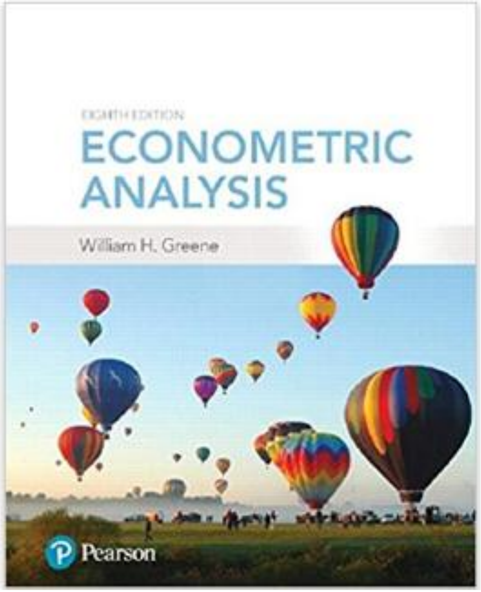
Econometrics I

Professor William Greene
Stern School of Business
Department of Economics



Econometrics I

Part 12 –Endogeneity and IV Estimation

Sources of “Endogeneity”

- Omitted Variables
- Ignored “Heterogeneity”
- Measurement Error
- Endogenous “Treatment Effects”
- Nonrandom Sampling (or Attrition)

Source of Endogeneity: Omitted Variable

Aggregate Data and Multinomial Choice: The Model of Berry, Levinsohn and Pakes

ECONOMETRICA
JOURNAL OF THE ECONOMETRIC SOCIETY

Automobile Prices in Market Equilibrium
Author(s): Steven Berry, James Levinsohn and Ariel Pakes
Source: *Econometrica*, Vol. 63, No. 4 (Jul., 1995), pp. 841-890
Published by: The Econometric Society
Stable URL: http://www.jstor.org/stable/2171802
Accessed: 08/12/2014 22:40

Theoretical Foundation

- Consumer market for J differentiated brands of a good
  - j =1,…, J_t brands or types
  - i = 1,…, N consumers
  - t = i,…,T "markets" (like panel data)
- Consumer i's utility for brand j (in market t) depends on
  - p = price
  - **x** = observable attributes
  - f = unobserved attributes
  - w = unobserved heterogeneity across consumers
  - ε = idiosyncratic aspects of consumer preferences
- Observed data consist of aggregate choices, prices and features of the brands.

BLP Automobile Market

J_t N P X

t

TABLE 1

DESCRIPTIVE STATISTICS

Year	No. of Models	Quantity	Price	Domestic	Japan	European	HP/Wt	Size	Air	MPG	MP$
1971	92	86.892	7.868	0.866	0.057	0.077	0.490	1.496	0.000	1.662	1.850
1972	89	91.763	7.979	0.892	0.042	0.066	0.391	1.510	0.014	1.619	1.875
1973	86	92.785	7.535	0.932	0.040	0.028	0.364	1.529	0.022	1.589	1.819
1974	72	105.119	7.506	0.887	0.050	0.064	0.347	1.510	0.026	1.568	1.453
1975	93	84.775	7.821	0.853	0.083	0.064	0.337	1.479	0.054	1.584	1.503
1976	99	93.382	7.787	0.876	0.081	0.043	0.338	1.508	0.059	1.759	1.696
1977	95	97.727	7.651	0.837	0.112	0.051	0.340	1.467	0.032	1.947	1.835
1978	95	99.444	7.645	0.855	0.107	0.039	0.346	1.405	0.034	1.982	1.929
1979	102	82.742	7.599	0.803	0.158	0.038	0.348	1.343	0.047	2.061	1.657
1980	103	71.567	7.718	0.773	0.191	0.036	0.350	1.296	0.078	2.215	1.466
1981	116	62.030	8.349	0.741	0.213	0.046	0.349	1.286	0.094	2.363	1.559
1982	110	61.893	8.831	0.714	0.235	0.051	0.347	1.277	0.134	2.440	1.817
1983	115	67.878	8.821	0.734	0.215	0.051	0.351	1.276	0.126	2.601	2.087
1984	113	85.933	8.870	0.783	0.179	0.038	0.361	1.293	0.129	2.469	2.117
1985	136	78.143	8.938	0.761	0.191	0.048	0.372	1.265	0.140	2.261	2.024
1986	130	83.756	9.382	0.733	0.216	0.050	0.379	1.249	0.176	2.416	2.856
1987	143	67.667	9.965	0.702	0.245	0.052	0.395	1.246	0.229	2.327	2.789
1988	150	67.078	10.069	0.717	0.237	0.045	0.396	1.251	0.237	2.334	2.919
1989	147	62.914	10.321	0.690	0.261	0.049	0.406	1.259	0.289	2.310	2.806
1990	131	66.377	10.337	0.682	0.276	0.043	0.419	1.270	0.308	2.270	2.852
All	2217	78.804	8.604	0.790	0.161	0.049	0.372	1.357	0.116	2.099	2.086

Note: The entry in each cell of the last nine columns is the sales weighted mean.

Random Utility Model

- Utility: $U_{ijt}=U(w_i,p_{jt},\mathbf{x}_{jt},f_{jt},\varepsilon_{ijt}\,|\theta)$, i = 1,…,(large) N, j=1,…,J
  - w_i = individual heterogeneity; time (market) invariant. w has a continuous distribution across the population.
  - p_{jt}, $\mathbf{x}_{jt}$, f_{jt}, = price, observed attributes, unobserved features of brand j; all may vary through time (across markets)
- Revealed Preference: Choice j provides maximum utility
- Across the population, given market t, set of prices $\mathbf{p}_t$ and features $(\mathbf{X}_t,\mathbf{f}_t)$, there is a set of values of w_i that induces choice j, for each j=1,…,J_t; then, $s_j(\mathbf{p}_t,\mathbf{X}_t,\mathbf{f}_t|\theta)$ is the market share of brand j in market t.
- There is an outside good that attracts a nonnegligible market share, j=0. Therefore, $\sum_{j=1}^{J_t} s_j(\mathbf{p}_t,\mathbf{X}_t,\mathbf{f}_t\,|\,\boldsymbol{\theta}) < 1$

Endogenous Prices: Demand side

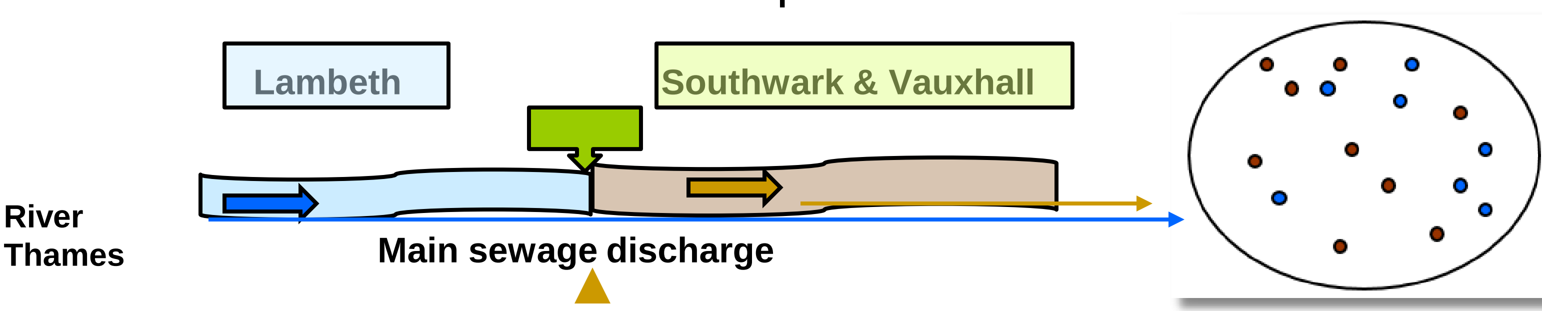
- $U_{ijt} = U(w_i, p_{jt}, \mathbf{x}_{jt}, f_{jt}, \varepsilon_{ijt} | \theta) = \mathbf{x}_{jt}'\boldsymbol{\beta}_i - \alpha p_j + f_{jt} + \varepsilon_{ijt}$
- f_{jt} is unobserved features of model j
- Utility responds to the unobserved f_{jt}
- Price p_{jt} is partly determined by features f_{jt}.
- In a choice model based on observables, price is correlated with the unobservables that determine the observed choices.

An Early Study of an Endogeneity Problem

(Snow, J., On the Mode of Communication of Cholera, 1855)
http://www.ph.ucla.edu/epi/snow/snowbook3.html

- London Cholera epidemic, ca 1853-4
- Cholera = f(Water Purity,u) + ε.
  - ‘Causal’ effect of water purity on cholera?
  - Purity=f(cholera prone environment (poor, garbage in streets, rodents, etc.). Regression does not work.

  Two London water companies



Paul Grootendorst: A Review of Instrumental Variables Estimation of Treatment Effects…
http://individual.utoronto.ca/grootendorst/pdf/IV_Paper_Sept6_2007.pdf

A review of instrumental variables estimation in the applied health sciences. *Health Services and Outcomes Research Methodology* 2007; 7(3-4):159-179.

Cornwell and Rupert Data

Cornwell and Rupert Returns to Schooling Data, 595 Individuals, 7 Years
Variables in the file are

EXP	= work experience
WKS	= weeks worked
OCC	= occupation, 1 if blue collar,
IND	= 1 if manufacturing industry
SOUTH	= 1 if resides in south
SMSA	= 1 if resides in a city (SMSA)
MS	= 1 if married
FEM	= 1 if female
UNION	= 1 if wage set by union contract
ED	= years of education
LWAGE	= log of wage = dependent variable in regressions

These data were analyzed in Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators," Journal of Applied Econometrics, 3, 1988, pp. 149-155. See Baltagi, page 122 for further analysis. The data were downloaded from the website for Baltagi's text.

Specification: Quadratic Effect of Experience

Ordinary	least squares regression					
LHS=LWAGE	Mean	=	6.67635			
	Standard deviation	=	.46151			
----------	No. of observations	=	4165	DegFreedom		Mean square
Regression	Sum of Squares	=	370.955	10		37.09546
Residual	Sum of Squares	=	515.950	4154		.12421
Total	Sum of Squares	=	886.905	4164		.21299
----------	Standard error of e	=	.35243	Root MSE		.35196
Fit	R-squared	=	.41826	R-bar squared		.41686
Model test	F[10, 4154]	=	298.66153	Prob F > F*		.00000

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	5.24547***	.07170	73.15	.0000	5.10493	5.38600
ED	.05654***	.00261	21.64	.0000	.05142	.06166
EXP	.04045***	.00217	18.61	.0000	.03619	.04471
EXP*EXP	-.00068***	.4783D-04	-14.24	.0000	-.00077	-.00059
WKS	.00449***	.00109	4.12	.0000	.00235	.00662
OCC	-.14053***	.01472	-9.54	.0000	-.16939	-.11167
SOUTH	-.07210***	.01249	-5.77	.0000	-.09658	-.04762
SMSA	.13901***	.01207	11.51	.0000	.11534	.16267
MS	.06736***	.02063	3.26	.0011	.02692	.10779
FEM	-.38922***	.02518	-15.46	.0000	-.43857	-.33987
UNION	.09015***	.01289	6.99	.0000	.06488	.11542

nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
***, **, * ==> Significance at 1%, 5%, 10% level.

The Effect of Education on LWAGE

$$\textbf{LWAGE} = \beta_1 + \beta_2\textbf{EDUC} + \beta_3\textbf{EXP} + \beta_4\textbf{EXP}^2 + \ldots + \varepsilon$$

What is ε? Ability,... + everything else

$$\textbf{EDUC} = f(\textbf{GENDER}, \textbf{SMSA}, \textbf{SOUTH}, \text{Ability}, \ldots, u)$$

What Influences LWAGE?

$$\textbf{LWAGE} = \beta_1 + \beta_2 \textbf{EDUC}(\textbf{X}, \text{Ability}, \ldots) + \beta_3 \textbf{EXP} + \beta_4 \textbf{EXP}^2 + \ldots + \varepsilon(\text{Ability})$$

Increased Ability is associated with increases in **EDUC**(**X**, Ability,...,u) and ε(Ability)

What looks like an effect due to increase in **EDUC** may be an increase in Ability. The estimate of β_2 picks up the effect of **EDUC** and the hidden effect of Ability.

An Exogenous Influence

$$\textbf{LWAGE} = \beta_1 + \beta_2 \textbf{EDUC}(\textbf{X}, \textbf{Z}, \text{Ability}, \ldots) + \beta_3 \textbf{EXP} + \beta_4 \textbf{EXP}^2 + \ldots + \varepsilon(\text{Ability})$$

Increased **Z** is associated with increases in $\textbf{EDUC}(\textbf{X}, \textbf{Z}, \text{Ability}, \ldots, u)$ and not $\varepsilon(\text{Ability})$

An effect due to the effect of an increase **Z** on **EDUC** will only be an increase in **EDUC**. The estimate of β_2 picks up the effect of **EDUC** only.

Z is an Instrumental Variable

Instrumental Variables

- Structure
  - LWAGE (**ED,EXP,EXPSQ,WKS,OCC, SOUTH,SMSA,UNION**)
  - ED (**MS**, **FEM**)

- Reduced Form:
  LWAGE[**ED** (**MS**, **FEM**), **EXP,EXPSQ,WKS,OCC, SOUTH,SMSA,UNION**]

Two Stage Least Squares Strategy

- Reduced Form:
  LWAGE[**ED** (**MS**, **FEM**,**X**),
  EXP,EXPSQ,WKS,OCC,
  SOUTH,SMSA,UNION]
- Strategy
  - (1) Purge ED of the influence of everything but MS, FEM (and the other variables). Predict ED using all exogenous information in the sample (**X** and **Z**).
  - (2) Regress LWAGE on this prediction of ED and everything else.
  - Standard errors must be adjusted for the predicted ED

OLS

```
----------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=LWAGE    Mean                    =       6.67635
             Standard deviation      =        .46151
----------   No. of observations     =          4165  DegFreedom   Mean square
Regression   Sum of Squares          =       291.042            8     36.38019
Residual     Sum of Squares          =       595.863         4156       .14337
Total        Sum of Squares          =       886.905         4164       .21299
----------   Standard error of e     =        .37865  Root MSE          .37824
Fit          R-squared               =        .32815  R-bar squared     .32686
Model test   F[  8,  4156]           =     253.74283  Prob F > F*       .00000
|--------+-------------------------------------------------------------------
        |                 Standard             Prob.       95% Confidence
   LWAGE|  Coefficient      Error       z      |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|    4.97986***     .07430    67.02   .0000      4.83424    5.12549
     EXP|     .04308***     .00232    18.54   .0000       .03853     .04764
   EXPSQ|    -.00070***   .5128D-04  -13.68   .0000      -.00080    -.00060
     WKS|     .00760***     .00116     6.53   .0000       .00532     .00988
     OCC|    -.11578***     .01578    -7.34   .0000      -.14672    -.08485
   SOUTH|    -.08207***     .01341    -6.12   .0000      -.10835    -.05578
    SMSA|     .09885***     .01285     7.69   .0000       .07367     .12403
   UNION|     .12891***     .01374     9.38   .0000       .10197     .15584
      ED|     .06365***     .00279    22.82   .0000       .05818     .06911
--------+-------------------------------------------------------------------
Note: nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
Note: ***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

```
-----------------------------------------------------------------------------
Two stage     least squares regression ............
LHS=LWAGE     Mean                    =         6.67635
              Standard deviation      =          .46151
              Number of observs.      =            4165
Model size    Parameters              =               9
              Degrees of freedom      =            4156
Residuals     Sum of squares          =         6921.67
              Standard error of e     =         1.29053
Fit           R-squared               =        -6.82120
              Adjusted R-squared      =        -6.83625
Not using OLS or no constant. Rsqrd & F may be < 0
Instrumental Variables:
ONE        MS         FEM        EXP        Intrct01   WKS
OCC        SOUTH      SMSA       UNION
```

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	-4.38670***	1.40197	-3.13	.0018	-7.13451	-1.63889
EXP	.06447***	.00852	7.56	.0000	.04777	.08117
EXP*EXP	-.00058***	.00018	-3.32	.0009	-.00093	-.00024
WKS	.01533***	.00413	3.72	.0002	.00725	.02342
OCC	1.71424***	.27473	6.24	.0000	1.17578	2.25270
SOUTH	.31274***	.07394	4.23	.0000	.16782	.45767
SMSA	-.13695**	.05588	-2.45	.0142	-.24647	-.02744
UNION	.37025***	.05879	6.30	.0000	.25502	.48548
ED	.65029***	.08689	7.48	.0000	.48000	.82059

***, **, * ==> Significance at 1%, 5%, 10% level.

4.97986***
.04308***
-.00070***
.00760***
-.11578***
-.08207***
.09885***
.12891***
.06365***

The weird results for the coefficient on ED happened because the instruments, MS and FEM are dummy variables. There is not enough variation in these variables.

The Ultimate Source of Endogeneity

- **LWAGE** = f(**ED,**
  EXP,EXPSQ,WKS,OCC,
  SOUTH,SMSA,UNION) + ε
- **ED** = f(**MS,FEM,**
  EXP,EXPSQ,WKS,OCC,
  SOUTH,SMSA,UNION) + u

Remove the Endogeneity

- **LWAGE** = f(**ED,** **EXP,EXPSQ,WKS,OCC,** **SOUTH,SMSA,UNION**) + u + ε

- Strategy
  - Estimate u
  - Add u to the equation. ED is uncorrelated with ε when u is in the equation.

Auxiliary Regression for ED to Obtain Residuals

```
----------------------------------------------------------------------------
Ordinary    least squares regression ............
LHS=ED      Mean                     =     12.84538
            Standard deviation       =      2.78800
----------  No. of observations      =         4165  DegFreedom   Mean square
Regression  Sum of Squares           =      14162.8           9    1573.64724
Residual    Sum of Squares           =      18203.6        4155       4.38113
Total       Sum of Squares           =      32366.4        4164       7.77292
----------  Standard error of e       =      2.09312  Root MSE         2.09060
Fit         R-squared                =       .43758  R-bar squared     .43636
Model test  F[  9,  4155]            =    359.18746  Prob F > F*       .00000
--------+-------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
      ED|  Coefficient       Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|    16.0756***      .34520    46.57  .0000     15.3990    16.7521
      MS|     .27698**       .12245     2.26  .0237      .03698     .51698
     FEM|    -.46653***      .14937    -3.12  .0018     -.75929    -.17376
     EXP|    -.04189***      .01290    -3.25  .0012     -.06716    -.01661
 EXP*EXP|    -.00014         .00028     -.50  .6181     -.00070     .00042
     WKS|    -.01810***      .00647    -2.80  .0051     -.03078    -.00543
     OCC|   -3.12102***      .07282   -42.86  .0000    -3.26376   -2.97829
   SOUTH|    -.65003***      .07349    -8.85  .0000     -.79407    -.50599
    SMSA|     .46655***      .07134     6.54  .0000      .32672     .60638
   UNION|    -.47323***      .07621    -6.21  .0000     -.62260    -.32385
--------+-------------------------------------------------------------------
***, **, * ==>  Significance at 1%, 5%, 10% level.
----------------------------------------------------------------------------
```

IVs (MS, FEM)

Exog. Vars (EXP, EXP*EXP, WKS, OCC, SOUTH, SMSA, UNION)

OLS with Residual (Control Function) Added

```
-----------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=LWAGE    Mean                   =      6.67635
             Standard deviation     =       .46151
----------   No. of observations    =         4165  DegFreedom   Mean square
Regression   Sum of Squares         =      367.888          9      40.87643
Residual     Sum of Squares         =      519.017       4155        .12491
Total        Sum of Squares         =      886.905       4164        .21299
----------   Standard error of e    =       .35343  Root MSE         .35301
Fit          R-squared              =       .41480  R-bar squared    .41353
Model test   F[  9,  4155]          =    327.23700  Prob F > F*      .00000
--------+--------------------------------------------------------------------
        |                Standard            Prob.      95% Confidence
   LWAGE|  Coefficient     Error       z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
Constant|   -4.38670***    .38395   -11.43  .0000     -5.13923  -3.63417
     EXP|     .06447***    .00233    27.62  .0000       .05990    .06904
 EXP*EXP|    -.00058***  .4810D-04  -12.13  .0000      -.00068   -.00049
     WKS|     .01533***    .00113    13.57  .0000       .01312    .01755
     OCC|    1.71424***    .07524    22.78  .0000      1.56678   1.86171
   SOUTH|     .31274***    .02025    15.44  .0000       .27305    .35243
    SMSA|    -.13695***    .01530    -8.95  .0000      -.16695   -.10696
   UNION|     .37025***    .01610    23.00  .0000       .33869    .40180
      ED|     .65029***    .02380    27.33  .0000       .60366    .69693
       U|    -.59376***    .02394   -24.80  .0000      -.64068   -.54684
--------+--------------------------------------------------------------------
nnnnn.D-xx or D+xx => multiply by 10 to -xx or +xx.
***, **, * ==>  Significance at 1%, 5%, 10% level.
-----------------------------------------------------------------------------
```

2SLS

```
-4.38670***
  .06447***
 -.00058***
  .01533***
 1.71424***
  .31274***
 -.13695**
  .37025***
  .65029***
```

A Warning About Control Function Estimators: The standard errors must be adjusted.

```
Two stage      least squares regression .............
               Standard error of e   =        1.29053
```

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	-4.38670***	1.40197	-3.13	.0018	-7.13451	-1.63889
EXP	.06447***	.00852	7.56	.0000	.04777	.08117
EXP*EXP	-.00058***	.00018	-3.32	.0009	-.00093	-.00024
WKS	.01533***	.00413	3.72	.0002	.00725	.02342
OCC	1.71424***	.27473	6.24	.0000	1.17578	2.25270
SOUTH	.31274***	.07394	4.23	.0000	.16782	.45767
SMSA	-.13695**	.05588	-2.45	.0142	-.24647	-.02744
UNION	.37025***	.05879	6.30	.0000	.25502	.48548
ED	.65029***	.08689	7.48	.0000	.48000	.82059

```
Residual augmented least squares regression .............
----------     Standard error of e   =         .35343
```

$$0.38395 \times \frac{1.29053}{0.35343} = 1.40197$$

LWAGE	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	-4.38670***	.38395	-11.43	.0000	-5.13923	-3.63417
EXP	.06447***	.00233	27.62	.0000	.05990	.06904
EXP*EXP	-.00058***	.4810D-04	-12.13	.0000	-.00068	-.00049
WKS	.01533***	.00113	13.57	.0000	.01312	.01755
OCC	1.71424***	.07524	22.78	.0000	1.56678	1.86171
SOUTH	.31274***	.02025	15.44	.0000	.27305	.35243
SMSA	-.13695***	.01530	-8.95	.0000	-.16695	-.10696
UNION	.37025***	.01610	23.00	.0000	.33869	.40180
ED	.65029***	.02380	27.33	.0000	.60366	.69693
U	-.59376***	.02394	-24.80	.0000	-.64068	-.54684

I am here to ask a little help for endogeneity.

I have a main regression, in which the independent variabels are lagged 1 year (this is an unbalanced panel dataset); I use fixed effect, xtreg:

Main Regression: Yt = Xt-1 + Qt-1 + Z3t-1

I suspect endogeneity: variable X may be itself determined by prior-year Y. As a solution, I read this strategy: regress the endogenous variable Xt-1 on the dependent variable (Yt-2) and other independent variables (i.e., Qt-2 and Zt-2); these Y Q and Z are all **in year t-2,** while X is in t-1. Then, from this regression, calculate the "**predicted" values for X, and include them as a control-for-endogeneity** (e.g., a variable named "Endogeneity-control") in the main regression above.

Question 1: in the Main Regression above, when including the control for endogeneity (i.e., the variable "Endogeneity-control"), do I have to lag its value? That is, do I have to include Endogeneity-control in t-1? or just the predicted values, without lagging?

The two stage LS strategy: (The two stage button in your software.) The software regresses EDUC on all independent variables plus the two instrumental variables (stage 1), then takes the predicted value on education and regresses lwage on that predicted value plus the original independent variables (stage 2). Is this correct?

Then the second method you showed is the same except the predicted residuals are included in the second stage OLS.

Is one method preferred over another? They produce the same results.

The General Problem

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta} + \mathbf{X}_2\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

$\text{Cov}(\mathbf{X}_1, \boldsymbol{\varepsilon}) = \mathbf{0}$, K_1 variables

$\text{Cov}(\mathbf{X}_2, \boldsymbol{\varepsilon}) \neq \mathbf{0}$, K_2 variables

$\mathbf{X}_2$ is **endogenous**

OLS regression of y on $(\mathbf{X}_1, \mathbf{X}_2)$ cannot estimate $(\boldsymbol{\beta}, \boldsymbol{\delta})$ consistently. Some other estimator is needed.

Additional structure:

$\mathbf{X}_2 = \mathbf{Z}\boldsymbol{\Pi} + \mathbf{V}$ where $\text{Cov}(\mathbf{Z}, \boldsymbol{\varepsilon}) = \mathbf{0}$.

An **instrumental variable (IV)** estimator based on $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{Z})$ may be able to estimate $(\boldsymbol{\beta}, \boldsymbol{\delta})$ consistently.

Instrumental Variables

- Fully General Framework: $\mathbf{y} = \mathbf{X}\beta + \varepsilon$, K variables in **X**.
- There exists a set of M=K variables, **Z** such that

  $$\text{plim}(\mathbf{Z'X/n}) \neq \mathbf{0} \quad \text{but} \quad \text{plim}(\mathbf{Z'}\varepsilon/n) = \mathbf{0}$$

  The variables in **Z** are called instrumental variables.
- An alternative (to least squares) estimator of β is

  $$\mathbf{b}_{IV} = (\mathbf{Z'X})^{-1}\mathbf{Z'y}$$

- We consider the following:
  - Why use this estimator?
  - What are its properties compared to least squares?
- We will also examine an important application

IV Estimators

Consistent

$$\mathbf{b}_{IV} = (\mathbf{Z'X})^{-1}\mathbf{Z'y}$$

$$= (\mathbf{Z'X}/n)^{-1} (\mathbf{Z'X}/n)\boldsymbol{\beta} + (\mathbf{Z'X}/n)^{-1}\mathbf{Z'}\boldsymbol{\varepsilon}/n$$

$$= \boldsymbol{\beta} + (\mathbf{Z'X}/n)^{-1}\mathbf{Z'}\boldsymbol{\varepsilon}/n \rightarrow \boldsymbol{\beta}$$

Asymptotically normal (same approach to proof as for OLS)

Inefficient – to be shown.

The General Result

By construction, the IV estimator is consistent. So, we have an estimator that is consistent when least squares is not.

LS as an IV Estimator

The least squares estimator is

$$(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\Sigma_i \mathbf{x}_i y_i$$

$$= \beta + (\mathbf{X}'\mathbf{X})^{-1}\Sigma_i \mathbf{x}_i \varepsilon_i$$

If plim($\mathbf{X}'\mathbf{X}$/n) = $\mathbf{Q}$ nonzero

plim($\mathbf{X}'\boldsymbol{\varepsilon}$/n) = $\mathbf{0}$

Under the usual assumptions LS is an IV estimator

X is its own instrument.

IV Estimation

Why use an IV estimator? Suppose that $\mathbf{X}$ and $\boldsymbol{\varepsilon}$ are *not* uncorrelated. Then least squares is neither unbiased nor consistent.

Recall the proof of consistency of least squares:

$$\mathbf{b} = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X}/n)^{-1}(\mathbf{X}'\boldsymbol{\varepsilon}/n).$$

Plim $\mathbf{b} = \boldsymbol{\beta}$ requires plim$(\mathbf{X}'\boldsymbol{\varepsilon}/n) = \mathbf{0}$. If this does not hold, the estimator is inconsistent.

A Popular Misconception

A popular misconception. If only one variable in **X** is correlated with ε, the other coefficients are consistently estimated. False.

Suppose only the first variable is correlated with $\boldsymbol{\varepsilon}$

Under the assumptions, plim(**X'ε**/n) = $\begin{pmatrix} \sigma_{1\varepsilon} \\ 0 \\ \cdots \\ . \end{pmatrix}$.

Then, plim **b** - **β** = plim(**X'X**/n)$^{-1}$ $\begin{pmatrix} \sigma_{1\varepsilon} \\ 0 \\ \cdots \\ . \end{pmatrix} = \sigma_{1\varepsilon} \begin{pmatrix} q^{11} \\ q^{21} \\ \cdots \\ q^{K1} \end{pmatrix}$

$= \sigma_{1\varepsilon}$ times the first column of $\mathbf{Q}^{-1}$.

The problem is "smeared" over the other coefficients.

Asymptotic Covariance Matrix of $\mathbf{b}_{IV}$

$$\mathbf{b}_{\text{IV}} - \beta = (\mathbf{Z'X})^{-1}\mathbf{Z}'\varepsilon$$

$$(\mathbf{b}_{\text{IV}} - \beta)(\mathbf{b}_{\text{IV}} - \beta)' = (\mathbf{Z'X})^{-1}\mathbf{Z}'\varepsilon\varepsilon'\mathbf{Z}(\mathbf{X'Z})^{-1}$$

$$E[(\mathbf{b}_{\text{IV}} - \beta)(\mathbf{b}_{\text{IV}} - \beta)' \mid \mathbf{X},\mathbf{Z}] = \sigma^2(\mathbf{Z'X})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{X'Z})^{-1}$$

Asymptotic Efficiency

Asymptotic efficiency of the IV estimator. The variance is larger than that of LS. (A large sample type of Gauss-Markov result is at work.)

(1) It's a moot point. LS is inconsistent.

(2) Mean squared error is uncertain:

MSE[estimator|**β**]=Variance + square of bias.

IV may be better or worse. Depends on the data

Two Stage Least Squares

How to use an "excess" of instrumental variables

(1) **X** is K variables. Some (at least one) of the K variables in **X** are correlated with **ε**.

(2) **Z** is now M > K variables. Some of the variables in **Z** are also in **X**, some are not. None of the variables in **Z** are correlated with **ε.**

(3) Which K variables to use to compute **Z'X** and **Z'y?**

Choosing the Instruments

- Choose K randomly?
- Choose the included Xs and the remainder randomly?
- Use all of them? How?
- A theorem: (Brundy and Jorgenson, ca. 1972) There is a most efficient way to construct the IV estimator from this subset:
  - (1) For each column (variable) in **X**, compute the predictions of that variable using all the columns of **Z**.
  - (2) Linearly regress **y** on these K predictions.
- This is two stage least squares

Algebraic Equivalence

- Two stage least squares is equivalent to
  - (1) each variable in **X** that is also in **Z** is replaced by itself.
  - (2) Variables in **X** that are not in **Z** are replaced by predictions of that **X** with all the variables in **Z**. Coefficients in augmented regression are added to match 2SLS. (They match if residuals are used instead of predictions.)

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it},$$

$$\ln Wage_{it} = \gamma_1 + \gamma_2 Ind_{it} + \gamma_3 Ed_i + \gamma_3 Union_{it} + \gamma_4 Fem_i + \gamma_5 SMSA_{it} + u_i.$$

```
name;w=one,ed,union,fem,ind,smsa$
name;x=one,lwage,ed,union,fem$
name;z=one,ind,ed,union,fem,smsa$
regr;lhs=lwage;rhs=one,ind,ed,union,fem,smsa
    ;keep=lwageh;res=u$
regr;lhs=wks;rhs=x,lwageh$
2sls;lhs=wks;rhs=x;inst=z$
```

Ordinary least squares regression ....

LHS=WKS	Mean	=	46.81152			
	Standard deviation	=	5.12910			
----------	No. of observations	=	4165	DegFreedom		Mean square
Regression	Sum of Squares	=	4640.01	5		928.00292
Residual	Sum of Squares	=	104905.	4159		25.22362
Total	Sum of Squares	=	109545.	4164		26.30765
----------	Standard error of e	=	5.02231	Root MSE		5.01869
Fit	R-squared	=	.04236	R-bar squared		.04121
Model test	F[5, 4159]	=	36.79103	Prob F > F*		.00000

WKS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	30.7044***	4.90997	6.25	.0000	21.0810	40.3277
LWAGE	.59245***	.20262	2.92	.0035	.19533	.98958
ED	-.31997***	.06489	-4.93	.0000	-.44714	-.19280
UNION	-2.19398***	.18262	-12.01	.0000	-2.55191	-1.83604
FEM	-.23784	.45954	-.52	.6048	-1.13852	.66284
LWAGEH	2.55937***	.86588	2.96	.0031	.86227	4.25646

Sum=2sls

Two stage least squares regression ....
Instrumental Variables:
ONE IND ED UNION FEM SMSA

WKS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	30.7044***	4.99966	6.14	.0000	20.9052	40.5035
LWAGE	3.15182***	.85722	3.68	.0002	1.47171	4.83193
ED	-.31997***	.06607	-4.84	.0000	-.44947	-.19048
UNION	-2.19398***	.18596	-11.80	.0000	-2.55845	-1.82951
FEM	-.23784	.46793	-.51	.6113	-1.15497	.67929

2SLS Algebra

$\hat{\mathbf{X}} = \mathbf{Z(Z'Z)^{-1}Z'X}$

$\mathbf{b_{2SLS}} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\mathbf{y}$

But, $\mathbf{Z(Z'Z)^{-1}Z'X} = (\mathbf{I} - \mathbf{M_Z})\mathbf{X}$ and $(\mathbf{I} - \mathbf{M_Z})$ is idempotent.

$\hat{\mathbf{X}}'\hat{\mathbf{X}} = \mathbf{X}'(\mathbf{I} - \mathbf{M_Z})(\mathbf{I} - \mathbf{M_Z})\mathbf{X} = \mathbf{X}'(\mathbf{I} - \mathbf{M_Z})\mathbf{X}$ so

$\mathbf{b_{2SLS}} = (\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\mathbf{y}$ = a real IV estimator by the definition.

Note, plim($\hat{\mathbf{X}}'\varepsilon/n$) = $\mathbf{0}$ since columns of $\hat{\mathbf{X}}$ are linear combinations of the columns of $\mathbf{Z}$, all of which are uncorrelated with ε.

$\mathbf{b_{2SLS}} = [\mathbf{X}'(\mathbf{I} - \mathbf{M_Z})\mathbf{X}]^{-1}\mathbf{X}'(\mathbf{I} - \mathbf{M_Z})\mathbf{y}$

Asymptotic Covariance Matrix for 2SLS

General Result for Instrumental Variable Estimation

$$E[(\mathbf{b}_{IV}-\beta)(\mathbf{b}_{IV}-\beta)'\,|\,\mathbf{X},\mathbf{Z}] = \sigma^2(\mathbf{Z'X})^{-1}\mathbf{Z'Z}(\mathbf{X'Z})^{-1}$$

Specialize for 2SLS, using $\mathbf{Z} = \hat{\mathbf{X}} = (\mathbf{I}-\mathbf{M_Z})\mathbf{X}$

$$E[(\mathbf{b}_{2SLS}-\beta)(\mathbf{b}_{2SLS}-\beta)'\,|\,\mathbf{X},\mathbf{Z}] = \sigma^2(\hat{\mathbf{X}}'\mathbf{X})^{-1}\hat{\mathbf{X}}'\hat{\mathbf{X}}(\mathbf{X}'\hat{\mathbf{X}})^{-1}$$

$$= \sigma^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}\hat{\mathbf{X}}'\hat{\mathbf{X}}(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

$$= \sigma^2(\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$$

2SLS has larger variance (around its mean) than LS has around its mean.

A comparison to OLS

$\text{Asy.Var[2SLS]} = \sigma^2 (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}$

Neglecting the inconsistency,

$\text{Asy.Var[LS]} \quad = \sigma^2 (\mathbf{X}'\mathbf{X})^{-1}$

(This is the variance of LS around its mean, not $\boldsymbol{\beta}$)

Asy.Var[2SLS] $\geq$ Asy.Var[LS] in the matrix sense.

To prove, compare the inverses:

$\{\text{Asy.Var[LS]}\}^{-1} - \{\text{Asy.Var[2SLS]}\}^{-1} = (1/\sigma^2)[\mathbf{X}'\mathbf{X} - \hat{\mathbf{X}}'\hat{\mathbf{X}}]$

$= (1/\sigma^2)[\mathbf{X}'\mathbf{X} - \mathbf{X}'(\mathbf{I} - \mathbf{M}_Z)\mathbf{X}] = (1/\sigma^2)[\mathbf{X}'\mathbf{M}_Z\mathbf{X}]$

This matrix is nonnegative definite. (Not positive definite as it might have some rows and columns which are zero.)

Implication for "precision" of 2SLS: Possibly very large variances.

The problem of "Weak Instruments"

Estimating σ^2

Estimating the asymptotic covariance matrix - a caution about estimating σ^2.

Since the regression is computed by regressing y on $\hat{\mathbf{x}}$, one might use

$$\hat{\sigma}^2 = \tfrac{1}{n}\Sigma_{i=1}^{n}(y_i - \hat{\mathbf{x}}'\mathbf{b}_{\mathbf{2sls}})$$

This is inconsistent. Use

$$\hat{\sigma}^2 = \tfrac{1}{n}\Sigma_{i=1}^{n}(y_i - \mathbf{x}'\mathbf{b}_{\mathbf{2sls}})$$

(Degrees of freedom correction is optional. Conventional, but not necessary.)

Two Problems with 2SLS

- **Z'X**/n may not be sufficiently large. The covariance matrix for the IV estimator is Asy.Cov(b) = $\sigma^2[(\mathbf{Z'X})(\mathbf{Z'Z})^{-1}(\mathbf{X'Z})]^{-1}$
  - If **Z'X**/n -> 0, the variance explodes.
  - Additional problems:
    - 2SLS biased toward plim OLS
    - Asymptotic results for inference fall apart.
- When there are many instruments, $\hat{\mathbf{x}}$ is too close to **x**; 2SLS becomes OLS.

Weak Instruments

- Symptom: The **relevance condition**, plim **Z'X**/n not zero, but is close to being violated.
- Detection:
  - Standard F test in the regression of x_k on Z. F < 10 suggests a problem.
  - F statistic based on 2SLS – see text p. 274.
- Remedy:
  - Not much – most of the discussion is about the condition, not what to do about it.
  - Use LIML? Requires a normality assumption. Probably not too restrictive.

Cornwell and Rupert Data

Cornwell and Rupert Returns to Schooling Data, 595 Individuals, 7 Years
Variables in the file are

EXP	= work experience
WKS	= weeks worked
OCC	= occupation, 1 if blue collar,
IND	= 1 if manufacturing industry
SOUTH	= 1 if resides in south
SMSA	= 1 if resides in a city (SMSA)
MS	= 1 if married
FEM	= 1 if female
UNION	= 1 if wage set by union contract
ED	= years of education
LWAGE	= log of wage = dependent variable in regressions

These data were analyzed in Cornwell, C. and Rupert, P., "Efficient Estimation with Panel Data: An Empirical Comparison of Instrumental Variable Estimators," Journal of Applied Econometrics, 3, 1988, pp. 149-155. See Baltagi, page 122 for further analysis. The data were downloaded from the website for Baltagi's text.

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it},$$
$$\ln Wage_{it} = \gamma_1 + \gamma_2 Ind_{it} + \gamma_3 Ed_i + \gamma_3 Union_{it} + \gamma_4 Fem_i + \gamma_5 SMSA_{it} + u_i.$$

Endogenous | Exogenous | Instruments

```
|-> regr;lhs=lwage;rhs=z;test:ind=0,smsa=0$
-----------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=LWAGE    Mean                      =      6.67635
             Standard deviation        =       .46151
----------   No. of observations       =         4165  DegFreedom   Mean square
Regression   Sum of Squares            =      272.516           5      54.50318
Residual     Sum of Squares            =      614.389        4159        .14773
Total        Sum of Squares            =      886.905        4164        .21299
----------   Standard error of e       =       .38435  Root MSE          .38407
Fit          R-squared                 =       .30727  R-bar squared     .30643
Model test   F[  5,  4159]             =    368.94981  Prob F > F*       .00000
Wald Test:   Chi-squared [   2]        =      240.932  Prob C2 > C2* =   .00000
F Test:      F ratio[ 2,  4159]        =      120.466  Prob F  > F*  =   .00000
--------+--------------------------------------------------------------------
        |                 Standard            Prob.      95% Confidence
   LWAGE|  Coefficient      Error       z    |z|>Z*         Interval
--------+--------------------------------------------------------------------
Constant|    5.71494***     .03299   173.25  .0000      5.65028   5.77959
     IND|     .08134***     .01278     6.36  .0000       .05629    .10640
      ED|     .06547***     .00232    28.24  .0000       .06093    .07002
   UNION|     .05859***     .01303     4.50  .0000       .03305    .08414
     FEM|    -.47009***     .01939   -24.24  .0000      -.50810   -.43208
    SMSA|     .18329***     .01287    14.24  .0000       .15807    .20851
--------+--------------------------------------------------------------------
```

8.4.3 LIMITED INFORMATION MAXIMUM LIKELIHOOD[6]

We have considered estimation of the two equation model,

$$Wks_{it} = \beta_1 + \beta_2 \ln Wage_{it} + \beta_3 Ed_i + \beta_4 Union_{it} + \beta_5 Fem_i + \varepsilon_{it},$$
$$\ln Wage_{it} = \gamma_1 + \gamma_2 Ind_{it} + \gamma_3 Ed_i + \gamma_3 Union_{it} + \gamma_4 Fem_i + \gamma_5 SMSA_{it} + u_i,$$

using 2SLS. In generic form, the equations are

$$y = \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda + \varepsilon,$$
$$x_2 = \mathbf{z}'\boldsymbol{\gamma} + u.$$

The control function estimator is always identical to 2SLS. They use exactly the same information contained in the moments and the two conditions, relevance and exogeneity. If we add to this system an assumption that (ε, u) have a bivariate normal density, then we can construct another estimator, the limited information maximum likelihood estimator. The estimator is formed from the joint density of the two variables, $(y, x_2|\mathbf{x}_1, \mathbf{z})$. We can write this as $f(\varepsilon, u|\mathbf{x}_1, \mathbf{z})\text{abs}|\mathbf{J}|$ where $\mathbf{J}$ is the Jacobian of the transformation from (ε, u) to (y, x_2),[7] $\text{abs}|\mathbf{J}| = 1$, $\varepsilon = (y - \mathbf{x}_1'\boldsymbol{\beta} + x_2\lambda)$, and $u = (x_2 - \mathbf{z}'\boldsymbol{\gamma})$. The joint normal distribution with correlation ρ can be written $f(\varepsilon, u|\mathbf{x}_1, \mathbf{z}) = f(\varepsilon|u, \mathbf{x}_1, \mathbf{z})f(u|\mathbf{x}_1, \mathbf{z})$, where $u \sim N[0, \sigma_u^2]$ and $\varepsilon|u \sim N[(\rho\sigma_\varepsilon/\sigma_u)u, (1 - \rho^2)\sigma_\varepsilon^2]$. (See Appendix B.9.) For convenience, write the second of these as $N[\tau\, u, \sigma_w^2]$. Then, the log of the joint density for an observation in the sample will be

$$\ln f_i = \ln f(\varepsilon_i|u_i) + \ln f(u_i) = -(1/2)\ln \sigma_w^2 - (1/2)\{[y_i - \mathbf{x}_1'\boldsymbol{\beta} - x_{2i}\lambda - \tau(x_{2i} - \mathbf{z}_i'\gamma)]/\sigma_w\}^2 \quad \textbf{(8-17)}$$

$$- (1/2)\ln \sigma_u^2 - (1/2)\{[x_{2i} - \mathbf{z}_i'\gamma]/\sigma_u\}^2.$$

Note, this suggests a two step estimator: (1) Estimate $[\boldsymbol{\gamma}, \sigma_u]$ by LS regression of $\mathbf{x}_2$ on $\mathbf{Z}$ then compute $\hat{\mathbf{u}} = (\mathbf{x}_2 - \mathbf{Z}\hat{\boldsymbol{\gamma}})/\hat{\sigma}_u$. (2) Estimate $[\boldsymbol{\beta}, \lambda, \tau]$ by regression of y on $(\mathbf{X}, \mathbf{x}_2, \hat{\mathbf{u}})$. This would be consistent, but would not be the same as 2SLS.

TABLE 8.2 Estimated Labor Supply Equation

	2SLS		LIML	
Variable	*Estimated Parameter*	*Standard Error*[a]	*Estimated Parameter*	*Standard Error*[a]
Constant	30.7044	8.25041	30.6392	5.05118
ln *Wage*	3.15182	1.41058	3.16303	0.87325
Education	−0.31997	0.11453	−0.32074	0.06755
Union	−2.19398	0.30507	−2.19490	0.19697
Female	−0.23784	0.79781	−0.23269	0.46572
σ_w	5.01870[b]		5.01865	0.03339
Constant			5.71303	0.03316
Ind			0.08364	0.01284
Education			0.06560	0.00232
Union			0.05853	0.01448
Female			−0.46930	0.02158
SMSA			0.18225	0.01289
σ_u			0.38408	0.00384
τ			−2.57121	0.90334

[a] Standard errors are clustered at the individual level using (8-8c).

[b] Based on mean squared residual.

Endogeneity Test? (Hausman)

	Exogenous	Endogenous
OLS	Consistent, Efficient	Inconsistent
2SLS	Consistent, Inefficient	Consistent

Base a test on $\mathbf{d} = \mathbf{b}_{2SLS} - \mathbf{b}_{OLS}$

Use a Wald statistic, $\mathbf{d}'[\text{Var}(\mathbf{d})]^{-1}\mathbf{d}$

What to use for the variance matrix?

Hausman: $\mathbf{V}_{2SLS} - \mathbf{V}_{OLS}$

Ordinary least squares regression ....

WKS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	44.7665***	1.21528	36.84	.0000	42.3846	47.1484
LWAGE	.73260***	.19718	3.72	.0002	.34614	1.11906
ED	-.15318***	.03206	-4.78	.0000	-.21601	-.09034
UNION	-1.99604***	.17006	-11.74	.0000	-2.32935	-1.66273
FEM	-1.34978***	.26417	-5.11	.0000	-1.86755	-.83200

Two stage least squares regression ....

Instrumental Variables:

ONE IND ED UNION FEM SMSA

WKS	Coefficient	Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Constant	30.7044***	4.99966	6.14	.0000	20.9052	40.5035
LWAGE	3.15182***	.85722	3.68	.0002	1.47171	4.83193
ED	-.31997***	.06607	-4.84	.0000	-.44947	-.19048
UNION	-2.19398***	.18596	-11.80	.0000	-2.55845	-1.82951
FEM	-.23784	.46793	-.51	.6113	-1.15497	.67929

```
Hausman Test
Name     ; X = one,lwage,ed,union,fem$
Name     ; Z = one,ind,ed,union,fem,smsa$
Regress ; Lhs = wks ; Rhs = X $
Matrix   ; Bols=b ; Vols=varb $
Calc     ; s2 = ssqrd $
2sls     ; Lhs = wks ; Rhs = X ; Inst = Z $
Matrix   ; b2sls = b ; v2sls = {s2/ssqrd}*varb $
Matrix   ; db = b2sls - bols ; dv = v2sls - vols $
Matrix   ; List ; Htest = db' * <dv> * db $
Matrix   ; List ; root(dv)$ (DV is singular. Rank=1)
Matrix   ; List ; Htest = db' * ginv(dv) * db $
```

```
|-> Matrix  ; List
   ; Htest = db' * <dv> * db $
---------+--------------
   HTEST|             1
---------+--------------
        1|       11.7918

|-> Matrix  ; List ; root(dv)$
---------+--------------
  RESULT|             1
---------+--------------
        1|       23.4963
        2|   -.451028E-16
        3|   -.222045E-15
        4|   -.333067E-15
        5|   -.763278E-15

|-> Matrix  ; List
   ; Htest = db' * ginv(dv) * db $
---------+--------------
   HTEST|             1
---------+--------------
        1|       11.7918
```

> 3.84

The matrix is not positive definite. It has a negative characteristic root. The matrix is indefinite. (Software such as Stata and NLOGIT find this problem and either use a generalized inverse or refuse to proceed.)

(Rank is not obvious by inspection.)

Matrix - DV

[5, 5] Cell: 22.6757

	1	2	3	4	5
1	22.6757	-3.90108	0.268964	0.319184	-1.79304
2	-3.90108	0.671132	-0.0462719	-0.0549116	0.30847
3	0.268964	-0.0462719	0.00319027	0.00378594	-0.0212678
4	0.319184	-0.0549116	0.00378594	0.00449284	-0.0252388
5	-1.79304	0.30847	-0.0212678	-0.0252388	0.141781

Hausman Test: One coefficient at a Time? No, use the full vector.

```
---- Coefficients ----
             |      (b)          (B)          (b-B)     sqrt(diag(V_b-V_B))
             |     Prior       Current      Difference      S.E.
-------------+--------------------------------------------------------
lpop         |   .5473182     1.477494    -.9301754      .1215583
eud          |  -.2723743     .0097496    -.2821239      .0350914
emud         |  -.9780319    -1.025233     .0472016      .0050788
trend        |   .1153878     .1032162     .0121716      .001261
----------------------------------------------------------------------
b= less efficient estimates obtained previously from xtreg
B= fully efficient estimates obtained from xtreg
Test:  Ho:  difference in coefficients not systematic
chi2(  5) = (b-B)'[(V_b-V_B)^(-1)](b-B)=   167.24
Prob>chi2 =      0.0000
```

Endogeneity Test: Wu

- Considerable complication in Hausman test (text, pp. 275-276)
- Simplification: Wu test.
- Regress **y** on **X** and **X**^ estimated for the endogenous part of **X**. Then use an ordinary Wald test.

Wu Test

```
|-> regr;lhs=wks;rhs=x,lwageh$

----------------------------------------------------------------------------
Ordinary     least squares regression ............
LHS=WKS      Mean                    =        46.81152
             Standard deviation      =         5.12910
-----------  No. of observations     =            4165  DegFreedom   Mean square
Regression   Sum of Squares          =         4640.01           5     928.00292
Residual     Sum of Squares          =         104905.        4159      25.22362
Total        Sum of Squares          =         109545.        4164      26.30765
-----------  Standard error of e     =         5.02231  Root MSE         5.01869
Fit          R-squared               =          .04236  R-bar squared     .04121
Model test   F[  5,  4159]           =        36.79103  Prob F > F*       .00000
--------+-------------------------------------------------------------------
        |                  Standard            Prob.      95% Confidence
     WKS|  Coefficient       Error       z    |z|>Z*         Interval
--------+-------------------------------------------------------------------
Constant|    30.7044***    4.90997     6.25  .0000     21.0810    40.3277
   LWAGE|     .59245***     .20262     2.92  .0035      .19533     .98958
      ED|    -.31997***     .06489    -4.93  .0000     -.44714    -.19280
   UNION|   -2.19398***     .18262   -12.01  .0000    -2.55191   -1.83604
     FEM|    -.23784        .45954     -.52  .6048    -1.13852     .66284
  LWAGEH|    2.55937***     .86588     2.96  .0031      .86227    4.25646
--------+-------------------------------------------------------------------
```