Lab 4: Endogeneity and Instrumental Variable Estimation in Panel Data

GECO 6281 Advanced Econometrics 1

Patrick Mokre

Fall 2020

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 3: 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?

Relationship between RE and FE

The relationship between the RE and FE setup is determined by the **heterogeneity in α_i and** σ_α^2. For maximum heterogeneity, RE converges to FE, for minimum heterogeneity, RE converges to the pooled OLS estimator.

Furthermore, for $T \to \infty$, RE and FE estimators converge.

Deciding on a model

There are different approaches to choosing between FE and RE setups. These are some:

1 Theoretical determinationn (Pesaran 2015): If we are interested in between-individual heterogeneity, FE makes sense. If N is large and you consider it a random sample from the population, RE is more appropriate. More technically, the decision variable is your **beliefs about the correlation between individual effects and covariates** x_{it}.

2 Hausman Test: The HT tests under the null that effects are random and compares the FE and RE estimators. Under the Null, the estimators converge. in STATA, you need to run both models, store the estimates, and use the `hausman` command.

3 Gelman's Rejection of fixed effects: Andrew Gelman, an important researcher into Bayesian multilevel modeling argues, that the notion of "fixed effects models" makes little effect in and of itself, and one should rather assume all downstream hierarchical coefficients are the product of **some** random distribution. However, this is easier said in Bayesian statistics, as it allows for distributions other than the Gaussian Normal.

Hausman Test: STATA

```
. quietly xtreg lwage exp ind, fe

. estimates store FE

. quietly xtreg lwage exp ind, re

. estimates store RE

. hausman FE RE

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |       FE           RE         Difference          S.E.
-------------+----------------------------------------------------------------
         exp |    .0969207     .0612741        .0356466        .0002741
         ind |     .022139    -.0123885        .0345275               .
------------------------------------------------------------------------------
                           b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =    15144.30
                Prob>chi2 =      0.0000
                (V_b-V_B is not positive definite)
```

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)
```