Lab 3: Endogeneity and Instrumental Variable Estimation in Panel Data

GECO 6281 Advanced Econometrics 1

Patrick Mokre

Fall 2019

Recapitulation

Lab 0: Introduction to the course

- What are the learning Outcomes expected for Advanced Econometrics 1?
- Which softwares are we using, what are their strengths and weaknesses?
- What is the main difference between STATA and RStudio regarding datasets?
- Which software do you use to load Google Drive files into **Apps Anywhere's** STATA 15 version?
- In which formats do we store data?

Lab 1: Panel Data

- What are the two dimensions of panel data?
- Which are the three main estimation methods we use for panel data? When are they consistent?
- How do we decide which estimation method to use?
- Why do degrees of freedom matter in statistical inference?
- How do first difference and pooled OLS estimation of a fixed effects model correspond?

Endogeneity 1

Both FE and RE models produce consistent estimators only if covariates x_{it} **are strictly exogenous**, i.e. $E(\epsilon_{it} \mid X) = E(\epsilon_{it}) = 0 \quad \forall i \in N, t \in T$. (Pesaran 2015, 635)

Consistency: $\hat{\beta} \to \beta$ for either $T \to \infty$ or $N \to \infty$. If an estimator is not consistent, it cannot be **unbiased**.

Endogeneity: There is an unobserved correlation between covariates x_{it} and residuals u_{it}. This lead to a bias in $\hat{\beta}$.

Endogeneity 2: Time-Series Example 1

Assume you have a model with:

$$y_t = \alpha + \beta_1 x_t + \epsilon_t$$

But x_t is endogenous:

$$x_t = y_t + z_t$$

The problem becomes obvious when the model is presented in **structural form**

$$x_t = \frac{\alpha}{1-\beta} + \frac{1}{1-\beta} z_t + \frac{1}{1-\beta}\epsilon_t$$
$$y_t = \frac{\alpha}{1-\beta} + \frac{\beta}{1-\beta} z_t + \frac{1}{1-\beta}\epsilon_t$$

From which it follows that:

$$cov(x_t, \epsilon_t) = \frac{1}{1-\beta} cov(z-t, \epsilon_t) + \frac{1}{1-\beta} V(\epsilon_t) = \frac{\sigma^2}{1-\beta}$$

Endogeneity 3: Biased Estimator

$$plim(\hat{\beta}) = \beta + \frac{cov(x_t, \epsilon_t)}{Vx_t}$$

$$V(x_t) = V(\frac{1}{1-\beta} z_t + \frac{1}{1-\beta} \epsilon_t) = \frac{1}{(1-\beta)^2}(V(z_t + \sigma^2))$$

$$plim(\hat{\beta}) = \beta + (1-\beta)\frac{\sigma^2}{V(z_t) + \sigma^2}$$

So for $\beta \in (0, 1)$, endogeneity produces overestimation of the effects.

Instrumental Variables

The **problem** with endogeneity is that you have a causal relationship from y_i to x_i. One possible solution is to find a **proxy** or **instrumental variable** z_i which helps explain x_i, but is not determined by y_i.

This allows for **2-step-least-suqare (2SLS)** estimation under two assumptions:

- relevance: $\frac{\partial X}{\partial Z} \neq 0$
- independence: $E((y_i - \alpha - x_i \beta) z_i) = 0$

In a 2SLS estimation, you first estimate the impact of z_i on x_i, and then the impact of z_i on y_i. Analytically, you derive the IV estimator as $\widehat{\beta_{IV}} = (\sum_i^N z_i x_i')^{-1} \sum_i^N z_i y_i$.

Caution: **forbidden regressions**: You must not apply 2SLS regressions to non-linear models, e.g. instrumentalizing a dummy variable in a PROBIT regression, since the first-stage residuals might be correlated with the second-stage fitted values and covariates. (Angrist and Prischke 2009, 190f)

2SLS in STATA

In STATA you use the ivregress 2sls command and assign instrumented as well as instrument variables in parentheses. The example from STATA help is intuitive, where you want to estimate the impact of housing value on rents. In orthodox economic theory, the value of an asset can be derived from the income one receives from it, i.e. $E((y_i - \beta x_i)x_i) \neq 0$

```
use http://www.stata-press.com/data/r13/hsng, clear

. ivregress 2sls ren pcturban (hsngval=faminc i.region)

Instrumental variables (2SLS) regression          Number of obs   =         50
                                                  Wald chi2(2)    =      90.76
                                                  Prob > chi2     =     0.0000
                                                  R-squared       =     0.5989
                                                  Root MSE        =     22.166

------------------------------------------------------------------------------
        rent |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hsngval |   .0022398   .0003284     6.82   0.000     .0015961    .0028836
    pcturban |    .081516   .2987652     0.27   0.785    -.504053     .667085
       _cons |   120.7065   15.22839     7.93   0.000     90.85942    150.5536
------------------------------------------------------------------------------
Instrumented:  hsngval
Instruments:   pcturban faminc 2.region 3.region 4.region
```

2SLS in STATA 2

2SLS estimates are only consistent and have reasonably small standard errors if the **instruments are strong**. This is measured by the **F-statistic** and may be retrieved using the `estat(firststage)` command.

```
. estat firststage

  First-stage regression summary statistics
  --------------------------------------------------------------------------
               |            Adjusted      Partial
      Variable |   R-sq.       R-sq.        R-sq.       F(4,44)   Prob > F
  -------------+------------------------------------------------------------
       hsngval |  0.6908      0.6557       0.5473       13.2978    0.0000
  --------------------------------------------------------------------------
```

The p-value for the F-statistic is most important to the frequentist logic in **weak instrument testing**.

Instrumental Variables in Panels

When dealing with both a cross-sectional and a time dimension, instrumenting becomes more difficult.

Your covariates need to be uncorrelated with your time-invariant and yout time-varying components of error for FE estimation. Then you can identify all time-varying estimators.

```
. use mus08psidextract.dta
. xtreg lwage ed exp wks, fe
note: ed omitted because of collinearity

Fixed-effects (within) regression               Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

R-sq:                                           Obs per group:
     within  = 0.6508                                         min =          7
     between = 0.0251                                         avg =        7.0
     overall = 0.0440                                         max =          7

                                                F(2,3568)         =    3325.13
corr(u_i, Xb)  = -0.9142                        Prob > F          =     0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          ed |          0  (omitted)
         exp |   .0969388    .001189    81.53   0.000     .0946077      .09927
```

IV in Fixed Effects Estimation 1

Problem: The FE estimation cannot identify the impact of time-invariant, such as years of education.

Further Problem: Assume that weeks worked `wks` is correlated with the time-varying part of the error (i.e. that workers who get paid more tend to stay on the job longer, or the other way around).

To solve the second problem, instrument weeks worked by marital status (**External Instrumentation**)

IV in Fixed Effect Estimation 2: STATA

```
. xtivreg lwage ed exp (wks=ms), fe

Fixed-effects (within) IV regression            Number of obs     =      4,165
Group variable: id                              Number of groups  =        595

R-sq:                                           Obs per group:
     within  =      .                                         min =          7
     between = 0.0126                                         avg =        7.0
     overall = 0.0223                                         max =          7

                                                Wald chi2(2)      =  641373.29
corr(u_i, Xb)  = -0.8570                        Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         wks |   -.120005   .2486092    -0.48   0.629    -.6072701    .3672601
          ed |          0  (omitted)
         exp |   .0962844   .0043809    21.98   0.000     .0876979    .1048709
       _cons |   10.38235   11.66474     0.89   0.373    -12.48011    33.24482
-------------+----------------------------------------------------------------
     sigma_u |  1.1547835
     sigma_e |  .53823759
         rho |  .82152826   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F  test that all u_i=0:     F(594,3568) =     3.34        Prob > F    = 0.0000
```

Hausman-Taylor Instrumentalization

Hausman and Taylor provide a instrumentalization procedure that allows for both endogenous time-varying and endogenous time-invariant variables.

Endogenous time-varying variables are estimated in a fixed effects procedure as their deviation from their individual mean over time. Endogenous time-invariant covariates are instrumentalized by exogenous time-invariant covariates. Note that there needs to be at least as many time-invariant exogenous as time-invariant endogenous variables, and they need to be **relevant** in estimation.

The procedure works without external instruments, and can be extended by using the non-diagonal covariance matrix of the error term to increase efficiency.

They distinguish four sets of variables, **time-varying exogenous** x_{1it}, **time-varying endogenous** x_{2it}, **time-invariant exogenous** w_{1it} and **time-invariant endogenous** w_{2it}.

Hausman-Taylor Instrumentalization 2

Consider an individual effects notation. x_{1it} and w_{1it} are exogenous (uncorrelated with α_i), x_{1it} and x_{2it} are time-varying. All are uncorrelated with ϵ_{it}. The challenge is to estimate both x_{2it} and w_{2it} consistently.

$$y_{it} = x_{it1}\beta_1 + x_{2it}\beta 2 + w_{1it}\gamma_1 + w_{2it}\gamma_2 + \alpha_i + \epsilon_{it}$$

Hausman and Taylor propose a **random effects** notation.

$$\tilde{y}_{it} = \tilde{x}_{it1}\beta_1 + \tilde{x}_{2it}\beta 2 + \tilde{w}_{1it}\gamma_1 + \tilde{w}_{2it}\gamma_2 + \tilde{\alpha}_i + \tilde{\epsilon}_{it}$$
$$\tilde{x}_{it} = x_{it} - \hat{\theta}_i \bar{x}_i$$

The random effects formulation with individual $\hat{\theta}_i$ allows for estimation of γ_1, γ_2 as $w_{1it}, w_{2it} \neq 0$.

However, $\tilde{\alpha}_i \neq 0$ and the individual effects are correlated with endogenous covariates $\tilde{x_{2it}}$ and $\tilde{w_{2it}}$. Here you need to **use instruments**.

Hausman-Taylor Instrumentalization 3

$\ddot{x}_{2it} = x_{2it} - \bar{x_{2i}}$ is uncorrelated with $\tilde{\alpha}_i$ and is used as an instrument for $\tilde{x}_{2it}$.

Exogenous and time-varying covariates x_{1it} are used as an instrument for time-invariant exogenous w_{2it} in a 2SLS procedure. Note that vector x'_{1it} has to be at least as long as w'_{2it}.

Hausman-Taylor Instrumentalization is STATA

```
use mus08psidextract.dta, clear
xthtaylor lwage occ sout smsa ind exp exp2 wks ms union fem blk ed,
(endog exp exp2 wks ms union ed)
```

Hausman-Taylor Instruments in STATA

The goal is to find a suitable estimation of years in education ed using the xthtaylor command in STATA, which is endogenous as it is correlated with individual effects α_i.

```
. xthtaylor lwage occ south smsa ind exp exp2 wks ms union fem blk ed, endog(ex

Hausman-Taylor estimation                       Number of obs      =      4,165
Group variable: id                              Number of groups   =        595

                                                Obs per group:
                                                             min =          7
                                                             avg =          7
                                                             max =          7

Random effects u_i ~ i.i.d.                     Wald chi2(12)      =    6891.87
                                                Prob > chi2        =     0.0000

------------------------------------------------------------------------------
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
TVexogenous  |
         occ |  -.0207047   .0137809    -1.50   0.133    -.0477149    .0063055
       south |   .0074398    .031955     0.23   0.816    -.0551908    .0700705
        smsa |  -.0418334   .0189581    -2.21   0.027    -.0789906   -.0046761
         ind |   .0136039   .0152374     0.89   0.372    -.0162608    .0434686
```

Hausman-Taylor Instruments in STATA 2

```
TVendogenous |
         exp |   .1131328    .002471    45.79   0.000     .1082898    .1179758
        exp2 |  -.0004189   .0000546    -7.67   0.000    -.0005259   -.0003119
         wks |   .0008374   .0005997     1.40   0.163    -.0003381    .0020129
          ms |  -.0298508     .01898    -1.57   0.116    -.0670508    .0073493
       union |   .0327714   .0149084     2.20   0.028     .0035514    .0619914
TIexogenous  |
         fem |  -.1309236    .126659    -1.03   0.301    -.3791707    .1173234
         blk |  -.2857479   .1557019    -1.84   0.066    -.5909179    .0194221
TIendogenous |
          ed |    .137944   .0212485     6.49   0.000     .0962977    .1795902
             |
       _cons |   2.912726   .2836522    10.27   0.000     2.356778    3.468674
-------------+----------------------------------------------------------------
     sigma_u |  .94180304
     sigma_e |  .15180273
         rho |  .97467788   (fraction of variance due to u_i)
------------------------------------------------------------------------------
Note: TV refers to time varying; TI refers to time invariant.
```

Dynamic Panels

As Panel Data is observed over time, including a lagged variable or **auto-regressive** term is an intuitive modeling choice.

Caution: OLS with a lagged variable and serially correlated errors leads to **inconsistent estimators** (as it does in the non-panel case).

When estimating a dynamic panel using fixed effects, **first differencing** must be used rather than **mean differencing**.

Arellano-Bond instrumentalization allows for efficient FD estimation in a dnymic model. Estimated parameters are consistent with both FE and RE models.

An AR(p) panel model

$$y_{it} = \gamma_1 y_{i,t-1} + ... + \gamma_p y_{i,t-p} + x'_{it}\beta + \alpha_i + \epsilon_{it}$$

3 channels of over-time correlation in y_i: **true state dependence** (directly $y_{i,t-1} \rightarrow y_{i,t}$), **observed heterogeneity** (directly through covariates $x_{i,t-1} \rightarrow x_{i,t} \rightarrow y_{i,t}$, or **unobserved heterogeneity** indirectly through α_i.

The within estimator (mean difference FE) is inconsistent with lags, as $y_{it} - \bar{y_i}$ is correlated with $\epsilon_{it} - \bar{\epsilon_i}$.

IV estimation using lags is also inconsistent, as $y_{i,t-s}$ is correlated with $\bar{\epsilon_i}$, and thus $\epsilon_{it} - \bar{\epsilon_i}$.

While first difference estimation will be inconsistent, using **appropriate lags of y_{it} as instruments** in FD estimation leads to consistent estimates.

First Difference Model

$$\Delta y_{it} = \gamma_1 \Delta y_{i,t-1} + ... + \gamma_p \Delta y_{i,t-p} + \Delta x'_{it}\beta + \Delta\epsilon_{it}$$

$\Delta y_{i,t-1}$ is correlated with $\Delta\epsilon_{i,t}$, but $y_{i,t-s}$ is not $\forall s > 2$. Anderson and Hsiao (1981) proposed using the second lag, while Arellano and Bond (1991) showed that efficiency is increased by using more lags as instruments, and that consistency holds under the assumption of **no serial correlation** in ϵ.

Regarding independent variables, you distinguish three categories. **Strictly exogenous covariates** (no problem), **weakly exogenous covariates** (correlated with past, but not with contemporaneous and future values of ϵ_{it}) and **temporarily endogenous covariates** (correlated with past and contemporaneous, but not future error terms).

You instrument accordingly with past values, and can also include external instruments.

Anderson Hsiao Instrumentalization

OLS estimates in short and broad panels will be upward biased due to correlation of the lagged coefficient with the error term.

Fixed effect estimate for laged covariate will be downward biased by size 1/T (**“Nickell bias”**)

Anderson Hsiao denotes a first-difference model, but instrumentalizes the first difference with 2- and 3-period lag differences.

```
regress n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr*, cluster(id)
estimates store OLS
xtreg n nL1 nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr*, fe cluster(id)
estimates store FE
ivregress 2sls D.n (D.nL1 = nL2) D.(nL2 w wL1 k kL1 kL2 ys ysL1 ysL2 yr1979 yr1
estimates store ahsiao1

esttab OLS FE
esttab ahsiao1
```

Anderson-Hsiao: Results 1

```
 esttab ahsiao1

----------------------------
                      (1)
                      D.n
----------------------------
D.nL1               2.308
                   (1.17)

D.nL2              -0.224
                  (-1.25)

D.w                -0.810**
                  (-3.10)

D.wL1               1.422
                   (1.21)

D.k                 0.253
                   (1.75)

D.kL1              -0.552
                  (-0.90)
```

Anderson-Hsiao: Results 2

D.kL2	-0.213
	(-0.89)
D.ys	0.991*
	(2.14)
D.ysL1	-1.938
	(-1.35)
D.ysL2	0.487
	(0.96)

Anderson-Hsiao: Results 3

```
D.yr1979              0.0467
                      (1.04)

D.yr1980              0.0761
                      (1.22)

D.yr1981              0.0226
                      (0.41)

D.yr1982              0.0128
                      (0.23)

D.yr1983             0.00991
                      (0.22)

_cons                 0.0159
                      (0.58)
----------------------------
N                        611
----------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001
```

“‘

Arellano-Bond Instrumentalization 1

In dynamic models which you estimate using FE, note the difference between **mean differencing** $x_{it} - \bar{x_i}$ and **first differencing $x_{it} - x_{i(t-1)}$.

Remember: when you include serially correlated errors and/or lagged dependent (autoregressive) variables, OLS estimation of an FE model is **inconsistent**.

Arellano-Bond estimation uses a sufficient number of lags as instruments for dependent variables, which is often more efficient than OLS estimation.

You just made the step to **dynamic panel modeling**.

Arellano-Bond Instrumentalization 2

$$y_{it} = \gamma_1 y_{i(t-1)} + ... + \gamma_p y_{i(t-p)} + x'_{it}\beta + \alpha_i + \epsilon_{it}$$

Note: Both **within-estimation** and **lag instrumentalization** will be inconsistent for correlation between mean differences $y_{it} - \bar{y_i}$ or lags $y_{i(t-p)}$ and $\epsilon_{it} - \bar{\epsilon_i}$. For the FD estimation, assume that ϵ_{it} is **serially uncorrelated.

$$\Delta y_{it} = \gamma_1 \Delta y_{i(t-1)} + ... + \gamma_{p-1} \Delta y_{i(t-p)} + \Delta x'_{it}\beta + \Delta\epsilon_{it}$$

You can instrument for $\Delta y_{i(t-1)}$ using enough lags $y_{i(t-2),...,y_{i(t-s)}}$, and Δx_{it} by x_{it} themselves, if x_{it} are exogenous. If x_{it} are not exogenous, they can be instrumented by enough lags of themselves.

Arellano Bond instrumentalization in STATA

```
. xtabond lwage, lags(2) vce(robust)

Arellano-Bond dynamic panel-data estimation     Number of obs     =      2,380
Group variable: id                              Number of groups  =        595
Time variable: t
                                                Obs per group:
                                                              min =          4
                                                              avg =          4
                                                              max =          4

Number of instruments =     15                  Wald chi2(2)      =    1253.03
                                                Prob > chi2       =     0.0000
One-step results
                                     (Std. Err. adjusted for clustering on id)
------------------------------------------------------------------------------
             |               Robust
       lwage |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       lwage |
         L1. |   .5707517   .0333941    17.09   0.000     .5053005    .6362029
         L2. |   .2675649   .0242641    11.03   0.000     .2200082    .3151216
             |
       _cons |   1.203588    .164496     7.32   0.000     .8811814    1.525994
------------------------------------------------------------------------------
Instruments for differenced equation
        GMM-type: L(2/.).lwage
```