University of Illinois
Fall 2013

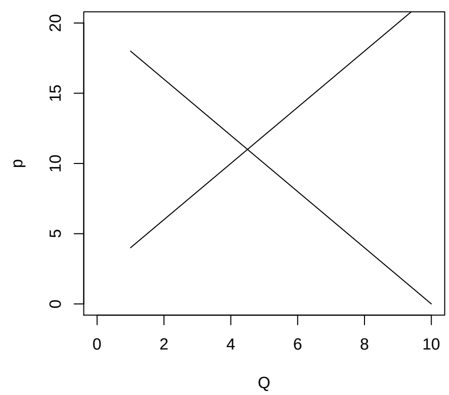
Department of Economics
Roger Koenker

Economics 508
Lecture 12

Introduction to Dynamic Simultaneous Equation Models

1. Structural Equation Models – An Overview

The classical problem in economics, and therefore in econometrics, is iconified in the so called supply and demand diagram.



Suppose we consider a commodity like oil, both demand and supply *functions* are subject to various shocks that shift the two curves around. What we see –at best– is a set of intersections. In the absence of any further information the problem of uncovering separate S and D curves is hopeless. This point was made in the classic (1928) paper of E. J. Working – "What do Statistical Demand Curves Show?"

Now suppose we consider, again classically, the demand for an agricultural commodity like wheat. Henry Schultz (1938) argued that for most of these commodities we could consider demand to be quite stable and supply to be shifting with weather condition, etc., so the locus of S-D intersections traced out (roughly) a demand curve. Likewise if supply were stable and demand very variable, we would be able to "see" supply.

Now consider something a little more modern, à la Heckman (2000),

$$\begin{cases} Q^D = Q^D(P^D, Z^D \\ Q^S = Q^S(P^S, Z^S) \\ Q^D = Q^S, P^D = P^S \end{cases}$$

Suppose we know, or at least believe, that Z_i^D influences Q^D, but not Q^S, we can write with impunity,

$$\frac{\partial Q^S}{\partial Z_i^D} = \frac{\partial Q^S}{\partial P^S}\frac{\partial P^S}{\partial Z_i^D}$$

and using the equilibrium conditions and assuming that the effect of Z_i^D on P is really non-negligible we have

$$\frac{\partial Q^S}{\partial P^S} = \frac{\partial Q/\partial Z_i^D}{\partial P/\partial Z_i^D}$$

In the simplest case where we assume these effects are constant (linearity of the reduced form model) this could be estimated by taking the ratio of two reduced form coefficient estimates:

$$\begin{aligned}
Q_i &= \alpha + \beta Z_i + u \\
P_i &= \gamma + \delta Z_i + v \\
\hat{\beta} &= \frac{\sum (Z_i - \bar{Z})(Q_i - \bar{Q})}{\sum (Z_i - \bar{Z})^2} \quad \text{estimates } \frac{\partial Q}{\partial Z_i} \\
\hat{\delta} &= \frac{\sum (Z_i - \bar{Z})(P_i - \bar{P})}{\sum (Z_i - \bar{Z})^2} \quad \text{estimates } \frac{\partial P}{\partial Z_i}
\end{aligned}$$

so

$$\hat{\eta} = \hat{\beta}/\hat{\delta} \qquad \text{estimates } \frac{\partial Q^S}{\partial P^S}.$$

This is the IV estimator,

$$\hat{\eta} = \frac{\sum \tilde{Z}_i \tilde{Q}_i}{\sum \tilde{Z}_i \tilde{P}_i}$$

where $\sim$'s denote variables in deviations from sample means. Likewise if we have Z_i^S we can recover $\partial Q^D/\partial P^D$. This can be greatly elaborated even under the restrictive linearity assumption and leads to the classical Cowles Commission theory of identifications in SEM's.

2. Causal Models and the DAG

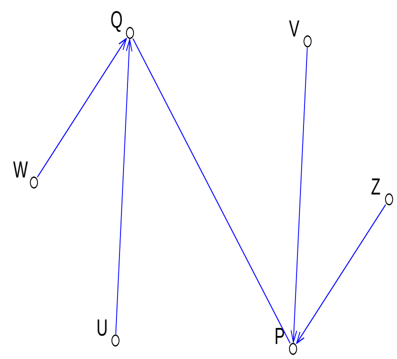
A basic question – one that children like to ask *all* the time is *why*?[1] To properly answer this question we need more than a theory of statistical correlation, or association. This leads us into the depths of philosophy, but it also forces us to think more clearly about problems and to try to avoid the naïve *post hoc ergo propter hoc* fallacy. (This before that, therefore that because of this).

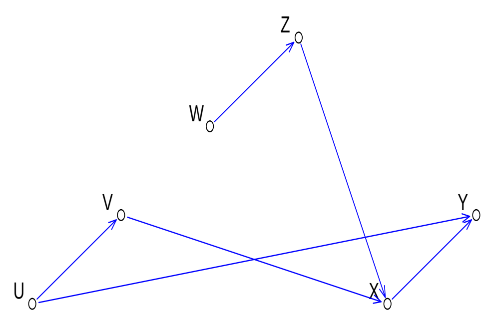
A tool for better understanding the structure of causal models is the path diagram introduced by the geneticist Sewell Wright in 1921. The modern version of this approach is the DAG championed by Judea Pearl.

Def: A *graph* is a set, V, of *nodes* and a set E of *edges*, or *links*. Edges can be directed, marked by a directed arrow, or undirected. If all edges are directed, then the graph is directed. A path is any sequence of edges, a directed path is any path that travels in accordance with the *directed edges*. Directed graphs may contain cycles, i.e., directed paths that returns to their originating node. Directed graphs that do not contain cycles are *acyclic*, i.e., *DAG*'s. DAG's can be used to represent causal models.

Note that the S-D example is not a DAG since it has the cycle $Q \leftrightarrow P$.

[1] See C. Tilly, 2006, Why? Princeton University Press.





Another crucial example is estimation of treatment effects with imperfect compliance, Pearl (Test, Fg. 5)

Assume Z, X, Y are observed, W, V, U are not, Z is randomized treatment assignment, X is actual treatment, Y is response. Dependence between V and U indicates that compliance may be influenced by some of the same factors that determine the response Y.

The objective is to estimate the effect of treatment on response abstracting from, that is to say adjusting for, non-compliance. In Pearl's notation this is

$$P(y|do(X = x))$$

Note that this can be quite different from the familiar conditional distribution function $P(y|x)$.

The classical example of this type in economics is the Roy model of occupational choice revived by Heckman and Honorè (1990). Suppose

$$Y_i = D_i Y_{1i} + (1 - D_i) Y_{0i}$$

where

$Y_{1i} \sim$ earnings of i if Hunter
$Y_{0i} \sim$ earnings of i if Fisherman

We would like to estimate the effect of X on earnings

$$\Delta^{ATE}(x) = E(\Delta = Y_1 - Y_0 | X = x)$$

but this average treatment effect is too difficult, so we try instead to estimate:

$$\Delta^{\pi}(x) = E(\Delta | X = x, D = 1)$$

How? Suppose we have covariates Z that effectively predict the "propensity score"

$$P(z) = P(D = 1 | Z = z)$$

Then we can define the local ATE

$$\begin{aligned}
&\Delta^{LATE}(x, P(z), P(z + \Delta_z)) \\
&\qquad = \frac{E(Y|X = x, P(Z) = P(z)) - E(Y|X = x, P(Z) = P(z + \Delta_z))}{P(z) - P(z + \Delta_z)} \\
&\qquad \to \frac{\partial E(Y|X = x, P(Z) = P(z))}{\partial P(z)}.
\end{aligned}$$

This also has a nice IV interpretation.

3. Classical Dynamic Simultaneous Equation Models

In this part of the lecture we will introduce some simple dynamic simultaneous equation models. Problem Set 3 will deal with two classical examples of this class of models which are typically used in studying partial equilibrium models of a single market. Let's begin by considering a rather general class of models of this form,

$$\Gamma y_t = A(L) y_{t-1} + B(L) x_t + u_t \tag{1}$$

where we will refer to y_t as a m-vector of endogenous variables, x_t as a q-vector of exogenous variables, and $A(L)$ and $B(L)$ denote matrix polynomials in the lag operator L, as usual. We will assume that given vectors y_{t-1}, and x_t, a realization of the vector u_t determines a unique realization of the response vector y_t, given the exogenous variables x_t and the past. This uniqueness requires that the matrix Γ be *invertible*. We may then "solve" the structural form of the model, (1), to obtain the reduced form,

$$y_t = \Psi(L) y_{t-1} + \Pi(L) x_t + v_t \tag{2}$$

where $\Psi(L) = \Gamma^{-1} A(L), \Pi(L) = \Gamma^{-1} B(L)$ and $v_t = \Gamma^{-1} u_t$. We can think of the structural form as representing an idealized version of how the model "really works", while the reduced form is a cruder version of the model which could be used for forecasting, for example.

It is helpful at this stage to have a more concrete example so let's consider the cobweb model from part one of PS 3.

$$\begin{aligned}
Q_t &= \alpha_1 + \alpha_2 P_{t-1} + \alpha_3 z_t + u_t \\
P_t &= \beta_1 + \beta_2 Q_t + \beta_3 w_t + v_t
\end{aligned} \tag{3}$$

To connect this model with the notation of (1) we may write,

$$\begin{aligned} x_t &= \begin{pmatrix} 1 \\ z_t \\ w_t \end{pmatrix} \qquad \text{exogenous variables} \\ y_t &= \begin{pmatrix} Q_t \\ P_t \end{pmatrix} \qquad \text{endogenous variables} \end{aligned}$$

$$\Gamma = \begin{pmatrix} 1 & 0 \\ -\beta_2 & 1 \end{pmatrix}, \quad A(L) = \begin{pmatrix} 0 & \alpha_2 \\ 0 & 0 \end{pmatrix} \quad \Gamma^{-1} = \begin{pmatrix} 1 & 0 \\ \beta_2 & 1 \end{pmatrix}, \quad \Psi(L) = \begin{pmatrix} 0 & \alpha_2 \\ 0 & \alpha_2\beta_2 \end{pmatrix}$$

$$B(L) = \begin{pmatrix} \alpha_1 & \alpha_3 & 0 \\ \beta_1 & 0 & \beta_3 \end{pmatrix}, \quad \Pi(L) = \begin{pmatrix} \alpha_1 & \alpha_3 & 0 \\ \alpha_1\beta_2 + \beta_1 & \beta_2\beta_3 & \beta_3 \end{pmatrix}$$

Now consider solving the model for an equilibrium form. This would yield, in the general case

$$y_e = (I - \Psi(1))^{-1}\Pi(1)x_e \tag{4}$$

In the specific model (3) form, we would have

$$I - \Psi(1) = \begin{pmatrix} 1 & -\alpha_2 \\ 0 & 1 - \alpha_2\beta_2 \end{pmatrix}$$

so

$$(I - \Psi(1))^{-1} = \begin{pmatrix} 1 - \alpha_2\beta_2 & \alpha_2 \\ 0 & 1 \end{pmatrix} / (1 - \alpha_2\beta_2)$$

In this case we must check to be sure that

$$(I - \Psi)^{-1} = I + \Psi + \Psi^2 + \cdots$$

converges in order to assure stability.

Having derived the equilibrium form of the model it is now an opportune moment to discuss forecasting. This is again conveniently done using the reduced form. The one-step ahead forecast in the simple case where $\Psi(L)$ and $\Pi(L)$ are matrices independent of L, yields

$$\hat{y}_{n+1} = \Psi y_n + \Pi x_{n+1}$$

and thus two-steps ahead,

$$\begin{aligned} \hat{y}_{n+2} &= \Psi(\Psi y_n + \Pi x_{n+1}) + \Pi x_{n+2} \\ &= \Psi^2 y_n + \Psi\Pi x_{n+1} + \Pi x_{n+2} \end{aligned}$$

and thus for s steps ahead,

$$\hat{y}_{n+s} = \Psi^s y_n + \sum_{i=0}^{s-1} \Psi^i \Pi x_{n+s-i}$$

Note that, although it might not be immediately obvious, this is completely consistent with the previous discussion of equilibrium. Suppose $x_{n+i} = x_e$ for all $i \geq 0$, then presuming that $\Psi^s \to 0$ as is obviously required for stability, we have

$$\hat{y}_{n+s} \to (\sum_{i=0}^{\infty} \Psi^i)\Pi x_e = (I - \Psi)^{-1}\Pi x_e$$

as previously discussed.

4. Estimation of Dynamic Simultaneous Equation Models

Estimation of simultaneous equation models pose some new problems which we will now gradually introduce. It is convenient to begin with the simple case of recursive models which bring us just to the edge of simultaneity, without quite plunging into it.

Definition Model (1) is recursive if the following conditions hold, or if the endogenous variables can be reordered so that these conditions are met: (i) The matrix Γ is lower triangular, and (ii) The covariance matrix $\Omega = Eu_t u_t'$ is diagonal, and $\{u_t\}$ is iid over time.

We hasten to point out that condition (i) is satisfied by the cobweb model (3), so provided we assume, in addition, in that model that $Eu_t v_t = 0$ so that it satisfies (ii),we can say that it is in recursive form. What is so special above recursive models? Closer examination reveals that recursivity implies an explicit *causal ordering* in the model. In our cobweb model (3) this is illustrated in the following diagram.

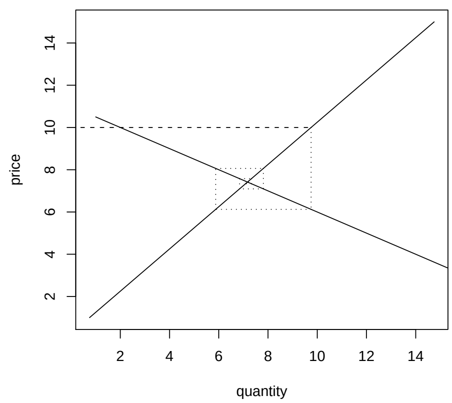


Figure 1. A simple linear cobweb model of supply and demand

Given an initial price p_0, supply determines the next periods quantity, Q_1. Demand in period one then determines a price that will clear the market and this, then, triggers a supply response for period two and the model continues to operate. In the figure we see that for fixed supply and demand functions (recall that in the constantly *model* as specified in (3), these functions are shifting around due both to changes in the exogenous variables x_t and w_t as well the random effects u_t and v_t), this leads to convergence to an equilibrium (P_t, Q_t) pair that looks a little like a spider web as it oscillates around the equilibrium point. At this point you might explore what happens if the demand curve is quite steep, i.e., inelastic, and supply is quite flat, i.e., elastic. Explain how this connects with the algebraic formulation of the model and the stability of equilibrium.

The crucial observation about causal ordering and estimation of recursive models concerns the orthogonality of the errors and explanatory variables in models of this form. Note that in model (3) p_{t-1} is clearly orthogonal to u_t provided the $\{u_t\}$ are iid, and in the demand equation Q_t is orthogonal to v_t provided that v_t and u_t are orthogonal. Clearly, Q_t depends upon u_t, so if u_t and v_t were correlated, then this orthogonality (i.e., $EQ_t v_t = 0$) would fail and ordinary least squares estimation would be biased. Obviously, the conditions for recursivity are rather delicate. We need, not only that the matrix Γ be triangular, a condition which itself is rather questionable, but we also rely on strong assumptions about the unobservable error effects in the system of equations. An illustration of how delicate things really are is provided by the effect of autocorrelation in u_t in the first equation of model (3). Note that if we perturb the model slightly and consider the possibility that

$$u_t = \rho u_{t-1} + \varepsilon_t$$

with $\{\varepsilon_t\}$ iid, now u_{t-1} clearly influences u_t, but it also helps to determine Q_{t-1}, thus p_{t-1}, so $Eu_t p_{t-1} = 0$ fails in this case and we need a more general estimation strategy that is capable of dealing with these failures of orthogonality. We will introduce a test for this kind of effect in the next lecture.

It is worth taking a few moments to try to clarify how this model differs from those models discussed earlier under the rubric "seemingly unrelated regressions." It may appear at first sight that a typical SUR application, like estimating a system of demand equations, is also a case of "simultaneous equations." After all, isn't the consumer simultaneously choosing his optimal consumption bundle based on the prices, income and the other factors he encounters? Why have we insisted that in this case we estimate a *reduced form* model of the SUR type while in the case of our supply and demand model we have posited a *structural model* which explicitly reflects the dependence of endogenous variable on one another?

A crucial distinction that is useful in unraveling this mystery involves asking the simple question: Is there a plausible thought experiment that would render the structural model meaningful. In the supply and demand model the answer is clearly "yes" – we can imagine putting various quantities "on the market" and observing the price that cleared the market. This thought experiment essentially reveals the demand equation. Here supply and demand processes are autonomous. In contrast, for the demand system example we would have to imagine that one was asked to optimize subject to fixed quantities of some subvector of commodities and this stretches credulity considerably. In the former case it may be quite important to know how demand and supply respond separately; whereas in the demand system example one agent in making a unified choice and there is no sense in compartmentalizing it. In both cases it is sensible to speak of endogenous variables, that is variables determined within the model, and of exogenous variables determined outside the model. However, in order that a meaningful structural interpretation of the model be possible we need to have some sort of equilibrium interpretation in which the structural equations reflect an autonomous behavioral relationship. This discussion leads naturally into the fundamental problem of identification in structural models. When can we infer knowledge of structural parameters from knowledge of the reduced form parameters? In keeping with the pragmatic, applied nature of the course we will focus exclusively on the elementary rule: we must have at least as many valid instrumental variables as included endogenous variables. This theme will be elaborated as we develop the simultaneous equation model.

5. Instrumental Variables and Two Stage Least Squares

If correlation (read: lack of orthogonality) between X and u is a potential problem lurking in the shadows of recursive models, it is fully armed and dangerous in the classical simultaneous equations model. For general Γ in (1), the model is truly simultaneous in that the realization of the error vector u_t jointly determines all of the endogenous variables and therefore, any contemporaneous endogenous variable appearing on the right hand side of an equation is bound to be correlated with the error term in that equation.

The remedy for this, fortunately, is rather simple. We must find, as we have already seen in discussion of estimating models with lagged dependent variables and autocorrelated errors, instrumental variables, z_i, that satisfy the conditions that (i) they are (asymptotically) uncorrelated with the errors in the equation of interest, and (ii) they *are* (asymptotically) correlated with the included endogenous variables in the equation.

Consider the problem of estimating one of a system of simultaneous equations, say the first one, which we might write as,

$$\begin{aligned} y_1 &= Y_1\gamma_1 + X_1\beta_1 + u_1 \\ &= Z_1\delta_1 + u_1. \end{aligned} \tag{3}$$

To connect this with our apparently more general class of models (1) take the first row of the matrix Γ in (1) to be $\Gamma_{1\cdot} = (1, -\gamma', 0')$ where the 0 corresponds to all of the endogenous variables that do not appear in the first equation. Let's begin by considering the simplest case in which (5) is exactly identified.

Partition the full set of exogenous variables as $X = (X_1 \vdots \tilde{X}_1)$ where X_1 is the set of included exogenous variables for equation one and $\tilde{X}_1$ and the excluded exogenous variables. In this case we may view $\tilde{X}_1$ as immediately available IV's for the Y_1's, since we have exactly, the same number of $\tilde{X}_1$'s as Y_1's.

Consider two apparently different instrumental variable estimators:

$$\hat{\delta}_1 = (\hat{Z}_1' Z_1)^{-1} \hat{Z}_1' y_1$$

that we will call the two stage least squares estimator, and

$$\tilde{\delta}_1 = (\tilde{Z}_1' Z_1)^{-1} \tilde{Z}_1 y_1$$

which is usually called the "indirect least squares" estimator, where $\hat{Z}_1 = P_X Z_1 = X(X'X)^{-1}X'Z_1$, and $\tilde{Z}_1 = (X_1 \vdots \tilde{X}_1) = X$. We will show that $\hat{\delta}_1 = \tilde{\delta}_1$.

We may write the claim more explicitly as

$$(Z_1' P_X Z_1)^{-1} Z_1' P_X y = (X'Z_1)^{-1} X' y$$

Now note that by assumption $X'Z_1$ is invertible so we can rewrite the lhs as

$$\begin{aligned} (Z_1' P_X Z_1)^{-1} Z_1' P_X y_1 &= (X'Z_1)^{-1} X'X (Z_1'X)^{-1} Z_1' X (X'X)^{-1} X' y_1 \\ &= (X'Z_1)^{-1} X' y_1 \end{aligned}$$

To explore the general case in which we have "more than enough" instrumental variables we consider the two stage least squares estimator in a bit more detail. Consider the model

$$y = Y_1\gamma + X_1\beta + u_1 = Z\delta + u_1.$$

and define instrumental variables $\tilde{Z} = XA$. Then it is easy to show that the instrumental variables estimator

$$\tilde{\delta} = (\tilde{Z}'Z)^{-1}\tilde{Z}'y$$

has the asymptotic linear representation

$$\sqrt{n}(\tilde{\delta} - \delta) = (n^{-1}A'X'Z)^{-1} n^{-1/2} A'X'u$$

and therefore,

$$\sqrt{n}(\tilde{\delta} - \delta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 (A'MD)^{-1} AMA (A'MD)^{-1})$$

where $\sigma^2 = Eu_i^2, M = \lim n^{-1}X'X$, $D = [\Pi_1 \vdots \Psi_1], X\Psi_1 = X_1$, and Π_1 is the Y_1 partition of the matrix of reduced form coefficients. The following result shows that among all possible choices of the matrix A, the two stage least squares choice has a claim to optimality.

Theorem.: The two stage least squares choice, $\hat{\delta}$ with $A = (X'X)^{-1}X'Z_1$ is optimal, i.e., $V(\hat{\delta}) \leq V(\tilde{\delta})$ for all A.

Proof:: Note $A \to D$ since, $(X'X)^{-1}X'Y_1 = \hat{\Pi}_1 \overset{P}{\to} \Pi_1$ and $(X'X)^{-1}X'X_1 = \Psi_1$. Thus, $\text{avar}(\hat{\delta}) \equiv \lim V(\sqrt{n}(\hat{\delta} - \delta)) = \sigma^2(D'MD)^{-1}$. We would like to show that for all $\alpha \in R^{p_1+q_1}$ $\alpha'(V(\hat{\delta}) - V(\tilde{\delta}))\alpha \leq 0$. It is equivalent, and slightly easier, to argue that for all $\alpha \in R^{p_1+q_1}$, $\alpha'(V(\hat{\delta})^{-1} - V(\tilde{\delta})^{-1})\alpha \geq 0$. Factor $M = NN'$(Cholesky decomposition) and set $h = ND\alpha$ as $h'h = \alpha'D'MD\alpha$ and

$$\alpha'V(\tilde{\delta})^{-1}\alpha = \alpha'D'MA(A'MA)^{-1}A'MD\alpha = h'G(G'G)^{-1}G'h$$

where $G = NA$, so we may write

$$\alpha'(V(\hat{\delta})^{-1} - V(\tilde{\delta})^{-1})\alpha = h'(I - G(G'G)^{-1}G')h \geq 0$$

as required, since the matrix in parentheses is a projection and therefore positive semidefinite.

5.1. **Control Variate Interpretation of 2SLS.** A somewhat less conventional interpretation of two-stage least-squares estimation reveals that the "endogeneity bias," the bias exhibited by the ordinary least-squares estimator of γ, can also be viewed as an omitted variable bias. Let

$$R = Y_1 - \hat{Y}_1 = Y_1 - P_X Y_1 \equiv M_X Y_1$$

denote the residuals from the least-squares regression of Y_1 on X, and consider the least squares estimator of γ in the model

$$y_1 = Z_1\delta + R\xi + \epsilon,$$

that is,

$$\hat{\delta}_{CV} = (Z_1^\top M_R Z_1)^{-1} Z_1^\top M_R y_1.$$

Here $M_R = I - R(R^\top R)^{-1}R^\top$ and CV stands for "control variate." Adding the columns of the matrix R to the structural form of the first equation "corrects" the ordinary least-squares estimator in precisely the same manner that the two-stage least-squares estimator does, that is,

$$\hat{\delta}_{CV} = \hat{\delta}_{2SLS}.$$

To see this it suffices to show that $Z_1^\top M_R = Z_1^\top P_X$. But since $M_R = I - P_R = I - P_{M_X Y_1}$, we have $X_1^\top M_R = X_1^\top = X_1^\top P_X$. This is clear from the fact that $P_R = P_{M_X Y_1} = M_X Y_1 (Y_1^\top M_X Y_1)^{-1} Y_1 M_X$ so $X_1^\top P_R = 0$. Also note that $Y_1^\top M_R = Y_1^\top (I - M_X) = Y_1^\top P_X$. Note that Y_1 may be multivariate in this formulation. After discussing this result, Blundell and Powell (2001), citing Dhrymes' 1974 text, comment that "it has been difficult to locate a definitive early reference to the control function version of two stage least squares." It seems plausible, however, that it is part of a quite ancient oral tradition in econometrics. Only relatively recently has it been widely recognized as a fruitful approach to more general models of endogeneity.

5.2. **Visual Instrumental Variables.** As a final installment in this rather loosely organized lecture, I'd like to try to describe a technique for visualizing the IV estimator in a scatter plot. What follows is my attempt to formalize somewhat the discussion in Section 4.1.3 of Angrist and Pischke (2009).

Consider the following simple model

$$y = \alpha + z\beta + u.$$

Suppose that z should be considered endogenous and for simplicity assume it is scalar. Supppose too that we have another variable, say f, that we would like to act like an instrumental variable. In `R` terminology f is a "factor," i.e., it takes discrete values $f \in \{1, \dots, J\}$. From f we can create a matrix F of indicator (dummy) variables $F_{ij} = 1$ if $f_i = j$, and $F_{ij} = 0$ otherwise. For example, f might be an occupational indicator, or in Angrist's context it might some grouped version of individual i's draft lottery number.

We have the following "reduced form" estimated equations

$$\begin{aligned} \hat{y} &= F\hat{\gamma} \\ \hat{z} &= F\hat{\delta} \end{aligned}$$

so we have two J-vectors $\hat{\gamma}$ and $\hat{\delta}$. Note that these estimates are simply the group means for y and z respectively determined by the F groups. That is, $\hat{\gamma}_j$ is just the mean of the n_j observations y_i that have $f_i = j$ for $j = 1, \cdots J$.

We now plot the J points $\bar{y} = \hat{\gamma}$ vs $\bar{z} = \hat{\delta}$ and overplot some sort of least squares line obtained from the regression,

$$(*) \qquad \hat{\gamma}_i = a + b\hat{\delta}_i + v_i$$

The question is what sort of least squares line would deliver an slope estimate equivalent to the 2SLS estimator? Suppose we consider OLS as a naive first thought:

$$\|\bar{y} - a - \bar{z}b\|^2 = \|(F^\top F)^{-1}F^\top y - a - (F^\top F)^{-1}F^\top zb\|^2$$

This is rather a mess, but if we modify it slightly, to do the GLS version,

$$\|\bar{y} - a - \bar{z}b\|^2_{(F^\top F)} = \|(F^\top F)^{-1}F^\top y - a - (F^\top F)^{-1}F^\top zb\|^2_{(F^\top F)} = \|y - a - zb\|^2_{P_F}$$

which is indeed the 2SLS estimator. This argument can be generalized somewhat to replace the intercept in our simple model with a vector of coefficients associated with some exogonous covariates and then apply the always useful Frisch-Waugh result to reduce the situation back to our simple case.

Obviously, the case that the IVs are just a set of discrete covariates is somewhat special, but it is a useful case to illustrate somewhat more geometrically how IV estimation works. By binning continuous covariates one can construct approximations to more general cases as well.

References

Blundell, R. and Powell, J.L., (2003) Endogeneity in nonparametric and semiparametric regression models, *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, 2, 655–679,

Dhrymes, P.J., (1974) *Econometrics: Statistical foundations and applications*, Springer-Verlag, New York.

Heckman, J.J.,(2000) Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective, *Quarterly Journal of Economics*, 115, 45–97.

Heckman, J.J. and Honore, B.E. (1990) The empirical content of the Roy model, *Econometrica*, 58, 1121–1149,

Pearl, J., (2003) Statistics and causal inference: A review, *Test*, 12, 281–345,

Schultz, H. (1938) *The theory and measurement of demand*, University of Chicago Press Chicago.

Working, E.J. (1928) What do statistical demand curves show? *The Quarterly Journal of Economics, 212–235.*