Endogeneity, Exogeneity and instrumental variables

Professor Bernard Fingleton

http://personal.strath.ac.uk/bernard.fingleton/

A typical regression model specification

$$Y_t = b_0 + b_1 X_{1t} + b_2 X_{2t} + b_3 X_{3t} + e_t$$

$Y_t =$ dependent variable

$X_{1t} =$ independent variable 1

$X_{2t} =$ independent variable 2

$X_{3t} =$ independent variable 3

$e_t =$ error term

Exogeneity failure

- Exogeneity means that each *X* variable does not depend on the dependent variable *Y*, rather *Y* depends on the *X*s and on *e*
- Since *Y* depends on *e*, this means that the *X*s are assumed to be independent of *Y* hence *e*
- It is a standard assumption we make in regression analysis
- required because if the 'independent variables' are not independent of *e* and *Y*, then the estimated regression coefficients are not consistent if we use the OLS estimating equations

$\hat{Y} = \hat{b}_0 + \hat{b}_1 X_1 + \hat{b}_2 X_2 .... + \hat{b}_{k-1} X_{k-1}$

$\hat{b}$ is an unbiased estimator of b if $\mathrm{E}(\hat{b}) = b$

$\hat{b}$ is a consistent estimator of b if $\hat{b} \xrightarrow{p} b$

this means that as the sample size T increases then

the probability approaches 1 that $\hat{b}$ lies

within the range $b + c$ to $b - c$

where c is a small constant > 0

the small p stands for 'converges in probability' to b

as T goes to infinity

Bias versus inconsistency

$\hat{b}$ is an unbiased estimator of b if $\mathrm{E}(\hat{b}) = b$

$\hat{b}$ is a biased estimator of b if $\mathrm{E}(\hat{b}) \neq b$

$\hat{b}$ is a consistent estimator of b if $\hat{b} \xrightarrow{p} b$

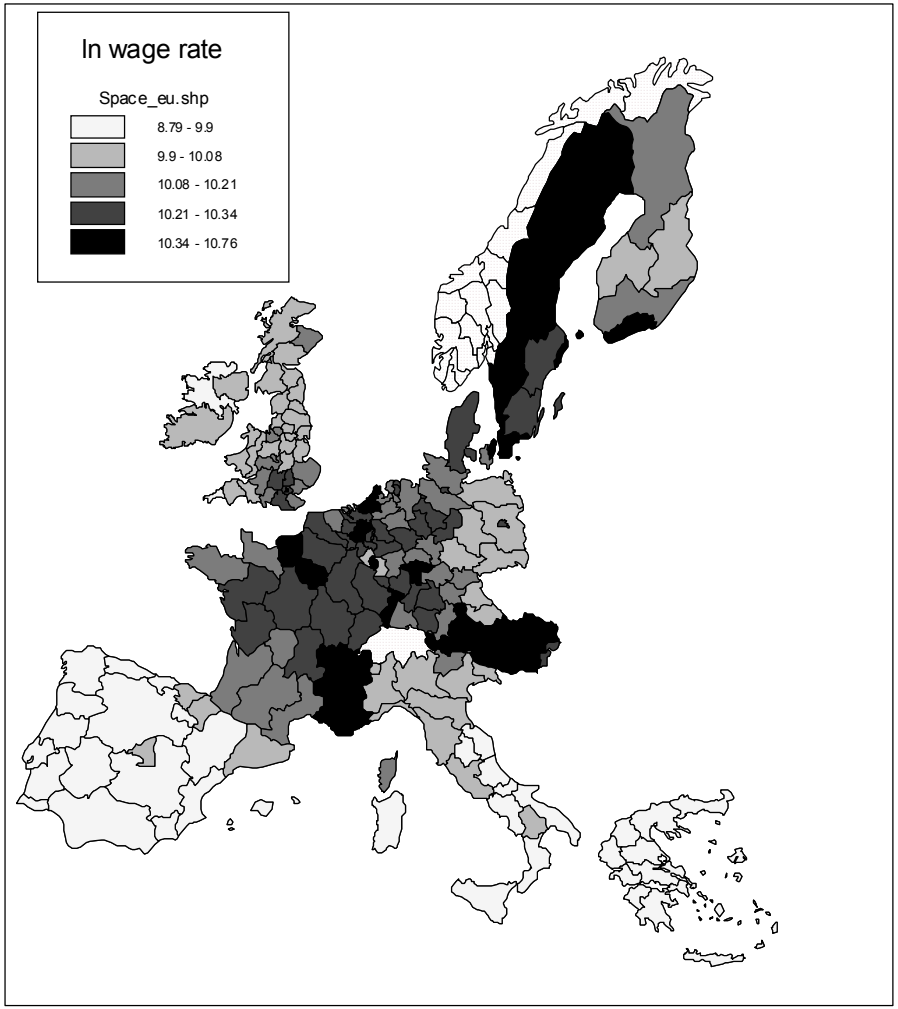
A typical biased estimator is the OLS estimator of b_1 which is the coefficient of Y_{t-1} in the autoregressive model

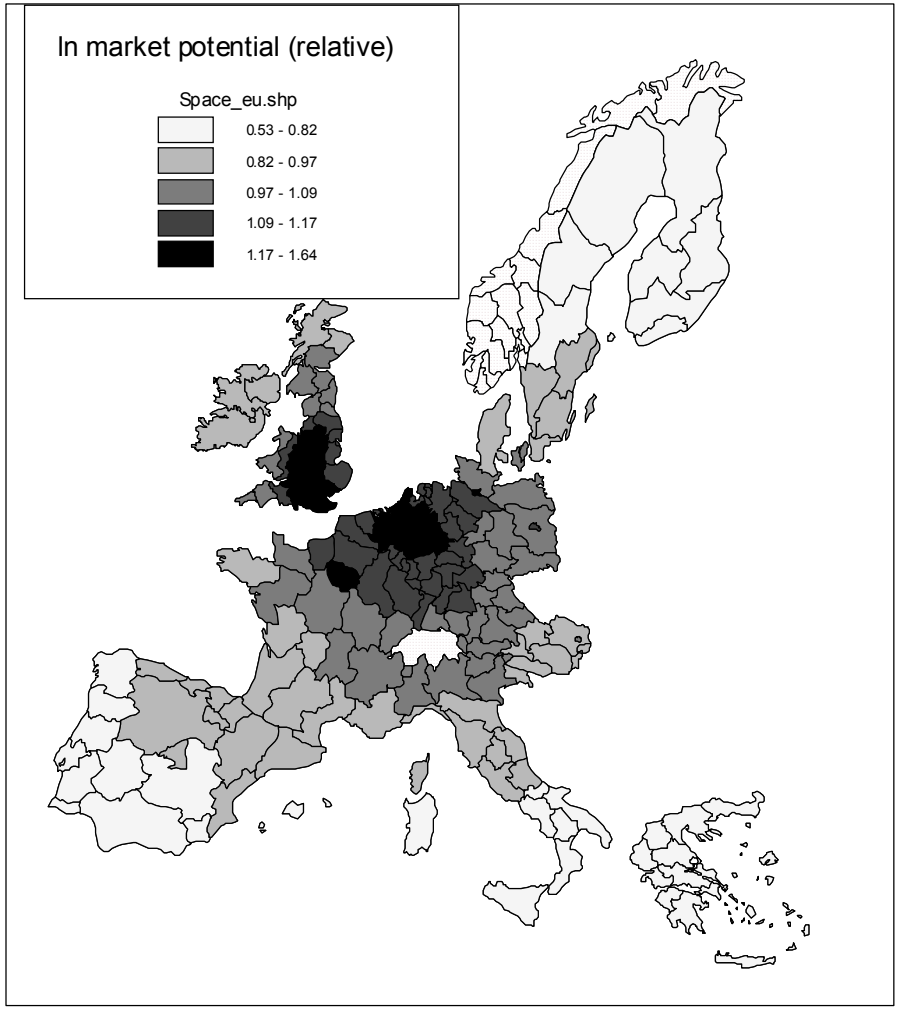
$$Y_t = b_0 + b_1 Y_{t-1} + b_2 X_{2t} + b_3 X_{3t} + e_t$$

Happily OLS can be biased and yet consistent, as with this autoregressive model, although For this to occur for the autoregressive model there is another condition we shall come to later

An empirical example

- Cross-sectional model


ln wage rate
Space_eu.shp
8.79 - 9.9
9.9 - 10.08
10.08 - 10.21
10.21 - 10.34
10.34 - 10.76


ln market potential (relative)
Space_eu.shp
0.53 - 0.82
0.82 - 0.97
0.97 - 1.09
1.09 - 1.17
1.17 - 1.64

What is market potential?

- Intuitively, it is the access to supply and demand at a particular location i.
- It depends on the on the level of income and prices in each area i,j,k,l,m….
- However remoter areas (eg m) add less to the market potential of location i because of transport costs between m and i.
- Where market potential is high, workers can bid up wage rates reflecting the advantages to producers in high market potential locations

Dependent variable *Y* = log(GVApw)

```
Model 2: OLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw

             coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -2.51682      1.19136     -2.113    0.0356    **
  lnMP         1.28870      0.117013    11.01     2.66E-023 ***
```

In general : 4 main reasons why *X* and *e* might be correlated

1. Simultaneous equations bias
2. Omitted variables bias
3. Regression model (time series) includes a lagged dependent variable and the error term is serially correlated.
   - Recall that estimate biased but consistent with a lagged dependent variable, but this assumes that the errors are independent of each other over time
4. Errors-in-variables
   - This is when we cannot measure the true *X* variable, so that there is uncertainty attached to the measured value

Simple linear regression model

$$Y_i = b_0 + b_1 X_i + e_i$$

$$i = 1, \ldots, N$$

or with time series, $Y_t = b_0 + b_1 X_t + e_t$ and $t = 1, \ldots, T$

- Data either time series or cross section
- *X* is exogenous if Corr(*X*, *e*) = 0
- *X* is endogenous if Corr(*X*, *e*) $\neq$ 0

- If OLS is to be unbiased and consistent, requires that *X* is exogenous.

Simple linear regression model

$$Y_i = b_0 + b_1 X_i + e_i$$

$$i = 1, \ldots, N$$

- If *X* is not exogenous (endogenous), i.e. Corr(*X*, *e*) $\neq$ 0
- then OLS is biased even in large samples and so is not consistent
- In this case IV(2sls) can produce consistent estimates

Consistency of OLS

$$Y_i = b_0 + b_1 X_i + e_i$$

$$\hat{b}_1 = \frac{S_{XY}}{S_{XX}} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)}{\sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2} = \frac{Cov\left(X_i, Y_i \right)}{Var\left(X_i \right)}$$

Consistency of OLS

$$Y_i = b_0 + b_1 X_i + e_i$$

$$\hat{b}_1 = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right)}{\sum_{i=1}^{N} \left(X_i - \bar{X} \right)^2} = \frac{Cov\left(X_i, Y_i \right)}{Var(X_i)}$$

$$= \frac{Cov\left(X_i, \left[b_0 + b_1 X_i + e_i \right] \right)}{Var\left(X_i \right)}$$

$$\hat{b}_1 = b_1 + \frac{Cov\left(X_i, e_i \right)}{Var\left(X_i \right)}$$

Derivation of the last step on the first slide

$$\frac{Cov\left(X_i, \left[b_0 + b_1 X_i + e_i\right]\right)}{Var\left(X_i\right)}$$

$$= \frac{Cov\left(X_i, b_0\right) + Cov\left(X_i, b_1 X_i\right) + Cov\left(X_i, e_i\right)}{Var\left(X_i\right)}$$

$$= \frac{0 + b_1 Var\left(X_i\right) + Cov\left(X_i, e_i\right)}{Var\left(X_i\right)}$$

$$= \frac{0}{Var\left(X_i\right)} + \frac{b_1 Var\left(X_i\right)}{Var\left(X_i\right)} + \frac{Cov\left(X_i, e_i\right)}{Var\left(X_i\right)}$$

$$= b_1 + \frac{Cov\left(X_i, e_i\right)}{Var\left(X_i\right)}$$

Consistency of OLS

Looking at this equation:

$$\hat{b}_1 = b_1 + \frac{Cov(X_i, e_i)}{Var(X_i)}$$

it is clear that for OLS to be consistent, $Cov(X_i, e_i)$ must be zero.

Inconsistency of OLS

$$\hat{b}_1 = b_1 + \frac{Cov(X_i, e_i)}{Var(X_i)}$$

$$Cov(X_i,\ e_i)\ \neq\ 0$$

Recall there are 4 main reasons why *X* and *e* might be correlated.

1) Simultaneous equations bias

$$Y_t = b_1 X_t + e_t \qquad (1)$$

$$X_t = \gamma_1 Y_t + v_t \qquad (2)$$

$$Y_t = \frac{(b_1 v_t) + e_t}{1 - (b_1 \gamma_1)} \qquad (3)$$

$$X_t = \frac{(\gamma_1 e_t) + v_t}{1 - (b_1 \gamma_1)} \qquad (4)$$

hence *X* is correlated with *e*

Simultaneous equations bias (Z and W exogenous variables)

$$Y_t = b_1 X_t + b_2 Z_t + e_t \qquad (1)$$

$$X_t = \gamma_1 Y_t + \gamma_2 W_t + v_t \qquad (2)$$

$$Y_t = \frac{(b_1\gamma_2)W_t}{1-(b_1\gamma_1)} + \frac{(b_2)Z_t}{1-(b_1\gamma_1)} + \frac{(b_1 v_t)+e_t}{1-(b_1\gamma_1)} \qquad (3)$$

$$X_t = \frac{(b_2\gamma_1)Z_t}{1-(b_1\gamma_1)} + \frac{(\gamma_2)W_t}{1-(b_1\gamma_1)} + \frac{(\gamma_1 e_t)+v_t}{1-(b_1\gamma_1)} \qquad (4)$$

hence *X* is correlated with *e*

2) Omitted variable bias

$$Y_t = b_1 X_t + b_2 W_t + e_t \qquad (True)$$

$$Y_t = b_1 X_t + (b_2 W_t + e_t)$$

$$Y_t = b_1 X_t + v_t \qquad \text{(We estimate)}$$

If $Corr(X, W) \neq 0$ then $Cov(X, v) \neq 0$

3) Lagged dependent variable

The model includes a lagged dependent variable AND has a serially correlated disturbance

Suppose we estimate the model

$$Y_t = b_0 + b_1 X_t + b_2 Y_{t-1} + e_t$$

with serially correlated errors

$$e_t = \rho e_{t-1} + u_t \text{ with } \rho \neq 0$$

Clearly $Y_{t-1} = b_0 + b_1 X_{t-1} + b_2 Y_{t-2} + e_{t-1}$

Y_{t-1} is correlated with e_{t-1}

e_{t-1} determines e_t since $e_t = \rho e_{t-1} + u_t$

therefore Y_{t-1} is correlated with e_t

this correlation will remain as T increases

4) Errors in variables

Suppose X_i is measured imprecisely by $\tilde{X}_i$ but we want to estimate the true relationship $Y_i = b_0 + b_1 X_i + e_i$

In fact using $\tilde{X}_i$ the true relationship becomes

$$Y_i = b_0 + b_1 \tilde{X}_i + [b_1 (X_i - \tilde{X}_i) + e_i]$$

since $b_1 \tilde{X}_i - b_1 \tilde{X}_i = 0$

Suppose we estimate

$$Y_i = b_0 + b_1 \tilde{X}_i + \nu_i$$

The error term $\nu_i = b_1 (X_i - \tilde{X}_i) + e_i$ contains the difference $(X_i - \tilde{X}_i)$

If $\text{corr}(\tilde{X}_i, (X_i - \tilde{X}_i)) \neq 0$ then OLS estimator $\hat{b}_1$ from $Y_i = b_0 + b_1 \tilde{X}_i + \nu_i$ is a biassed and inconsistent estimator of the true b_1 in $Y_i = b_0 + b_1 X_i + e_i$

Solving the problem

- All 4 sources of endogeneity lead to inconsistent OLS estimation
- Ideally we should eliminate measurement error, introduce omitted variables, estimate a system of simultaneous equations etc.
- Often these solutions are not achievable in practice, thus……
- The solution is to use an alternative estimation method known as instrumental variables (IV) or equivalently two-stage least squares (2sls)
- this involves replacing the endogenous variable *X* (which is correlated with the error term) by a 'proxy' variable. To do this we make use of variable *Z*, known as an instrumental variable, that is independent of the error term.

Two conditions for a valid instrument

- **Instrument relevance**: $\text{corr}(Z_i, X_i) \neq 0$
- **Instrument exogeneity**: $\text{corr}(Z_i, e_i) = 0$

- Suppose for now that you have such a Z_i (we'll discuss how to find instrumental variables later).
- How can you use Z_i to estimate b_1 consistently?

Explanation 1: Two Stage Least Squares (TSLS)

- **Stage 1: Isolate the part of *X* that is uncorrelated with *e***

We do this by regressing X on Z using OLS

$$X_i = \pi_0 + \pi_1 Z_i + v_i \qquad (1)$$

because Z_i is uncorrelated with e_i

$\pi_0 + \pi_1 Z_i$ is uncorrelated with e_i

we don't know π_0 or π_1 but we have estimated them

so as to obtain the predicted values of X

$$\hat{X}_i = \hat{\pi}_0 + \hat{\pi}_1 Z_i$$

Stage 2: Replace X_i by the predicted values of X_i in the regression of interest

Next regress Y on $\hat{X}$ (the predicted X from the first stage regression)

$$Y = b_0 + b_1 \hat{X} + e \qquad (2)$$

because $\hat{X}$ is uncorrelated with e in large samples

then b_1 can be estimated consistently by OLS using

this second stage regression

IV or 2sls Estimator

This argument relies on large samples

so that π_0, π_1 are well estimated using regression (1)

The resulting estimator is called the two-stage least squares

(2sls or TSLS) estimator

2sls is a consistent estimator of b_1

Recall that an estimator is consistent if the probability that it is in error by more than a given amount
tends to zero as the sample become larger.

Two-stage least squares summary	
Preliminaries: Seek out an appropriate instrument Z Generally this is not easy because 1)It has to be exogenous, that is uncorrelated with the error term 2) It has to be relevant to the endogenous variable	
Stage 1:	Regress X_i on Z_i using OLS to obtain predicted values $\hat{X}_i$
Stage 2:	Using OLS, regress Y_i on $\hat{X}_i$; the estimated coefficient on $\hat{X}_i$ is the 2sls estimator of b_1
Postscript: Generally we want more than one instrument, so as to improve the prediction $\hat{X}_i$ Also, there may be more than one endogenous variable, e.g. $X_{1i}, X_{2i} \ldots$	

Inference using TSLS

- Statistical inference proceeds in the usual way.
- The justification is (as usual) based on large samples
- In large samples, the sampling distribution of the IV/TSLS estimator is normal.
- Inference (hypothesis tests, confidence intervals) proceeds in the usual way, e.g. estimated coefficient value ± 1.96SE
- This all assumes that the instruments are valid
- Note however that the standard errors from the second-stage OLS regression are not valid, because they do not take account of the fact that the first stage is also estimated
- So it is necessary to use a dedicated regression package that carries out 2sls with correct standard errors and hence t-ratios, rather than do two separate OLS regressions manually (see Stock and Watson, 2007, p.429 for details)

An example: the wage equation from NEG theory

- Dependent variable *Y*
  - *Y* = log(GVApw)
  - 255 values, one for each NUTS 2 EU region across 25 countries
- One endogenous regressor *X*
  - *X* = *ln MP*
    - Suggested by theory
- Other variables
  - *W* = new entrants
    - a dummy variable = 1 when a region is in a 'new entrant' country, 0 otherwise
    - Wages lower in new entrant countries due to legacy of inefficiency under command economy, different institutions etc
  - Z_1 =ln area of region in sq. km = ln(sqkm)
  - Z_2 =weighted average of ln of areas of surrounding regions in sq. km = Wa(ln(sqkm))
  - Z_3 = weighted average of new entrants in surrounding regions = Wa(new entrants)

MPexample.xls

lnGVApw	constant	new_entrant	lnMP	WA(new_entrant)	ln_sqkm	WA(ln_sqkm)	ln_empdens
10.82691	1	0	10.01332564	0.6	8.28538723	9.288957551	3.195355895
10.80377	1	0	9.993008009	0.5	9.861508639	8.746668059	3.501206181
11.12049	1	0	10.83037023	0	6.027555367	9.861508639	7.611072417
10.81516	1	0	10.04351981	0.166666667	9.162829389	9.455268725	3.220321926
10.77102	1	0	9.988936451	0.166666667	9.704542589	9.248893521	3.485632696
10.86415	1	0	10.03403943	0.166666667	9.391135765	9.53876825	3.951149982
10.88408	1	0	10.08467595	0	8.875454876	9.498954035	3.635665679
10.87223	1	0	10.09963093	0	9.445230659	9.196703992	3.278343278
10.95135	1	0	10.33777588	0	7.863843481	9.335565816	4.154669915
11.24226	1	0	11.13331466	0	5.081404365	7.652545693	8.25928136
11.13279	1	0	10.48334659	0	7.961021466	7.992550559	5.452423698
10.94566	1	0	10.4319876	0	7.792348924	8.021262834	4.75138097
11.00476	1	0	10.44130762	0	8.000349495	7.977994304	5.084874681
11.21462	1	0	10.50919595	0	7.652545693	7.475425719	5.074627607
10.92203	1	0	10.41779631	0	8.053251154	8.412520712	4.972138661
11.1821	1	0	10.56125076	0	6.994849986	8.089351979	4.617395789
10.89881	1	0	10.34747743	0	8.239065332	8.54474896	4.578187645
10.91588	1	0	10.38790472	0	8.258940463	8.001029897	4.439704219
10.74328	1	0	10.29706175	0	8.398409655	8.908157554	3.010121494
10.8725	1	0	10.29169784	0	8.206856428	8.408369483	3.652823939
11.32281	1	0	10.24961326	0	9.073213954	9.168187316	4.221240494
10.97004	1	0	10.25499486	0	9.216541108	8.755119559	4.467090406
11.23771	1	0	10.53475477	0	7.579780963	8.649074934	5.650367189
11.19267	1	0	10.615395	0	7.455240647	8.620925591	6.162758774
11.22232	1	0	10.22702244	0	9.351926736	8.732056987	3.672969337
11.22785	1	0	10.33344285	0	8.408114661	8.436404297	4.295062908
11.08884	1	0	10.27735402	0	7.941722374	9.411184592	3.85652488
10.12844	1	1	10.19094851	1	6.206374293	9.307113118	7.229593257
9.742717	1	1	9.990118241	1	9.307113118	8.804614869	3.720328404
9.571378	1	1	10.00058476	0.428571429	9.776659357	9.370406696	3.424674153

OLS vs TSLS

```
Model 1: OLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw

            coefficient   std. error   t-ratio    p-value
  ---------------------------------------------------------
  const       -2.51682      1.19136     -2.113    0.0356    **
  lnMP         1.28870      0.117013    11.01     2.66E-023 ***


Model 2: TSLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw
Instruments: ln_sqkm const

            coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------
  const       3.69262       1.61533      2.286    0.0223   **
  lnMP        0.678655      0.158671     4.277    1.89E-05 ***

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 49.5432
  with p-value = 1.94052e-012

First-stage F-statistic (1, 253) = 383.11
  A value < 10 may indicate weak instruments
```

Reasons why *X* and *e* might be correlated

- Omitted variables bias
  - *New Entrants* have low ln*MP*, so
  - corr(*New Entrants* ,ln*MP*) < 0
  - Since *New Entrants is* in *e*, corr(*e* ,ln*MP*) not 0
- Simultaneous equations bias
  - Market potential (ln*MP*) depends on wages as well as determines them

Why is *ln MP* endogenous?

NEG (new economic geography) theory gives a set on nonlinear simultaneous equations

involving wage rates w_i^M and market potential *MP*

wage rates depend on *MP*

but *MP* is partially determined by wage rates

in theory $\ln w_i^M = b_1 \ln MP$

$$b_1 = \frac{1}{\sigma}$$

$$w_i^M = [MP]^{\frac{1}{\sigma}}$$

$$w_i^M = [\sum_r Y_r (G_r^M)^{\sigma-1} T_{Mir}{}^{1-\sigma}]^{\frac{1}{\sigma}}$$

$$G_i^M = [\sum_r \lambda_r (w_r^M T_{Mir})^{1-\sigma}]^{\frac{1}{1-\sigma}}$$

Adding omitted variable to the model

```
Model 3: TSLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw
Instruments: ln_sqkm new_entrant const

               coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const           7.72764     0.868184       8.901   5.54E-019 ***
  lnMP            0.300959    0.0848567      3.547   0.0004    ***
  new_entrant    -1.24618     0.0487433    -25.57    3.63E-144 ***

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 1.49897
  with p-value = 0.220829

First-stage F-statistic (1, 252) = 504.878
  A value < 10 may indicate weak instruments
```

Endogenous MP?

- Assuming the variable new_entrant is exogenous, adding it to the model now means that OLS is now acceptable, as indicated by the Hausman test
- But there is a theoretical reason why MP is endogenous, because by definition it depends on the dependent variable, so we prefer to instrument it anyway
- Also is new_entrant exogenous?
- Also the results may differ with different /more instruments
- And we may also prefer to use > 1 instrument since then we can also test the validity of the instruments via the Sargan overidentification test

Some instruments

- Z_1 =ln area of region in sq. km = ln(sqkm)
  - Sqkm is fixed, it is the area of the region and will not change in response to wage rates, or as a result of taking logs
  - Regions with smaller areas are cities, which are concentrations of economic activity with high market potential

- Z_2 =weighted average of log of areas of surrounding regions in sq. km = Wa(ln(sqkm))
  - Likewise, we do not alter the exogeneity by taking the weighted mean of ln(sqkm)
  - Having ‘cities’ nearby will add to an areas market potential

Some instruments

- Z_3 = Wa(new entrants)
  - we have assumed that the dummy variable, new entrants is exogenous. It simply takes the value 1 or zero according to whether a region is in a new entry country.
  - Simply taking the weighted average of new entrants in surrounding regions = Wa(new entrants) will not change this fact
  - An area surrounded by new entrants will have lower market potential than one that is not surrounded

Why include three instruments (the *Z*s)?

- One instrument will suffice, but better prediction of the endogenous variable with more than one instrument (the coefficient is said to be overidentified in this case)
- In the case of just one instrument and one endogenous variable, 2sls will work, we have in this case exact identification.
- but if we were to introduce a second endogenous variable, then one instrument is not enough because the coefficient to be estimated is underidentified

Identification

The coefficients $b_1, \ldots, b_k$ are said to be:

- **exactly identified** if m = k. (There are just enough instruments to estimate $b_1, \ldots, b_k$)
- **overidentified** if m > k. There are more than enough instruments to estimate $b_1, \ldots, b_k$. If so, you can test whether the instruments are valid (a test of the "overidentifying restrictions")
- **underidentified** if m < k. There are too few enough instruments to estimate $b_1, \ldots, b_k$. If so, you need to get more instruments!

The General IV Regression Model

- Usually we have more than one rhs endogenous variable
- Usually we want to use more than one instrumental variable

The General IV Regression Model

The general IV regression model	$Y_i = b_0 + b_1 X_{1i} + ... + b_k X_{ki} + b_{k+1} W_{1i} + ... + b_{k+r} W_{ri} + e_i$
Dependent variable	Y_i
k endogenous regressors (potentially correlated with *e*)	$X_{1i}, ..., X_{ki}$
r included exogenous variables (regressors) uncorrelated with *e*	$W_{1i}, ..., W_{ri}$
m instrumental variables (or excluded exogenous regressors)	$Z_{1i}, ..., Z_{mi}$
Unknown regression coefficients	$b_0, b_1, ..., b_{k+r}$

tsls with overidentification, one endogenous *X*, one or more *W* variable

- The 2sls method is the ‘same’ as before
- in stage 1 regress the endogenous variable *X* on all the exogenous variables (*W*s) and all the instruments (*Z*s),
- in stage 2 regress *Y* on the exogenous (*W*) variables and the fitted values from stage 1.

Preliminaries: Check that X_i is correlated with e_i (Hausman test, see later) Seek out m appropriate instruments $Z_1,...,Z_m$ So that 1) they are exogenous, that is uncorrelated with the error term (Sargan test, see later) 2) they are correlated with the endogenous variable	
Stage 1:	Regress X_i on $W_1,...,W_r,Z_1,...,Z_m$ using OLS to obtain predicted values $\hat{X}_i$
Stage 2:	Using OLS, regress Y_i on $\hat{X}_i,W_1,...,W_r$; the estimated coefficient on $\hat{X}_i$ is the 2sls estimator of b_1

Gretl output

```
Model 3: TSLS, using observations 1-255
Dependent variable: lnGVApw
Instrumented: lnMP
Instruments: ln_sqkm WA_ln_sqkm_ WA_new_entrant_ new_entrant const

                coefficient   std. error      z       p-value
  ------------------------------------------------------------
  const           7.48301     0.842887       8.878   6.82e-019 ***
  lnMP            0.324873    0.0823834      3.943   8.03e-05  ***
  new_entrant    -1.23822     0.0482409    -25.67    2.70e-145 ***

Mean dependent var   10.60041   S.D. dependent var   0.541194
Sum squared resid    13.12181   S.E. of regression   0.228190
R-squared            0.823621   Adjusted R-squared   0.822222
F(2, 252)            582.7264   P-value(F)           3.06e-95

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(1) = 0.639271
  with p-value = 0.423975

Sargan over-identification test -
  Null hypothesis: all instruments are valid
  Test statistic: LM = 5.92972
  with p-value = P(Chi-Square(2) > 5.92972) = 0.0515677
```

Gretl output

```
Weak instrument test -
  First-stage F-statistic (3, 250) = 200.482
  Critical values for TSLS bias relative to OLS:

     bias        5%       10%       20%       30%
    value     13.91      9.08      6.46      5.39

  Relative bias is probably less than 5%

  Critical values for desired TSLS maximal size, when running
  tests at a nominal 5% significance level:

     size       10%       15%       20%       25%
    value     22.30     12.83      9.54      7.80

  Maximal size is probably less than 10%
```

Critical value for F is 13.91, 200.482>13.91 so TSLS estimator bias < 5% of OLS bias

Also 200.482>22.30 so tests of significance of individual variables have 'size' of < 10% (S&W p.79)
This means that we have a less than 10% chance of wrongly 'accepting' a variable as significant using the standard rules (t > 2 roughly, so nominal size = 5%)

interpretation

- Sargan test suggests (marginally) that all instruments are not valid, perhaps new_entrant is endogenous
- Weak instruments can lead to serious problems in IV regression: biased estimates and/or incorrect size of hypothesis tests, with rejection rates well in excess of the nominal significance level

2sls with > 1 endogenous *X* variable

- Consider next that whether or not a country is a new entrant depends on its GVA per worker
- Then we have 2 endogenous variables. lnMP, new_entrant
- The 2 stages are as before but
- Take care that there are enough *Z* variables so as to avoid under-identification.
- So we add an additional exogenous variable (ln_empdens) to make 3 instruments for our 2 endogenous variables
- Now we have overidentification and can test for the validity of the instruments via the Sargan test

Gretl output

```
Model 5: TSLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw
Instruments: ln_sqkm WA_ln_sqkm_ ln_empdens const

                coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const           7.74865      1.10673      7.001    2.53E-012 ***
  new_entrant    -1.21021      0.327567    -3.695    0.0002    ***
  lnMP            0.298355     0.105019     2.841    0.0045    ***

Hausman test -
  Null hypothesis: OLS estimates are consistent
  Asymptotic test statistic: Chi-square(2) = 5.87046
  with p-value = 0.0531184

Sargan over-identification test -
  Null hypothesis: all instruments are valid
  Test statistic: LM = 0.318418
  with p-value = P(Chi-Square(1) > 0.318418) = 0.57256
```

Testing with 2 endogenous variables

- Hausman test is borderline, indicating that possibly we would have inconsistency if we used OLS and did not use instruments
- Sargan test indicates that the instruments are valid
- BUT the weak instrument test suggests that the size of tests on coefficients, nominally 5% size, may actually be > 25%

Checking the validity of instruments : Sargan

- Instruments should be independent of the errors
- To test whether this is the case, we take the 2sls residuals as the dependent variable
  - 2sls residuals use the 2sls coefficient estimates and the original variables, not the instruments
- Then take the instruments (*Z*s) and the other exogenous variables (*W*s) as regressors
- For valid instruments, the *Z*s should be unrelated to the 2sls residuals
  - This assumes that the set of *W*s is correct. If not then this may cause a significant result, but in this case because the basic model is mispecified rather than invalid instruments
- Details are given in 12.3, S&W (2007)

Sargan test also called overidentifying restrictions test

- Overidentification is when we have more Instruments than endogenous variables
- On its own each instrument will give a different estimate
- But we expect valid individual instruments to give more or less the same estimates
- If they differ, that suggests ‘something is wrong with one or the other of the instruments-or both’
- To check we need different instruments, at least two when we have one endogenous variable

Checking the validity of instruments : Sargan

- They are called ‘over-identifying restrictions’

because we test the null hypothesis that, in the regression of the 2sls residuals depending on W and Zs, the coefficients on the whole set of instruments (the Zs) can be restricted to zero

  - This is what we would expect of all the instruments were valid, that is valid Zs should be unrelated to the residuals

Checking the validity of instruments : Sargan

- It only works with over-identification, the test cannot be carried out with exact identification
  - If you have exact identification, and regress the instrument(s) on the 2sls residuals, the coefficient(s) is(are) exactly zero.
  - The same thing happens if you regress an exogenous variable on OLS residuals. By definition, the residuals are independent of the regressor, so you cannot test whether this is the case
- Thus we need more *Z*s (instruments) than *X*s (endogenous variables)

Checking the validity of instruments : Sargan

```
Model 6: OLS estimates using the 255 observations 1-255
Dependent variable: tslsres

               coefficient   std. error   t-ratio    p-value
  ------------------------------------------------------------
  const          0.00959931   0.285198     0.03366   0.9732
  ln_sqkm       -0.00997415   0.0213319   -0.4676    0.6405
  WA_ln_sqkm_    0.0101892    0.0224299    0.4543    0.6500
  ln_empdens    -0.00325940   0.0176229   -0.1850    0.8534

  F-statistic (3, 251) = 0.104605 (p-value = 0.957)
```

Checking the validity of instruments : Sargan

The test statistic is $J = mF$

m is the number of instruments

F is the F statistic

Here $m = 3, F = 0.1046, J = 0.314$

This is referred to the χ^2_{m-k} distribution

k is the number of endogenous variables

$m - k$ is the degree of overidentification

equal to the number of instruments minus the number of endogenous regressors

So $J = 0.314$ has a p-value of 0.57 in χ^2_1

do not reject the null that the instruments are valid

Checking the exogeneity of variables : Hausman

- An exogenous variable does not need to be instrumented, an endogenous one does
- Sometimes theory tells us that a variable is endogenous (eg *lnMP*)
- But we can also use diagnostics to tell us whether a variable is endogenous

Checking the exogeneity of variables : Hausman

- The test, often referred to as the Wu-Hausman test, comprises 2 regressions
  - Wu(1973) is responsible for the simpler regression-based version described here
- The first takes the suspect endogenous *X* variable as the dependent variable and the *W*s and the instruments *Z* as independent variables, saving the *fitted values OR* the *residuals (*both give identical conclusions*)*

Checking the exogeneity of variables : Hausman

- The 2nd regression takes the *Y* variable as the dependent variable and *X* , *Ws* and *fitted values* (or *residuals*) as independent variables
- If the effect of *fitted values* (or equivalently *residuals*) is significant, that indicates that they carry explanatory information additional to that that already contained in *X* and *W*.
- That suggests that we get different results instrumenting *X* than simply using *X* per se as an independent variable, thus pointing to the endogeneity of *X*

Checking the exogeneity of variables : Hausman

```
ols lnMP const  ln_sqkm WA_ln_sqkm_ ln_empdens
genr fvMP = $yhat

ols new_entrant const  ln_sqkm WA_ln_sqkm_ ln_empdens
genr fv_ne = $yhat

ols lnGVApw const new_entrant lnMP fvMP fv_ne
omit fvMP fv_ne
```

Gretl output, two regressions for Wu-Hausman test

```
Model 9: OLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw

               coefficient   std. error   t-ratio     p-value
  ------------------------------------------------------------
  const          7.74865     1.09497        7.077    1.48E-011 ***
  new_entrant   -1.13052     0.0610374    -18.52     8.43E-049 ***
  lnMP           0.742390    0.173650       4.275    2.72E-05  ***
  fvMP          -0.444035    0.202362      -2.194    0.0291    **
  fv_ne         -0.0796919   0.329784      -0.2416   0.8093


Model 10: OLS estimates using the 255 observations 1-255
Dependent variable: lnGVApw

               coefficient   std. error   t-ratio    p-value
  -----------------------------------------------------------
  const           7.12069    0.708099      10.06    3.20E-020 ***
  new_entrant    -1.22644    0.0458748    -26.73    1.56E-075 ***
  lnMP            0.360292   0.0692045      5.206   4.00E-07  ***

Comparison of Model 9 and Model 10:

  Null hypothesis: the regression parameters are zero for the variables

    fvMP
    fv_ne

  Test statistic: F(2, 250) = 2.87768, with p-value = 0.0581311
```

Checking the exogeneity of variables : Hausman

- This reaffirms that there might be some indication (say at the 10% significance level) that the two variables MP and new entrants are endogenous
- The results obtained by this regression approach are (almost) identical to the output for the Hausman test given by Gretl

Gretl code

```
open
C:\dad\courses\Strathclyde\MSc_appliedEconometrics\week4\MPexample.gdt

ols lnGVApw const lnMP

#exact identification
tsls lnGVApw const lnMP ; ln_sqkm const

tsls lnGVApw const lnMP new_entrant ; ln_sqkm   new_entrant const
#over identification
tsls lnGVApw const new_entrant lnMP ; ln_sqkm WA_ln_sqkm_ \
  WA_new_entrant_ new_entrant const

# with > 1 endogenous variable

tsls lnGVApw const new_entrant lnMP ; ln_sqkm WA_ln_sqkm_ \
  ln_empdens const
genr tslsres = $uhat

# Sargan manual version

ols tslsres const ln_sqkm WA_ln_sqkm_ ln_empdens
```

Gretl code

```
# Wu-Hausman test of exogeneity of variables

ols lnMP const  ln_sqkm WA_ln_sqkm_ ln_empdens
genr fvMP = $yhat

ols new_entrant const  ln_sqkm WA_ln_sqkm_ ln_empdens
genr fv_ne = $yhat

ols lnGVApw const new_entrant lnMP fvMP fv_ne
omit fvMP fv_ne

# repeat using residuals rather than fitted values

ols lnMP const  ln_sqkm WA_ln_sqkm_ ln_empdens
genr r_MP = $uhat

ols new_entrant const  ln_sqkm WA_ln_sqkm_ ln_empdens
genr r_ne = $uhat

ols lnGVApw const new_entrant lnMP r_MP r_ne
omit r_MP r_ne
```