# Homebaked *Wooldridge* Econometrics Revision Guide

Xiaolong Yang Intermediate Econometrics
National School of Development, Peking University

1/1/2022

## Contents



# 1 DiD: pooled cross-sectional and panel data

## 1.1 Repeated cross-sections

Let us start by thinking, what is *not* Panel Data?

- The Chow test for structural change across time
  - two time periods
  - many time periods and explanatory variables
- Policy Analysis with Pooled Cross-sectional Data
- Identification strategies
  - cross-section comparison
    * Assumption: $E\left(y_t \mid x_1, ..., x_k,\ D=1\right) = E\left(y_t \mid x_1, ..., x_k,\ D=0\right)$
    * Regression model: $y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_k x_{ik} + \alpha D_i + u_i$
    * Mean effect: $\hat{a} = E\left(y_t + \triangle \mid x_1, ..., x_k,\ D=1\right) - E\left(y_t \mid x_1, ..., x_k,\ D=0\right)$

    - problem: self-selection/unobserved heterogeneity
  - before and after comparison
    - Assumption: $E\left(y_t \mid x_1, ..., x_k,\ D=1\right) = E\left(y_{t'} \mid x_1, ..., x_k,\ D=1\right)$
    - Regression model: $y_{it} \ = \beta_0 + \beta_1 x_{it1} + ... + \beta_k x_{itk} + \gamma T_{it} + u_{it}$
    - Mean effect: $\hat{\gamma} = E\left(y_t + \triangle \mid x_1, ..., x_k,\ D=1\right) - E\left(y_{t'} \mid x_1, ..., x_k,\ D=1\right)$
    - problems: pooled cross-sectioinal or panel data are needed; business cycle sensitivity; Ashenfelter's Dip (units have prior knowledge of treatment assignment hence are prone to change their behaviour accordingly; think about posttreatment bias)
  - DiD estimation
    - Assumption (problematic in practice):

$$E\left(y_t + \triangle \mid x_1, ..., x_k,\ D=1\right) - E\left(y_t \mid x_1, ..., x_k,\ D=1\right)$$

$$= E\left(y_t + \triangle - y_{t'} \mid x_1, ..., x_k,\ D=1\right) - E\left(y_t - y_{t'} \mid x_1, ..., x_k,\ D=0\right)$$

    - Advantage: elimination of the unwanted influence of unobserved heterogeneity
    - Regression model: $y_{it} \ = \beta_0 + \beta_1 x_{it1} + ... + \beta_k x_{itk} + \delta_1 D_i + \delta_2 T_{it} + \delta_3 (D_i \cdot T_{it}) + u_{it}$
    - Mean effect:

$$\begin{aligned} \hat{\delta_3} = &\left[E\left(y_t + \triangle \mid x_1, ..., x_k,\ D=1\right) - E\left(y_{t'} \mid x_1, ..., x_k,\ D=1\right)\right] \\ &-\left[E\left(y_t \mid x_1, ..., x_k,\ D=0\right) - E\left(y_{t'} \mid x_1, ..., x_k,\ D=0\right)\right] \end{aligned}$$

- Problems: pooled cross-sectioinal or panel data are needed; temporary economic fluctuations that affect outcomes of participants and non-participants differently; Ashenfelter's Dip (units have prior knowledge of treatment assignment hence are prone to change their behaviour accordingly; think about posttreatment bias)

### 1.2 Panel Data

- Balanced panel = 0 attrition rate of data
- Advantages
  - reduces data needs
  - could control for unobserved heterogeneity
  - possible to identify the direction of causation
  - study the importance of time dimension in decision making
- Limits
  - collection over long time
  - simple panel analysis may exacerbate measurement error (twice than corss-section)
  - still has selectivity problem (attrition could introduce severe selection problems)
  - what if the main variables of interest do not vary across time

Consider a *fixed-effects* two-period panel data model

$$y_{it} \ = \beta_0 + \beta_1 x_{it1} + \delta_0 d2_t + \alpha_i + u_{it}$$

- $\alpha_i$ are the things vary across individuals but not over time, which are referred to as
  - Fixed effect
  - Unobserved heterogeneity

    - Unobserved individual effect
- The primary strength of panel data analysis is the *ability to remove* $\alpha_i$
- Two techniques
    - *Differencing* (First-differences model): $\triangle y_i = \delta_o + \beta_0 \triangle x_i + \triangle u_i$
        * better than OLS
    - *Demeaning* (Fixed-effect model): $(y_{it} - \bar{y_i}) = \beta_1(x_{it} - \bar{x_i}) + \delta_o(d2_t - \bar{d2}) + (u_{it} - \bar{u_i})$
        * use all nT observations unlike differencing
        * if T=2, first differencing and demeaning produce identical coefficient estimates and s.e.
        * $\alpha_i$ is swept out of the model $\Rightarrow$ unbiased estimators even if $Cov(\alpha, \mathbf{X}) \neq 0$
        * cannot estimate anything that is constant over time or has a constant rate of change
    - If $T = 2$, the estimates and test statistics between FE and FD are the same
    - If $T = 3$, FE is more efficient if the $u_{it}$ are serially correlated
    - Good to compare FE and FD $\Rightarrow$ assumptions could be wrong if observe a difference in estimates

# 2 Instrumental Variables

## 2.1 Omitted Variables in a Simple Regression Model

### 2.1.1 Four ways of dealing with Omitted Variables problem

- Do nothing in estimation but argue about the possible bias
- Proxy variable
- Panel data
- Instrumental variables approach (IV)

### 2.1.2 Two assumptions for IV

- Instrument exogeneity : $Cov(z, u) = 0 \Rightarrow$ *empirically un-testable*; use logic and intuition
    - $z$ has no partial effect on $y$
    - $z$ should be uncorrelated with the omitted variable
- Instrument relevance: $Cov(z, x) \neq 0 \Rightarrow x = \pi_i + \pi_1 z + v$ and see if $\pi_1 \neq 0$

### 2.1.3 Identification with IV

- Write $\beta_1$ as population covariances, then

$$y = \beta_0 + \beta_1 x + u$$

$$cov(z, y) = \beta_1 cov(z, x) + cov(z, u),$$

  where $cov(z, u) = 0$, hence

$$\beta_1 = \frac{cov\,(z, y)}{cov\,(z, x)},$$

  and

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- When $z = x$, meaning x is exogeneous, we obtain the OLS estimator of $\beta_1$
- $plim(\hat{\beta}_1) = \beta_1 \Rightarrow$ consistent if assumptions satisfied

#### 2.1.4 Inference with IV:

Need a s.e. to compute $t$ statistics and confidence intervals $\Rightarrow$ homoscesdasticity assumption conditional on $z$.

$$E(u^2 \mid z) = \sigma^2 = Var(u)$$

hence the asymptotic variance of $\hat{\beta}_1$ is

$$\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$$

where $\rho_{x,z}^2$ is the square of the populartion correlation between $x$ and $z$.

- If $Cov(z, x)$ is weak, then $R^2$ for $x, z$ regression can be small $\Rightarrow$ large sampling variance for the IV estimtaor
- The asymptotic variance of the IV estimator is always larger when $Cov(x, u) \neq 0$

#### 2.1.5 $R^2$ of IV estimation

$$R^2 = 1 - \frac{SSR_{iv\ residuals}}{SST_y}$$

- Can be negative
- Cannot be used to compute $F$ tests of joint restrictions
- No natural interpretation when $x$ and $u$ are correlated

### 2.2 IV Estimation of the Multiple Regression Model

A structural equation (emphasise on $\beta_s$)

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

to obtain consistent estimators for $\beta_s$ we need an instrument $z_2$ that satisfies

$$E(u_1) = 0,\ Cov(z_1, u_1) = 0,\ and\ Cov(z_2, u_1) = 0$$

A reduced form equation is

$$y_2 = \pi_0 + \pi_0 z_1 + \pi_2 z_2 + v_2,\ \pi_2 \neq 0$$

and we use this to state the key identification condition that a valid instrument needs to be correlated with the endogenous variables.

### 2.3 Two Stage Least Squares

Often more than 1 valid IVs for the single endogenous variable $\Rightarrow$ how to use multiple IVs $\Rightarrow$ *the Two Stage least sqaures (2SLS) estimator.*

Consider the structural equation

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

and suppose we have two exogenous variables: $z_2, z_3$ and they satisfy *exclusion restrictions*

- $z_2, z_3$ do not appear in structural equation

- $z_2$ and $z_3$ are uncorrelated with the error $u_1$

The best IV for $y_2$ is hence the linear combination of the $z_j$

$$y_2^* = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3$$

where $\pi_2 \neq 0, \pi_3 \neq 0$. The structural equation is not identified if $\pi_2 = 0$ and $\pi_3 = 0$; we can also use $F$ statistic to test $H_0$:$\pi_2 = 0$ and $\pi_3 = 0$.

*1st Stage.* Obtain the fitted model with our sample:

$$\hat{y_2^*} = \hat{\pi}_0 + \hat{\pi}_1 z_1 + \hat{\pi}_2 z_2 + \hat{\pi}_3 z_3$$

*2nd Stage 2.* The OLS regression of $y_1$ on $\hat{y_2}$ and $z_1$:

$$y_1 = \beta_0 + \beta_1 \hat{y_2} + \beta_2 z_1 + u_1$$

- The 2SLS estimates can differ substantially from the OLS estimates.
- Avoid doing the second stage manually as the standard errors and test statistics obtained in this way are not valid.

The asymptotic variance of the *2SLS* estimator of $\beta_1$ approximated as $\frac{\sigma^2}{\widehat{SST_2}(1-\hat{R_2^2})}$ is greater than that of *OLS*, because

- $\hat{y_2}$ has less variation than $y_2$
- Multicollinearity problem in 2SLS: correlation between $\hat{y_2}$ and the exogenous variables is often much higher than the correlation between $y_2$ and these variables.
- 2SLS can also be used in model with more than one endogenous explanatory variable

However, we need at least two exogenous variables that do not appear in the structural equation but are correlated with the endogenous variables $y_2$ and $y_3$

- order condition
- rank condition

## 2.4 Testing for Endogeneity

When the explanatory variables are exogenous 2SLS is less efficient than OLS (large s.e.) $\Rightarrow$ test for endogeneity of an explanatory variable

Consider

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$$

where $y_2$ is the suspected endogenous explanatory variable; we also have two additional exogeneous variables $z_3, z_4$.

*1st Step.*

$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2$$

*2nd Step.*

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta_1 \hat{v}_2 + error$$

and test for $H_0 : \delta_1 = 0$ using a $t$ statistic. If we reject $H_0$ at a small significant level, we conclude that $y_2$ is endogenous because $Cov(v_2, u_1) \neq 0$.

# 3 Specification and Data Issues

## 3.1 Functional Form Misspecification

- *Omitted variable bias*: if $Cov(u, x_j) \neq 0 \Rightarrow x_j$ is endogenous, which leads to
  - biasedness
  - inconsistency in all OLS estimators
- Functional Form Misspecification is a special case of omitted variable bias
  - omit the *squared* terms
  - omit the *interaction* terms
  - use the *level* of a variable rather than its *log* form

### 3.1.1 Tests for Functional Form Misspecification

*RESET* (Ramsey's (1969) Regression Specification Error Test)

- Logic: adds polynomials in the OLS fitted values to the original regression to detect general kinds of functional form misspecification

1. Estimate $y = \beta_0 + \beta_1 x1 + \beta_2 x2 + ... + \beta_k x_k + u$ to obtain $\hat{y}$
2. Estimate the expanded function

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + \delta_1 \hat{y}^2 + \delta_2 \hat{y}^3 + error$$

3. test $H_0 : \delta_1 = \delta_2 = 0$

- $\Rightarrow$ apply $F$ test with 2 and $(n - k - 1) - 2 = n - k - 3$ degrees of freedom
- If $\delta_1$ and $\delta_2$ are *jointly insignificant*, then the original model is correctly specified

Limitations of the RESET test

- No implication on the correct specification even if misspeecification is detected
- Has no power for detecting omitted variables or heteroscedasticity whenever they have expectations that are linear in the included independent variables
- No power for detecting heteroskedasticity if the functional form is correctly specified

Tests against Nonnested (not F) Alternatives: *Davidson-MacKinnon test*

- Logic: to decide whether an independent variable should appear in level or logarithmic form

We test model 1 against model 2:

***Model 1.***

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$$

***Model 2.***

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u$$

1. Estimate model 1. to obtain the predicted values $\hat{y}$
2. Estimate model 2. to obtain the predicted values $\hat{\hat{y}}$
3. Estimate the models

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \theta_1 \hat{\hat{y}} + error$$

and

$$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u + \theta_2 \hat{y} + error$$

The Davidson-MacKinnon test is based on the $t$ statistic on $\hat{\hat{y}}$ and $\hat{y}$ in the two separated equations

- If $\theta_1$ is significant, then the level equation is rejected
- If $\theta_2$ is significant, then the log equation is rejected

Limitation of The Davidson-MacKinnon test

- The test cannot be applied if the sets of independent variables are different
- The test is not helpful if both models are rejected
  - if both models are not rejected, we take the model with the higher (adjusted) $R^2$
  - if the effects of key independent variables on $y$ are not very different, then it does not really matter which model is used
- The level model can be rejected for a variety of functional form misspecification (not necessarily for the log form)
- Obtaining nonnested tests when the leading case is $y$ versus $\log(y)$ is difficult

## 3.2 Proxy Variables

A proxy variable is a variable that is related to the *unobserved variable* that we would like to include in our model e.g., IQ as a proxy variable for ability.

*Formal setup*

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

where $x3^*$ is unobserved and $x_3$ is the observed proxy variable, and

$$x3^* = \delta_0 + \delta_3 x_3 + v_3$$

### 3.2.1 Minimum requirements to obtain unbiased estimates of $\beta_1$ and $\beta_2$

- $u$ has to be uncorrelated with $x_1, x_2;, x_3^*$ and $x_3$
- $E(x_3^* \mid x_1, x_2, x_3) = E(x_3^* \mid x_3) = \delta_0 + \delta_3 x_3 \Rightarrow v_3$ has to be uncorrelated with $x_1, x_2, x_3$ ($x_3$ must be a *good proxy* for $x_3^*$.

### 3.2.2 Use $x_3$ to get unbiased estimators of $\beta_1$ and $\beta_2$

Plugging in $\delta_0 + \delta_3 x3 + v_3$ for $x3^*$ in popultion model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3^* + u$$

this yields

$$y = (\beta_0 + \beta_3 \delta_0) + \beta_1 x_1 + \beta_2 x_2 + \beta_3 \delta_3 x_3 + u + \beta_3 v_3$$

where

$$y = \alpha_0 + \beta_1 x_1 + \beta_2 x_2 + \alpha_3 x_3 + e$$

with $\alpha_0 = \beta_0 + \beta_3 \delta_0,\ \alpha_3 = \beta_3 \delta_3$, and $e = u + \beta_3 v_3$

$\Rightarrow$ since both $u$ and $v_3$ i.e., $e$ are uncorrelated with $x_1, x_2, x_3$, we have unbiased estimates of $\beta_1$ and $\beta_2$.

- and unbiased estimates of $\alpha_0, \alpha_3$

### 3.2.3 Using Lagged Dependent Variables as Proxy Variables

Using lagged dependent variable in a cross-sectional equation provides a simple way to account for *historical factors* that cause *current differences* in the dependent variable.

$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 y_{t-1} + u$$

where the lagged dependent variable $y_{t-1}$ could control for historical confounders.

## 3.3 Measurement error

Imprecise measurement $\Rightarrow$ measurement error

- Measurement error in $x$
- Measurement error in $y$

### 3.3.1 Measurement Error in the Dependent Variable ($y$)

$$e_0 = y - y^*$$

where $y^*$ is the unobserved actual dependent variable.

We then obtain an estimable regression model by plugging $e_0 = y - y^*$ into a regression equation that satisfies the Gauss-Markov assumptions (MLR.1-4):

$$y_t = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u + e_0$$

OLS estimators will be unbiased

- If $E(e_0) \neq 0$ which is naturally the case
- If $E(e_0 \mid x_1, ..., x_k) = 0 \rightsquigarrow$ if the the measurement error is systematically related to one or more of the explanatory variables, it can cause biased OLS estimators
- If $e_0$ and $u$ are uncorrelated, then $Var(u + e_0) = \sigma_u^2 + \sigma_e^2 > \sigma_u^2$, which sacrifices efficiency (statistical significance) due to larger error variance

### 3.3.2 Measurement Error in the Dependent Variable ($x$)

For a regression model that satisfies the Gauss-Markov assumptions, the measurement error of independent variable is

$$e_k = x_k - x_k^*$$

Assumptions

- $E(e_k) = 0 \rightsquigarrow$ the average measurement error in the population is zero
- $E(u \mid x_k) = E(u \mid x_k^*) = E(u \mid x_k, x_k^*) = 0 \Rightarrow E(y \mid x_k, x_k^*) = E(u = y \mid x_k^*) \rightsquigarrow x_k$ does not affect $y$ after $x_k^*$ has been controlled for

Actual model estimated:

$$y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$$

Whether we can obtain unbiased estimators after replacing $x_1^*$ with $x_1$ depends on the assumptions made about the correlation between measurement error $e_1$ and $x_1$:

- $Cov(x_1, e_1) = 0$

- $E(u) = E(e_1) = 0$ and $Cov(x_1, u) = Cov(x_1, e_1) = 0$, $E(u - \beta_1 e_1) = 0$ and $Cov(x_1, u - \beta_1 e_1) = 0 \Rightarrow$ unbiased estimates of $\beta_0$ and $\beta_1$
- since $u$ is uncorrelated with $e_1$, $Var(u - \beta_1 e_1) = \sigma_u^2 + \beta_1^2 \sigma_{e1}^2 \Rightarrow$ measurement error increases the error variance (unless $\beta_1 = 0$) but this does not affect the OLS properties

- $Cov(x_1, e_1) \neq 0$
  - *Classical errors-in-variable (CEV) assumption*

$$Cov(x_i^*, e_1) = 0 \Rightarrow Cov(x_1, e_1) = E(x_1, e_1) = E(x_1^*, e_1) + E(e_1^2) = 0 + \sigma_{e1}^2 = \sigma_{e1}^2$$

- Since $y = \beta_0 + \beta_1 x_1 + (u - \beta_1 e_1)$, *the OLS estimates will be biased and inconsistent*

$$Cov(x_1, u - \beta_1 e_1) = -\beta_1 Cov(x_1, e_1) = -\beta_1 \sigma_{e1}^2$$

- This leads to *attenuation bias* i.e., $\beta_1$ is biased towards zero, because

$$plim \hat{\beta}_1 = \beta_1 (\frac{\sigma_{x_1^*}^2}{\sigma_{x_1^*}^2 + \sigma_{e_1}^2}) = \beta_1 \frac{Var(x_1^*)}{Var(x_1)}$$

and

$$Var(x_1) > Var(x_1^*) \Rightarrow \frac{Var(x_1^*)}{Var(x_1)} < 1$$

## 3.4 Missing Data, Nonrandom Samples and Outliers

It is said to be a data problem when the random sampling assumption is violated. Assumption MLR.2: We have a random sample of n observations, $\{(x_{i1}, x_{i2}, ..., x_{ik}, y_1) : i = 1, ..., n\}$, following the population model $y = \beta_0 x_1 + \beta_1 x_2 + ... + \beta_k x_k + u$.

There are three situations of violation:

- Missing data
- Nonrandom sampling
- Outliers

### 3.4.1 Missing data

- Missing at random $\Rightarrow$ no bias but $\downarrow$ sample size hence $\downarrow$ precise estimation
- Missing systematically $\Rightarrow$ biased estimates

### 3.4.2 Nonrandom Samples

Missing data is more problematic when it results in a nonrandom sample from the population

- If the sample selection based on the **independent variables**, the estimators are **unbiased** $\Rightarrow$ **exogenous sample selection**.
- If the sample selection based on the **dependent variables**, the estimators are **biased and inconsistent** $\Rightarrow$ **endogenous sample selection**.
  - possible solution: Tobit model

**Sample selection bias**

- Endogenous selection can result in a sample selection bias in the OLS estimates
- Possible solution: Heckman selection model

### 3.4.3 Outliers/Influential observations

An observation is an outlier if dropping it from the analysis changes the **key OLS estimates** by a practically **large** amount

Why there are outliers:

- Data entry mistakes
- One or several members of the population are very different in some relevant aspect from the rest of the population
  - OLS should be reported with and without outlying observations

Ways to Deal with Outliers:

- Drop the outliers when comprise $< 5\%$ of the sample population
- Functional form transformation (to forms less sensitive to outliers)
  - Log forms
- Use method that is less sensitive to outliers than OLS i.e., Least absolute deviations (LAD)

# Citation

Wooldridge, Jeffrey M. 2016. *Introductory Econometrics : A Modern Approach.* Sixth edition..; Student edition..