text{argmin}_j\|\hat{\boldsymbol{\gamma}}_\alpha^{(j)}\|_{\mathbf{W}}$.

The IVQR estimator is consistent and asymptotically normal. In comparing IVQR with 2SQR, Kostov (2009) notes that the former offers better performance in finite-sample inference with weak instruments. This, along with the higher computational demand, makes IVQR more suitable for smaller datasets.

2.2.3 Application: Housing Prices and Spatial Quantile Regression

Liao and Wang (2012) study how the implicit prices of housing characteristics vary across the conditional quantiles of housing prices in an emerging Chinese city. They are interested in households' willingness to pay for proximity to "green space" such as public parks. They estimate the spatial lag model discussed in the previous section by 2SQR, using the spatial lags of the exogenous housing characteristics as instruments for the spatial lag of housing price. Some of their graphical results are presented below.

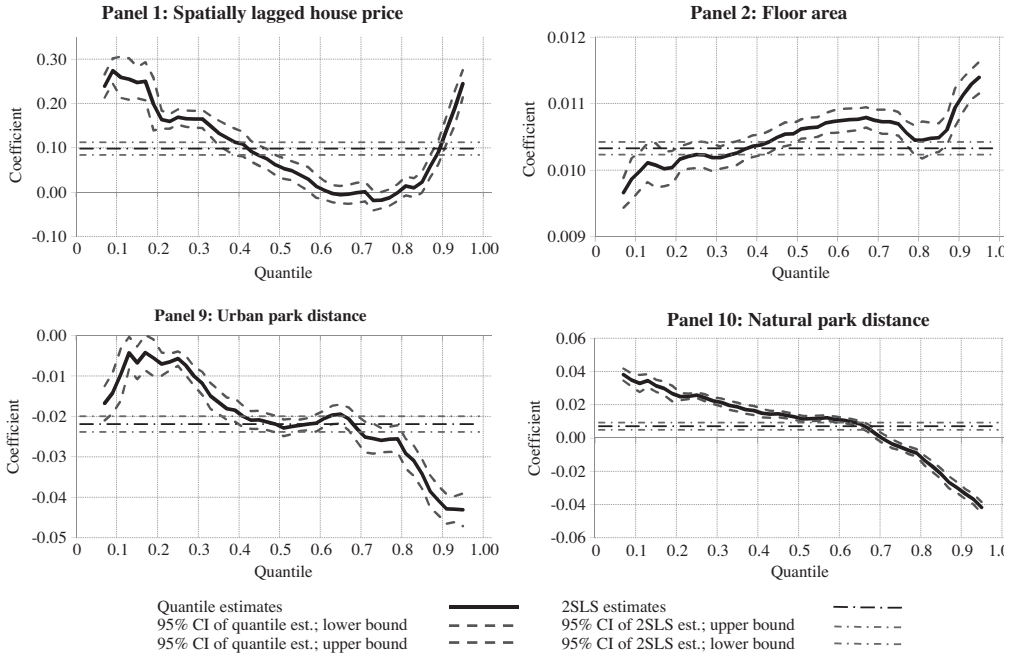

Figure 1: 2SQR and 2SLS coefficient estimates by Liao and Wang (2012)

The coefficient estimates[20] for park distances (panels 9 and 10) support the idea that households with more expensive housing have a higher marginal willingness to pay for environmental amenities, which cannot be captured by the 2SLS estimates. A bit more unexpected is the distinct U-shape in the estimated degree of spatial autocorrelation for different quantiles (panel 1). Liao and Wang (2012) propose some explanations involving local policy decisions and note that it warrants further research.

2.3 Simultaneity

Thus far we have focused on issues related to omitted variables in single-equation models. We now consider a different approach, common in studies concerning supply and demand dynamics, designed to account for feedback simultaneity, which is also known as reverse causation.

Take, for example, the relationship between housing prices and migration. When additional people migrate to a location, this increases the demand for housing in that area, which would generally cause the local housing prices to increase. But when housing prices increase, people have an incentive to move to nearby, more affordable neighborhoods. Thus, our observations of population changes and housing prices are the result of an equilibrium between the two variables. To see why this induces endogeneity, consider the relationship in terms of the two linear models

$$Price = Popchange * \gamma_1 + \mathbf{x}'\beta_1 + u$$

$$Popchange = Price * \gamma_2 + \mathbf{x}'\beta_2 + v$$

where u and v are independent error terms and $\mathbf{x}$ contains exogenous predictors. Substituting the top equation into the bottom one and simplifying, we get

$$Popchange = (1 - \gamma_1\gamma_2)^{-1}(\mathbf{x}'(\beta_1\gamma_2 + \beta_2) + v + u\gamma_2)$$

from which $Cov(Popchange, u) = (1-\gamma_1\gamma_2)^{-1}\gamma_2 \neq 0$ in general, so $Popchange$ is endogenous in the top equation. Thus, we cannot simply use OLS to estimate the causal effect of a change in one variable on the other, but we could, for example, use 2SLS with valid instruments found for each equation individually. From the above expression, we see that the exogenous predictors $\mathbf{x}$ would be natural candidates to satisfy the valid instrument conditions.

[20] The units here are in percentage change of housing price associated with a unit change in the predictor, since the response is log-transformed.

In the more general case, we have the linear simultaneous equations model (SIM) of G equations, where the ith of n iid observations from equation g has the form

$$y_{i,g} = \begin{bmatrix} \mathbf{y}_i \\ \mathbf{x}_{i,g} \end{bmatrix}' \boldsymbol{\beta}_g + u_{i,g}, \quad \mathbf{y}_i = \begin{bmatrix} y_{i,1} & \dots & y_{i,G} \end{bmatrix}'$$

By stacking the G equations for a single observation, we get

$$\mathbf{y}_i = \mathbf{B} \begin{bmatrix} \mathbf{y}_i \\ \mathbf{x}_i \end{bmatrix} + \mathbf{u}_i, \quad \mathbf{B} = \begin{bmatrix} \boldsymbol{\beta}_1 & \dots & \boldsymbol{\beta}_G \end{bmatrix}'$$

where we assume $\mathbf{x}_i$ contains the exogenous predictors across all equations, $\mathbf{y}_i$ contains the endogenous variables, the parameter matrices $\mathbf{B}$ and $\boldsymbol{\Gamma}$ can be determined uniquely from the data, and the errors $u_{i,g}$ are uncorrelated across observations, but possibly correlated across equations.

Although 2SLS can be applied to each of the G equations, using the exogenous predictors $\mathbf{x}_i$ as instruments, a system-wide estimator may be more efficient.

2.3.1 The Three-Stage Least Squares (3SLS) Estimator

We outline the three-stage least squares (3SLS) estimation procedure for simultaneous equation models, first proposed by Zellner and Theil (1962).

To simplify our notation, group the observations by equation as $\mathbf{y}_g = \begin{bmatrix} y_{1,g} & \dots & y_{n,g} \end{bmatrix}'$, letting $\mathbf{X}_g$ contain the corresponding observations of predictors $\{y_{i,s}, s \neq g\}$ and $\mathbf{x}_{i,g}$, and stack them as

$$\begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_G \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1 & 0 & \dots & 0 \\ 0 & \mathbf{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \mathbf{X}_G \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \\ \vdots \\ \boldsymbol{\beta}_G \end{bmatrix} + \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_G \end{bmatrix}$$

with the equivalent matrix equation

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$$

Also, let the instrument matrix be $\mathbf{Z}$, containing the observations of exogenous predictors. Then the three stages are:

1. Project each endogenous variable in $\mathbf{X}$ on $\mathbf{Z}$ to get the predicted values $\hat{\mathbf{X}}$.

2. Use the stage-one $\hat{\mathbf{X}}$ to obtain single-equation 2SLS estimates $\hat{\boldsymbol{\beta}}_{2SLS}$.

3. The 2SLS residuals $\hat{\mathbf{U}} = \begin{bmatrix} \hat{\mathbf{u}}_1 & \dots & \hat{\mathbf{u}}_G \end{bmatrix}$ can be used to calculate a consistent estimate for the error variance via

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\mathbf{U}}'\hat{\mathbf{U}}$$

   Then obtain the 3SLS estimator by generalized least squares[21]

$$\hat{\boldsymbol{\beta}}_{3SLS} = \left[\hat{\mathbf{X}}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_n)\hat{\mathbf{X}}\right]^{-1} \hat{\mathbf{X}}'(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_n)\mathbf{y}$$

Note that, with normal errors, 2SLS is asymptotically equivalent to so-called limited information maximum likelihood estimation, which is efficient among single-equation estimators. Likewise, 3SLS is asymptotically equivalent to full information maximum likelihood estimation. Although the 2SLS estimates are consistent, they are less efficient than 3SLS. This, of course, depends on the consistency of $\hat{\boldsymbol{\Sigma}}$, so *all* the 3SLS estimates may be inconsistent if even a single equation is specified incorrectly,[22] while only that equation's estimates will be impacted when using 2SLS. The error variance estimator may also perform poorly in small samples. For these reasons, 2SLS is often preferred.

2.3.2 A Generalized S2SLS (GS2SLS) Procedure

The GS2SLS and GS3SLS procedures proposed by Kelejian and Prucha (2004) generalize their previously discussed single-equation estimator to obtain consistent estimates of the parameters in spatial autoregressive simultaneous equation models. It allows for spatial dependence in the response variables $y_{i,g}$ as well as in the disturbances $u_{i,g}$, and optionally in some of the exogenous predictors. The model for the gth equation has the familiar form

$$\mathbf{y}_g = \mathbf{X}_g\boldsymbol{\beta}_g + \begin{bmatrix}\mathbf{W}\mathbf{y}_1 & \dots & \mathbf{W}\mathbf{y}_G\end{bmatrix}\boldsymbol{\rho}_g + \mathbf{u}_g, \quad \mathbf{u}_g = \lambda_g\mathbf{W}\mathbf{u}_g + \boldsymbol{\varepsilon}_g$$

[21] Here $\otimes$ denotes the Kronecker product.

[22] As with single-equation models, these instrumental variable estimates can have a causal interpretation under milder conditions than linearity. See Angrist et al. (2000).

where $\mathbf{X}_g$ contains, as rows, the observations of predictors $\{y_{i,s}, s \neq g\}$ and $\mathbf{x}_{i,g}$, and the innovations $\boldsymbol{\varepsilon}$ are generated as

$$\begin{bmatrix} \boldsymbol{\varepsilon}_1 \\ \vdots \\ \boldsymbol{\varepsilon}_G \end{bmatrix} = \boldsymbol{\varepsilon} = (\boldsymbol{\Sigma} \otimes \mathbf{I}_n)\mathbf{v}$$

for an iid mean zero random vector $\mathbf{v}$ with finite fourth moments and non-singular $G \times G$ matrix $\boldsymbol{\Sigma}$. The chosen weight matrices are all identical to simplify notation.

The parameters to be estimated are the degrees of autocorrelation $\boldsymbol{\rho}_g$ and λ_g, the coefficient vectors $\boldsymbol{\beta}_g$, and, in the full-information case, the cross-equation innovation variance matrix $\boldsymbol{\Sigma}$. Note that the vector $\boldsymbol{\rho}_g$ captures spatial dependence of the gth endogenous variable on itself as well as on the other response variables. The full model assumptions are analogous to the single-equation case, so they are omitted here.

The same instrument matrix $\mathbf{Z}$ is used for the endogenous variables $\mathbf{y}_g$ and their spatial lags in each equation. As before, it is augmented with spatial lags of exogenous predictors. Define the matrix $\mathbf{H}_g$ to contain all the regressors in the gth model equation and $\boldsymbol{\delta}_g = \begin{bmatrix} \boldsymbol{\beta}_g \\ \boldsymbol{\rho}_g \end{bmatrix}$ to contain their respective parameters, so

$$\mathbf{y}_g = \mathbf{H}_g \boldsymbol{\delta}_g + \mathbf{u}_g$$

with spatial autoregressive disturbances. The limited information GS2SLS procedure is as follows:

1. Proceed by 2SLS, which is still consistent even though the disturbances are autocorrelated spatially and across equations, and use the instrument matrix $\mathbf{Z}$ to obtain an inefficient estimate $\hat{\boldsymbol{\delta}}_{g,2SLS}$.

2. Use the estimated residuals from the previous step in a GMM procedure to obtain a consistent estimate of the autoregressive disturbance parameter λ_g. The details of the procedure are omitted here.

3. Apply a Cochrane-Orcutt type transformation to obtain the final GS2SLS estimate. As before, we subtract $\lambda_g \mathbf{W}\mathbf{y}_g = \lambda_g \mathbf{W}\mathbf{H}_g \boldsymbol{\delta}_g + \lambda_g \mathbf{W}\mathbf{u}_g$ from both sides of the model equation to get:

$$\mathbf{y}_g^* = \mathbf{y}_g - \lambda_g \mathbf{W}\mathbf{y}_g = (\mathbf{H}_g - \lambda_g \mathbf{W}\mathbf{H}_g)\boldsymbol{\delta}_g + \boldsymbol{\varepsilon}_g = \mathbf{H}_g^* \boldsymbol{\delta}_g + \boldsymbol{\varepsilon}_g$$

where the transformed model is now in terms of the same parameters $\boldsymbol{\delta}_g$ but with iid errors. Substitute the consistent estimate of λ_g into the above equation and then get the final estimate by 2SLS

$$\boldsymbol{\delta}_{g,GS2SLS} = (\mathbf{H}_g^{*\prime}\mathbf{P}_Z\mathbf{H}_g^*)^{-1}\mathbf{H}_g^{*\prime}\mathbf{P}_Z\mathbf{y}_g^*$$

The GS2SLS estimator is consistent and asymptotically normal.

In the full information GS3SLS procedure, there is an additional step:

4. After applying GS2SLS to each of the G equations, use the estimated innovations $\hat{\mathbf{E}} = \begin{bmatrix}\hat{\boldsymbol{\varepsilon}}_1 & \dots & \hat{\boldsymbol{\varepsilon}}_G\end{bmatrix}$ for the consistent estimate

$$\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\hat{\mathbf{E}}'\hat{\mathbf{E}}$$

Stack the differenced model equations $\mathbf{y}_g^* = \mathbf{H}_g^*\boldsymbol{\delta}_g + \boldsymbol{\varepsilon}_g$ as

$$\mathbf{y}^* = \mathbf{H}^*\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

where $\mathbf{H}^* = diag_{g=1}^G(\mathbf{H}_g^*)$. Denoting $\hat{\mathbf{H}}^* = diag_{g=1}^G(\mathbf{P}_Z\mathbf{H}_g^*)$, obtain the 3SLS estimator by generalized least squares

$$\hat{\boldsymbol{\delta}}_{GS3SLS} = \left[\hat{\mathbf{H}}^{*\prime}(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_n)\hat{\mathbf{H}}^*\right]^{-1}\hat{\mathbf{H}}^{*\prime}(\hat{\boldsymbol{\Sigma}}^{-1} \otimes \mathbf{I}_n)\mathbf{y}^*$$

which is efficient relative to GS2SLS but with the same caveat regarding specification errors.

2.3.3 Application: Modeling Population Migration and Housing Price Dynamics

Jeanty et al. (2010) study the relationship between population migration and housing prices in their dataset of Michigan census tracts, using two simultaneous equations. Since migration and housing price changes are likely to have spillover effects onto neighboring areas, they compare the standard OLS and 2SLS estimates with GS2SLS. We summarize their results for the following variables: log median housing value (lnval), 10-year population change (popch), log household average income (lincome) and log population density (lpopden).

The SARLS model allows for spatial autocorrelation but not feedback simultaneity, while the FS-SARLS model allows for both and OLS allows neither. 2SLS allows only feedback simultaneity.

Variable	OLS		2SLS	
	lnval	popch	lnval	popch
lnval		0.2657***		0.0832
popch	0.3538***		0.7671***	
lincome	0.5835***	-0.2116***	0.5203***	- 0.1186***
lpopden	-0.0223***	-0.0731***	0.0109	- 0.0877***

Table 2: Summarized non-spatial results from Jeanty et al. (2010)

Variable	GS2SLS (SARLS)		GS2SLS (FS-SARLS)	
	lnval	popch	lnval	popch
lnval		0.2647***		-0.1189**
popch	0.3312***		0.1607**	
Spatial Lags				
wlnval	0.2830***		0.3220***	
wpopch		0.0627		0.7270***
lincome	0.4837***	-0.1567***	0.5014***	0.0202
lpopden	-0.0054	-0.0705***	-0.0186**	-0.0766***

Table 3: Summarized spatial results from Jeanty et al. (2010)

We see substantial variation in the estimates' magnitude, sign, and significance across the four specifications. Recall that we noted increases in housing prices tend to incentivize people to migrate to other neighborhoods. This corresponds to a negative coefficient of lnval in the equation for popch, which *only* appears in the FS-SARLS model estimate. This is evidence that failing to account for the feedback simultaneity and cross-sectional spatial dependence produces biased results.

2.4 Binary Response

We now move beyond standard linear models into the classification setting, which explicitly incorporates the fact that the response variable y is binary. Given iid data $(\mathbf{x}_i, y_i)$, $i = 1, \dots n$, the linear classification model is

$$y_i = \mathbb{1}(y_i^* > 0), \quad y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$

where $E[y_i|\mathbf{x}_i] = P(y_i = 1|\mathbf{x}_i) = F(\mathbf{x}_i'\boldsymbol{\beta})$ for some link function F mapping the real number line to the interval $(0, 1)$. The probit[23] model uses $F = \Phi$, the standard normal cumulative distribution function. Then the conditional error distribution is $u_i \sim N(0, 1)$. This follows from the simple calculation

$$P(u_i \leq \mathbf{x}_i'\boldsymbol{\beta}) = P(y_i = 0|-\mathbf{x}_i) = 1-P(y_i = 1|-\mathbf{x}_i) = 1-\Phi(-\mathbf{x}_i'\boldsymbol{\beta}) = \Phi(\mathbf{x}_i'\boldsymbol{\beta})$$

Setting the conditional error variance to one guarantees that $\boldsymbol{\beta}$ is uniquely specified, since multiplying $\mathbf{x}_i'\boldsymbol{\beta}+u_i$ by any positive constant does not change the value of y_i. For exogenous predictors, the most used estimator is maximum likelihood, using the specified Bernoulli distribution

$$\hat{\boldsymbol{\beta}}_{ML} = \text{argmax}_{\mathbf{b}} \sum_{i=1}^{n} [y_i \log \Phi(\mathbf{x}_i'\mathbf{b}) + (1 - y_i) \log(1 - \Phi(\mathbf{x}_i'\mathbf{b}))]$$

Note that the causal effects are

$$\frac{\partial P(y = 1|\mathbf{x})}{\partial x_j} = \Phi'(\mathbf{x}'\boldsymbol{\beta})\beta_j$$

whereas in standard linear models they are

$$\frac{\partial E[y|\mathbf{x}]}{\partial x_j} = \beta_j$$

In practice, the estimates from using a probit model may not differ much from OLS. Nevertheless, in the structural approach to causal inference, where we assume a correctly specified model and our objective is to estimate its parameters, OLS is fundamentally inappropriate.[24] In general, researchers use the probit model, or another link function, when there is a binary response.

2.4.1 IV Probit Estimation

Many instrumental variable estimators have been developed to deal with endogenous predictors in the probit model. We describe an approach that

[23] Which is often favored over logit models in economics, due in part to its similarity to tobit models for censored data.

[24] Angrist (2001) argues, however, that the case of a binary response is not fundamentally different from the continuous case and so OLS and 2SLS estimates may still have a causal interpretation

is analogous to the exogenous probit maximum likelihood. Given iid data $(\mathbf{x}_i, \mathbf{z}_i, y_i)$, $i = 1, \ldots, n$, where we partition $\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_i^e \\ \mathbf{w}_i \end{bmatrix}$ so that $\mathbf{w}_i$ contains the exogenous predictors and $\mathbf{x}_i^e$ contains the endogenous ones, we have the model

$$y_i = \mathbb{1}(y_i^* > 0), \quad y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + u_i$$
$$\mathbf{x}_i^e = \boldsymbol{\Pi}_1 \mathbf{w}_i + \boldsymbol{\Pi}_2 \mathbf{z}_i + \mathbf{v}_i$$

where the second equation is simply the relevance condition for the valid instruments $\mathbf{z}_i$. Assume, conditional on $\mathbf{w}_i$ and $\mathbf{x}_i^e$, iid errors $[u_i \ \mathbf{v}_i']' \sim N(0, \boldsymbol{\Sigma})$ with, as before, the conditional variance of u_i set to one.

We can factor the joint density function into

$$f(y_i, \mathbf{x}_i^e | \mathbf{w}_i, \mathbf{z}_i; \boldsymbol{\beta}, \boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2, \boldsymbol{\Sigma})$$
$$= g(y_i | \mathbf{x}_i, \mathbf{z}_i; \boldsymbol{\beta}, \boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2, \boldsymbol{\Sigma}) h(\mathbf{x}_i^e | \mathbf{w}_i, \mathbf{z}_i; \boldsymbol{\Pi}_1, \boldsymbol{\Pi}_2, \boldsymbol{\Sigma}_v)$$

where g is a probit likelihood and h is a normal likelihood. We can then maximize the log-likelihood to estimate all parameters *jointly*. This seems to be the most used estimator, by far.[25]

A popular two-stage IV probit estimator is an extension[26] of the two-stage conditional maximum likelihood (2SCML) procedure of Rivers and Vuong (1988). The latter proposed:

1. First maximize the likelihood h by OLS regression of $\mathbf{x}_i^e$ on $\mathbf{w}_i$ and $\mathbf{z}_i$ to estimate $\boldsymbol{\Pi}_1$ and $\boldsymbol{\Pi}_2$. Use the estimated residuals in the variance estimate
$$\hat{\boldsymbol{\Sigma}}_v = \frac{1}{n} \sum_{i=1}^{n} \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i'$$

2. Using the stage-one estimates, maximize the likelihood g over the remaining parameters by probit maximum likelihood

Strictly speaking, these estimators assume that the endogenous predictors are continuous, since otherwise the model would be a system of probit equations.

[25] It is the default in the Stata command `ivprobit`, for example, with the option of using Newey's two-step estimator instead.

[26] The minimum chi-squared estimator of Newey (1987).

2.4.2 Application: The Effect of Housing Wealth on Labor Force Participation

Fu et al. (2016) study the effect that a change in housing wealth has on the choice to participate in the labor force in urban China. Since they have strong reasons to suspect that an individual's housing wealth is correlated with some influential unobserved variables, such as housing and income preferences, they treat it as endogenous. Their instrumental variable is the average change in housing value for other houses in the same neighborhood since it should not directly affect the individual's decision to work. Their estimation results, for treating labor force participation as continuous (OLS and 2SLS) as well as under the probit model (probit and IV probit) are summarized in the table below.

Effect of housing wealth change on labor force participation, IV estimation.

Variables	(1) Full sample	(2) Female sample	(3) Male sample	(4) Full sample	(5) Female sample	(6) Male sample
Panel 1: linear probability model results						
	OLS			2SLS		
HousingWealthChange	−0.0011	−0.0013	−0.0002	−0.0038	−0.0137**	0.0055
	(0.0010)	(0.0017)	(0.0017)	(0.0050)	(0.0056)	(0.0069)
Sample size	4332	1896	2436	4332	1896	2436
R^2 (Centered R^2)	0.1820	0.2139	0.1953	0.1642	0.1734	0.1556
First-stage regression						
Instrumental variable for				0.7500***	0.7517***	0.7513***
HousingWealthChange				(0.1080)	(0.1403)	(0.0924)
First stage F test				48.19	28.69	66.07
Panel 2: probit results						
	Probit			IV Probit		
HousingWealthChange	−0.0011	−0.0018	−0.0005	−0.0034	−0.0143**	0.0025
	(0.0010)	(0.0019)	(0.0014)	(0.0043)	(0.0057)	(0.0053)
Sample size	4317	1889	2356	4317	1889	2356
Pseudo R^2	0.2472	0.2470	0.2694			

Figure 2: Results summary from Fu et al. (2016)

Note that, both with and without instruments, using the probit model produces roughly the same[27] estimate signs and significance as treating the response as continuous. There are, however, substantial differences between estimates with and without accounting for endogeneity. The IV probit results suggest that a gain in housing wealth significantly decreases labor force participation for women, while the standard probit results show no significant impact.

[27] Although they are estimates of different quantities.

References

Angrist, J. D. (2001). Estimation of limited dependent variable models with dummy endogenous regressors: simple strategies for empirical practice. *Journal of Business & Economic Statistics*, 19(1):2–28.

Angrist, J. D., Graddy, K., and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442.

Anselin, L. (1988). Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, 20(1):1–17.

Bound, J., Jaeger, D. A., and Baker, R. M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

Chernozhukov, V. and Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. *Journal of Econometrics*, 132(2):491–525.

Fu, S., Liao, Y., and Zhang, J. (2016). The effect of housing wealth on labor force participation: evidence from China. *Journal of Housing Economics*, 33:59–69.

Jeanty, P. W., Partridge, M., and Irwin, E. (2010). Estimation of a spatial simultaneous equation model of population migration and housing price dynamics. *Regional Science and Urban Economics*, 40(5):343–352.

Kelejian, H. H. and Prucha, I. R. (1998). A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *The Journal of Real Estate Finance and Economics*, 17(1):99–121.

Kelejian, H. H. and Prucha, I. R. (2004). Estimation of simultaneous systems of spatially interrelated cross sectional equations. *Journal of Econometrics*, 118(1-2):27–50.

Kelejian, H. H. and Prucha, I. R. (2010). Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics*, 157(1):53–67.

Kim, C. W., Phipps, T. T., and Anselin, L. (2003). Measuring the benefits of air quality improvement: a spatial hedonic approach. *Journal of Environmental Economics and Management*, 45(1):24–39.

Kim, T.-H. and Muller, C. (2004). Two-stage quantile regression when the first stage is based on quantile regression. *The Econometrics Journal*, 7(1):218–231.

Kostov, P. (2009). A spatial quantile regression hedonic model of agricultural land prices. *Spatial Economic Analysis*, 4(1):53–72.

Lee, L.-F. (2003). Best spatial two-stage least squares estimators for a spatial autoregressive model with autoregressive disturbances. *Econometric Reviews*, 22(4):307–335.

Lee, L.-F. (2007). GMM and 2SLS estimation of mixed regressive, spatial autoregressive models. *Journal of Econometrics*, 137(2):489–514.

Liao, W.-C. and Wang, X. (2012). Hedonic house prices and spatial quantile regression. *Journal of Housing Economics*, 21(1):16–27.

Newey, W. K. (1987). Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics*, 36(3):231–250.

Ord, K. (1975). Estimation methods for models of spatial interaction. *Journal of the American Statistical Association*, 70(349):120–126.

Rivers, D. and Vuong, Q. H. (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics*, 39(3):347–366.

Su, L. (2012). Semiparametric GMM estimation of spatial autoregressive models. *Journal of Econometrics*, 167(2):543–560.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838.

Zellner, A. and Theil, H. (1962). Three-stage least squares: simultaneous estimation of simultaneous equations. *Econometrica*, 30(1):54–78.

Zietz, J., Zietz, E. N., and Sirmans, G. S. (2008). Determinants of house prices: a quantile regression approach. *The Journal of Real Estate Finance and Economics*, 37(4):317–333.