

# Notes on Regression

Alex Young

April 22, 2012

## Contents

<b>1</b>	<b>Preface</b>	<b>4</b>
<b>2</b>	<b>The Experimental Ideal [Angrist and Pischke, 2008]</b>	<b>4</b>
2.1	Selection Bias . . . . .	4
<b>3</b>	<b>The Classical Linear Regression Model</b>	<b>5</b>
3.1	Derivation of OLS . . . . .	5
3.1.1	The Argument That Minimizes the SSR . . . . .	6
3.1.2	The Method of Moments . . . . .	6
3.1.3	Maximum Likelihood . . . . .	7
3.2	Interpretation of Coefficients [Stock and Watson, 2007] . . . . .	8
3.3	Finite and Asymptotic Properties . . . . .	9
3.3.1	Unbiasedness . . . . .	9
3.3.2	Consistency . . . . .	9
3.3.3	Normality . . . . .	10
<b>4</b>	<b>Instrumental Variables</b>	<b>10</b>
4.1	Motivation: Simultaneity . . . . .	10

4.1.1	Working’s Example . . . . .	10
4.1.1.1	Preview of 2SLS . . . . .	14
4.1.2	Haavelmo’s Example . . . . .	15
4.2	Motivation: Errors-in-variables . . . . .	16
4.3	Motivation: Omitted Variables Bias [Angrist and Pischke, 2008] . . . . .	19
4.3.1	Wages and education . . . . .	19
4.3.2	Caveat: Are omitted variables always a “problem”? . . . . .	20
4.4	Properties . . . . .	21
4.4.1	Consistency . . . . .	21
4.4.2	Normality . . . . .	23
4.5	Forbidden Regressions [Angrist and Pischke, 2008] . . . . .	23
4.5.1	Example: Founder-CEO Endogeneity [Adams et al., 2009] . . . . .	24
<b>5</b>	<b>The Generalized Method of Moments: Single-Equation</b>	<b>25</b>
5.1	OLS and IV as GMM estimators . . . . .	26
5.2	Sampling Error and Consistency . . . . .	26
5.3	Efficient GMM, or “Making the formulas look nicer” . . . . .	27
5.3.1	Imposing Conditional Homoskedasticity . . . . .	28
5.3.2	2SLS as an IV estimator or as Two Separate Regressions . . . . .	29
<b>6</b>	<b>Panel Data and Clustering</b>	<b>31</b>
6.1	OLS . . . . .	31
6.1.1	Clustering . . . . .	32
6.1.2	Special Case: The individual unit of analysis is the cluster . . . . .	33
6.1.3	A more special case: Plain ol’ heteroskedasticity-robust standard errors	35
6.2	Random and Fixed Effects . . . . .	36
6.2.1	Random Effects . . . . .	37
6.2.2	Fixed Effects . . . . .	39

6.2.2.1	Interlude: Fixed Effects and Least-Squares Dummy Variables	43
6.3	Differences-in-Differences . . . . .	45
6.3.1	Minimum wage on unemployment [Angrist and Pischke, 2008] . . . . .	45
6.3.2	Empirical Corporate Finance: Collateral on Lending [Assunção et al., 2011] . . . . .	47
6.3.3	Empirical Corporate Finance: Antitakeover legislation on managers’ preferences [Bertrand and Mullainathan, 2003] . . . . .	47
<b>7</b>	<b>The Fama-MacBeth Approach [Campbell et al., 1997]</b>	<b>48</b>

# 1 Preface

These notes largely follow Hayashi [2000]. I wrote these because fundamental econometric theory (i.e. first-year course coverage) is important—you should know on paper what your code is doing—and I habitually forget things.

## 2 The Experimental Ideal [Angrist and Pischke, 2008]

### 2.1 Selection Bias

Think about treatment as described by a binary random variable:  $D_i = \{0, 1\}$ . The outcome of interest is denoted by  $Y_i$ . The question is whether  $Y_i$  is *affected* by the treatment. To address this question, assume we can imagine what might have happened to someone who was treated if that person had not been (treated), and vice versa.

Hence, for any individual, there are two potential outcomes:  $1 \cdot Y_{1i}[D_i = 1]$ . We would like to know the difference between  $Y_{1i}$  and  $Y_{0i}$ , which can be said to be the causal effect of the treatment for individual  $i$ .

The observed outcome,  $Y_i$ , can be written in terms of potential outcomes as

$$Y_i = Y_{0i} + (Y_{1i} - Y_{0i}) \cdot D_i$$

A naïve comparison of averages by treatment tells something about potential outcomes, though not necessarily what we want to know, which is the causal effect (of treatment). The observed difference in average outcome is formally linked to the average causal effect by the

equation

$$\begin{aligned} E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 0) &= \{E(Y_{1i}|D_i = 1) - E(Y_{0i}|D_i = 1)\} \\ &\quad + \{E(Y_{0i}|D_i = 1) - E(Y_{0i}|D_i = 0)\} \end{aligned}$$

The first term in braces is the average treatment effect on the treated. The observed difference in treatment, however, adds to this causal effect *selection bias*, which is the difference in average  $Y_{0i}$  between those who were and were not treated.

### 3 The Classical Linear Regression Model

The classical regression model is a set of joint distributions satisfying the following assumptions:

A1 **Linearity:**  $y_i = \sum_{j=1}^K \beta_j x_{ij} + \varepsilon_i$  ( $i \in \mathbb{N}$ ) =  $\mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i$ .

A2 **Strict exogeneity:**  $\mathbb{E}(\boldsymbol{\varepsilon} | \mathbf{X}) = \mathbf{0}$ .

A3 **No multicollinearity:** The rank of the  $n \times K$  data matrix,  $\mathbf{X}$ , is  $K$  with probability 1.

A4 **Spherical error variance:**  $\text{Var}(\boldsymbol{\varepsilon} | \mathbf{X}) = \sigma^2 I$ . That is,  $\mathbb{E}(\varepsilon_i^2 | \mathbf{X}) = \sigma^2 > 0$  ( $i \in \mathbb{N}$ ) and  $\mathbb{E}(\varepsilon_i \varepsilon_j | \mathbf{X}) = 0$  ( $i, j \in \mathbb{N}$  and  $i \neq j$ ).

The last assumption is also known as *homoskedasticity*.

#### 3.1 Derivation of OLS

The error term is unobservable. Instead, we calculate the residual implied by a hypothetical value  $\tilde{\boldsymbol{\beta}}$  of  $\boldsymbol{\beta}$  as

$$e_i \equiv y_i - \mathbf{x}_i' \tilde{\boldsymbol{\beta}}$$

Accordingly, the sum of squared residuals (SSR) is

$$\begin{aligned}
 SSR(\tilde{\boldsymbol{\beta}}) &\equiv \sum_{i=1}^n e_i^2 \\
 &= \sum_{i=1}^n (y_i - \mathbf{x}'_i \tilde{\boldsymbol{\beta}})^2 \\
 &= (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})
 \end{aligned}$$

### 3.1.1 The Argument That Minimizes the SSR

The Ordinary Least Squares estimate,  $\mathbf{b}$ , is the  $\tilde{\boldsymbol{\beta}}$  that minimizes the SSR:

$$\begin{aligned}
 \mathbf{b} &\equiv \arg \min_{\tilde{\boldsymbol{\beta}}} SSR(\tilde{\boldsymbol{\beta}}) \\
 &= \arg \min_{\tilde{\boldsymbol{\beta}}} (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \\
 &= \arg \min_{\tilde{\boldsymbol{\beta}}} \mathbf{y}'\mathbf{y} - 2\mathbf{y}'\mathbf{X}\tilde{\boldsymbol{\beta}} + \tilde{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}}
 \end{aligned}$$

Taking the FOC and setting it equal to zero produces the fabled OLS expression:

$$\begin{aligned}
 \frac{\partial SSR(\tilde{\boldsymbol{\beta}})}{\partial \tilde{\boldsymbol{\beta}}} &= -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\tilde{\boldsymbol{\beta}} \\
 &= 0 \\
 \Rightarrow \tilde{\boldsymbol{\beta}} &\equiv \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}
 \end{aligned}$$

### 3.1.2 The Method of Moments

But OLS can be derived another way through the **method of moments**, which relies on the principle of choosing the parameter estimate so that if a set of population moments are all equal to zero, the corresponding sample moments are also equal to zero.

The strict exogeneity assumption implies that  $\mathbf{x}_i$  and  $\varepsilon_i$  are uncorrelated (i.e. orthogonal):

$$\begin{aligned}\mathbb{E}(\mathbf{x}_i \varepsilon_i) &= \mathbb{E}(\mathbb{E}(\mathbf{x}_i \varepsilon_i) \mid \mathbf{X}) \\ &= \mathbb{E}(\mathbf{x}_i \mathbb{E}(\varepsilon_i \mid \mathbf{X})) \\ &= 0\end{aligned}$$

We set the corresponding sample moment equal to zero:

$$\begin{aligned}\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i &= 0 \\ \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{x}_i' \tilde{\boldsymbol{\beta}}) &= 0 \\ \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i - \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \tilde{\boldsymbol{\beta}} &= 0 \\ \tilde{\boldsymbol{\beta}} \equiv \mathbf{b} &= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{x}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) \\ &= (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\mathbf{y}\end{aligned}$$

### 3.1.3 Maximum Likelihood

There is still another way to derive the OLS estimator: the method of **maximum likelihood**.

For the derivation, we add an additional assumption, which is also used to conduct statistical inference in finite samples:

(A5) **Normality**: The distribution of  $\varepsilon$  conditional on  $\mathbf{X}$  is jointly normal.

Since  $\mathbf{y} = \mathbf{X}\beta + \varepsilon$ ,  $\mathbb{E}(\mathbf{y} | \mathbf{X}) = \mathbf{X}\beta$ . The conditional density of  $\mathbf{y}$  given  $\mathbf{X}$  is then

$$\begin{aligned} f(\mathbf{y} | \mathbf{X}) &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbb{E}(\mathbf{y} | \mathbf{X}))'(\mathbf{y} - \mathbb{E}(\mathbf{y} | \mathbf{X}))\right] \\ &= \frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)\right] \end{aligned}$$

We replace the true parameters by their hypothetical values and take logs for simplification to obtain the **log likelihood function**:

$$\begin{aligned} \log L(\tilde{\boldsymbol{\beta}}, \tilde{\sigma}^2) &= \log \left\{ \frac{1}{(2\pi\tilde{\sigma}^2)^{\frac{n}{2}}} \exp\left[-\frac{1}{2\tilde{\sigma}^2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})\right] \right\} \\ &= -\frac{n}{2} \log(2\pi) - \frac{n}{2} \log(\tilde{\sigma}^2) - \frac{1}{2\tilde{\sigma}^2}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}) \end{aligned}$$

As the name “maximum” likelihood would suggest, we take the FOC of the log likelihood function with respect to  $\tilde{\boldsymbol{\beta}}$ . Maximizing it is equivalent to minimizing  $(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$ . But we know that the  $\arg \min (\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})'(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$  is none other than  $\mathbf{b}$ .

### 3.2 Interpretation of Coefficients [Stock and Watson, 2007]

$$Y_i = \beta_0 + \beta_1 \ln(X_i) + u_i \tag{1}$$

$$\ln(Y_i) = \beta_0 + \beta_1 X_i + u_i \tag{2}$$

$$\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i \tag{3}$$

1. A 1% change in  $X$  is associated with a change in  $Y$  of  $0.01\beta_1$ .
2. A change in  $X$  by 1 unit ( $\Delta X = 1$ ) is associated with a  $100\beta_1\%$  change in  $Y$ .
3. A 1% change in  $X$  is associated with a  $\beta_1\%$  change in  $Y$ , so  $\beta_1$  is the elasticity of  $Y$  with respect to  $X$ .

### 3.3 Finite and Asymptotic Properties

#### 3.3.1 Unbiasedness

We have an estimate,  $\mathbf{b}$ , of  $\boldsymbol{\beta}$ , but how “good” is our estimate? With assumptions (A1) through (A3), we can show in finite samples that the OLS estimate of  $\boldsymbol{\beta}$  is unbiased:

$$\begin{aligned}\mathbb{E}(\mathbf{b} \mid \mathbf{X}) &= \mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}) \\ &= \mathbb{E}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})) \\ &= \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbb{E}(\boldsymbol{\varepsilon} \mid \mathbf{X}) \\ &= \boldsymbol{\beta}\end{aligned}$$

Thus, conditional on the data matrix  $\mathbf{X}$ , on average the OLS estimate of the population regression vector *is* the population regression vector.

#### 3.3.2 Consistency

The strict exogeneity assumption is quite strong. Is there a weaker measure of how “good” an estimator is? Yes, consistency. If we maintain the rank condition assumption but weaken strict exogeneity to orthogonality, we can show that OLS is consistent:

$$\begin{aligned}\mathbf{b} - \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\varepsilon} \\ &= \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i'\right)^{-1} \left(\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i\varepsilon_i\right) \\ &= S_{xx}^{-1}\bar{\mathbf{g}} \\ &\xrightarrow{p} \Sigma_{xx}^{-1}\mathbf{0} \\ &= \mathbf{0}\end{aligned}$$

**Warning:** Finite sample properties hold for any sample size  $n$ . Asymptotic properties are valid only as  $n \rightarrow \infty$ . In practice, the latter never happens. Thus, estimation and inference

based on asymptotic theory is always “wrong.” That’s just something we have to live with as empiricists.

### 3.3.3 Normality

We can also show that OLS is asymptotically normal by using Slutsky’s Theorem:

$$\begin{aligned} \mathbf{S}_{xx}^{-1} &\xrightarrow{p} \boldsymbol{\Sigma}_{xx}^{-1} && \text{(Converges in probability)} \\ \sqrt{n}\bar{\mathbf{g}} &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{S}) && \text{(Converges in distribution)} \\ \sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) &= \mathbf{S}_{xx}^{-1}(\sqrt{n}\bar{\mathbf{g}}) \\ &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1}\mathbf{S}\boldsymbol{\Sigma}_{xx}^{-1}) \end{aligned}$$

To derive the OLS estimator through the method of moments, we relied on the assumption of strict exogeneity, which implied orthogonality between the error term and the regressors. What happens if the error term and the regressors are *not* orthogonal?

## 4 Instrumental Variables

### 4.1 Motivation: Simultaneity

#### 4.1.1 Working’s Example

Consider the following model of supply and demand:

$$\begin{aligned} q_i^s &= \beta_0 + \beta_1 p_i + v_i && \text{(Supply)} \\ q_i^d &= \alpha_0 + \alpha_1 p_i + u_i && \text{(Demand)} \\ q_i^s &= q_i^d && \text{(Equilibrium)} \end{aligned}$$

We assume that

$$\mathbb{E}(v_i) = 0$$

$$\mathbb{E}(u_i) = 0$$

$$\text{Cov}(u_i, v_i) = 0$$

Letting  $q \equiv q_i^s = q_i^d$ , we have

$$q = \beta_0 + \beta_1 p_i + v_i \quad (\text{Supply})$$

$$q = \alpha_0 + \alpha_1 p_i + u_i \quad (\text{Demand})$$

Recall that a regressor is **endogenous** if it is not predetermined, or not orthogonal to the error term:

$$\mathbb{E}(x_i \cdot \varepsilon_i) \neq 0$$

In the present example,  $p_i$  is necessarily endogenous in both equations. Why? Solving for  $(p_i, q_i)$ , we have

$$p_i = \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} - \frac{v_i - u_i}{\alpha_1 - \beta_1} \quad (4)$$

$$q_i = \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 v_i - \beta_1 u_i}{\alpha_1 - \beta_1} \quad (5)$$

From (4), we can already see that price is a function of both error terms. Therefore, it must be correlated with them. To see this, calculate the covariance of  $p_i$  with  $u_i$  and  $v_i$ . Recalling

that  $\mathbb{E}(u_i) = \mathbb{E}(v_i) = 0$ ,

$$\begin{aligned}\text{Cov}(p_i, u_i) &= \mathbb{E}(p_i \cdot u_i) = -\frac{\text{Var}(u_i)}{\alpha_1 - \beta_1} \\ \text{Cov}(p_i, v_i) &= \mathbb{E}(p_i \cdot v_i) = \frac{\text{Var}(v_i)}{\alpha_1 - \beta_1}\end{aligned}$$

In general, neither covariance is zero. Therefore,  $p_i$  is not orthogonal to the error term in either equation. Hence,  $p_i$  is **endogenous**. But what does that imply?

In the least squares projection of  $q_i$  on a constant and  $p_i$ , the coefficient of  $p_i$  is

$$\frac{\text{Cov}(p_i, q_i)}{\text{Var}(p_i)}$$

To rewrite this ratio in relation to the price effect in the demand curve ( $\alpha_1$ ), use the demand equation to calculate  $\text{Cov}(p_i, q_i)$ :

$$\begin{aligned}\text{Cov}(p_i, q_i) &= \text{Cov}(p_i, \alpha_0 + \alpha_1 p_i + u_i) \\ &= \alpha_1 \text{Var}(p_i) + \text{Cov}(p_i, u_i)\end{aligned}$$

The coefficient of  $p_i$  is then rewritten as

$$\begin{aligned}\frac{\text{Cov}(p_i, q_i)}{\text{Var}(p_i)} &= \frac{\alpha_1 \text{Var}(p_i) + \text{Cov}(p_i, u_i)}{\text{Var}(p_i)} \\ &= \alpha_1 + \frac{\text{Cov}(p_i, u_i)}{\text{Var}(p_i)} \\ &\Rightarrow \\ \text{plim of the OLS estimate of the price coefficient} - \alpha_1 &= \frac{\text{Cov}(p_i, u_i)}{\text{Var}(p_i)} \\ &\neq 0\end{aligned}$$

The same is true for estimating  $\beta_1$  with OLS. Thus, in general, OLS will not consistently

estimate  $\alpha_1$  due to **simultaneity** or **endogeneity bias**. In English, neither supply nor demand is consistently estimated in general because we cannot infer from data whether the change in  $(p_i, q_i)$  is due to a supply shift or a demand shift. As a mathematical exercise, note that when both curves can shift, the OLS estimate is consistent for a weighted average of  $\alpha_1$  and  $\alpha_2$ :

$$\begin{aligned}
\frac{\text{Cov}(p_i, q_i)}{\text{Var}(p_i)} &= \frac{\text{Cov}\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}, \frac{\alpha_1\beta_0 - \alpha_0\beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1v_i - \beta_1u_i}{\alpha_1 - \beta_1}\right)}{\text{Var}\left(\frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{v_i - u_i}{\alpha_1 - \beta_1}\right)} \\
&= \frac{1}{(\alpha_1 - \beta_1)^2} \text{Cov}(v_i - u_i, \alpha_1v_i - \beta_1u_i) \\
&= \frac{1}{(\alpha_1 - \beta_1)^2} \text{Var}(v_i - u_i) \\
&= \frac{\text{Cov}(v_i, \alpha_1v_i) + \text{Cov}(u_i, \beta_1u_i)}{\text{Var}(v_i) + \text{Var}(u_i)} \quad (\text{Errors assumed to be uncorrelated}) \\
&= \frac{\alpha_1 \text{Var}(v_i) + \beta_1 \text{Var}(u_i)}{\text{Var}(v_i) + \text{Var}(u_i)}
\end{aligned}$$

Might it be possible to estimate demand if some of the factors that shift supply are observable? Suppose the supply shifter,  $v_i$ , can be divided into an observable factor  $x_i$  and an unobservable factor  $\zeta_i$ ,  $\mathbb{E}(x_i \cdot \zeta_i) = 0$ :

$$\begin{aligned}
q_i &= \beta_0 + \beta_1 p_i + v_i \\
&= \beta_0 + \beta_1 p_i + (\beta_2 x_i + \zeta_i) \quad (\text{Supply})
\end{aligned}$$

Suppose that  $\mathbb{E}(x_i \cdot u_i) = 0$ . That is, the observed supply shifter does not affect demand. Then it should be possible to extract from price movements a component that is related to  $x_i$  but uncorrelated with  $u_i$ . We could then estimate demand by examining the relationship between  $q_i$  and  $x_i$ ; we trace out the demand curve through shifts in the supply curve.

For the equation in question, a predetermined variable that is correlated with  $p_i$  is an **in-**

**strumental variable.** In our example,  $x_i$  can serve as an instrument for  $p_i$  in the demand equation. To see this, solve Demand and Supply for  $(p_i, q_i)$ :

$$\begin{aligned} p_i &= \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} x_i + \frac{\zeta_i - u_i}{\alpha_1 - \beta_1} \\ q_i &= \frac{\alpha_1 \beta_0 - \alpha_0 \beta_1}{\alpha_1 - \beta_1} + \frac{\alpha_1 \beta_2}{\alpha_1 - \beta_1} x_i + \frac{\alpha_1 \zeta_i - \beta_1 u_i}{\alpha_1 - \beta_1} \end{aligned}$$

Since  $\text{Cov}(x_i, \zeta_i) = \text{Cov}(x_i, u_i) = 0$ , it follows that

$$\begin{aligned} \text{Cov}(x_i, p_i) &= \text{Cov}\left(x_i, \frac{\beta_0 - \alpha_0}{\alpha_1 - \beta_1} + \frac{\beta_2}{\alpha_1 - \beta_1} x_i + \frac{\zeta_i - u_i}{\alpha_1 - \beta_1}\right) \\ &= \frac{\beta_2}{\alpha_1 - \beta_1} \text{Var}(x_i) \\ &\neq 0 \end{aligned}$$

Use Demand to calculate  $\text{Cov}(x_i, q_i)$ :

$$\begin{aligned} \text{Cov}(x_i, q_i) &= \text{Cov}(x_i, \alpha_0 + \alpha_1 p_i + u_i) \\ &= \alpha_1 \text{Cov}(x_i, p_i) \end{aligned}$$

It follows that

$$\alpha_1 = \frac{\text{Cov}(x_i, q_i)}{\text{Cov}(x_i, p_i)}$$

and a natural estimator is

$$\hat{\alpha}_{1,IV} = \frac{\widehat{\text{Cov}}(x_i, q_i)}{\widehat{\text{Cov}}(x_i, p_i)}$$

Recall that  $x_i$  instrumented for  $p_i$ . If we did not have to do that, then the numerator and denominator in  $\hat{\alpha}_{1,IV}$  would be  $\widehat{\text{Cov}}(p_i, q_i)$  and  $\widehat{\text{Var}}(p_i)$ , respectively. But then  $\hat{\alpha}_{1,IV} = \hat{\alpha}_{1,OLS}$ .

**4.1.1.1 Preview of 2SLS** We can also consistently estimate  $\alpha_1$  through **two-stage least squares**, which we will discuss again later. In the first stage, the endogenous regressor  $p_i$  is regressed on a constant and the instrument,  $x_i$ , to obtain the fitted value  $\hat{p}_i$ . Then, in

the second stage, the dependent variable of interest,  $q_i$ , is regressed on a constant and  $\widehat{p}_i$ :

$$\widehat{\alpha}_{1,2SLS} = \frac{\widehat{\text{Cov}}(\widehat{p}_i, q_i)}{\widehat{\text{Var}}(\widehat{p}_i)}$$

To relate the second stage regression to the demand equation of interest, rewrite the demand equation:

$$\begin{aligned} q_i &= \alpha_0 + \alpha_1 p_i + u_i \\ &= \alpha_0 + [\alpha_1 \widehat{p}_i - \alpha_1 \widehat{p}_i] + \alpha_1 p_i + u_i \\ &= \alpha_0 + \alpha_1 \widehat{p}_i + [u_i + \alpha_1 (p_i - \widehat{p}_i)] \end{aligned}$$

The estimate of  $\alpha_1$  is consistent in this second stage regression. Suppose that  $\widehat{p}_i = \widehat{\mathbb{E}}^*(p_i | 1, x_i)$  exactly. Then  $u_i$  is uncorrelated with  $\widehat{p}_i$  because  $\mathbb{E}(x_i \cdot u_i) = 0$  and  $\widehat{p}_i = f(x_i)$ . Moreover,  $(p_i - \widehat{p}_i)$  is the least squares projection error and hence is uncorrelated with  $\widehat{p}_i$ . As we will mention later on, the standard errors from the second-stage regression are not correct.

#### 4.1.2 Haavelmo's Example

Consider the following (simple) macroeconomic model:

$$C_i = \alpha_0 + \alpha_1 Y_i + u_i \quad (\text{Consumption function})$$

$$Y_i = C_i + I_i \quad (\text{GNP identity})$$

$\alpha_1$  is the marginal propensity to consume (MPC) out of income. What happens if you try

to estimate  $\alpha_1$  through OLS?

$$\begin{aligned}
 \text{plim } \hat{\alpha}_{1,\text{OLS}} &= \frac{\text{Cov}(Y_i, C_i)}{\text{Var}(Y_i)} \\
 &= \frac{\text{Cov}(Y_i, \alpha_0 + \alpha_1 Y_i + u_i)}{\text{Var}(Y_i)} \\
 &= \frac{\alpha_1 \text{Var}(Y_i) + \text{Cov}(Y_i, u_i)}{\text{Var}(Y_i)} \\
 &= \alpha_1 + \frac{\text{Cov}(Y_i, u_i)}{\text{Var}(Y_i)} \\
 &\neq \alpha_1 \qquad \qquad \qquad (\text{If } \text{Cov}(Y_i, u_i) \neq 0)
 \end{aligned}$$

Just as before, the estimate is not consistent because of simultaneity bias. Solving for  $Y_i$  results in equilibrium GNP:

$$Y_i = \frac{\alpha_0}{1 - \alpha_1} + \frac{I_i}{1 - \alpha_1} + \frac{u_i}{1 - \alpha_1}$$

$Y_i$  is a function of  $u_i$ , therefore the two are correlated. But if  $\text{Cov}(I_i, u_i) = 0$ , then  $I_i$  can instrument for  $Y_i$ :

$$\begin{aligned}
 \text{Cov}(Y_i, u_i) &= \frac{1}{1 - \alpha_1} \text{Var}(u_i) > 0 && \text{(Endogenous)} \\
 \text{Cov}(Y_i, I_i) &= \frac{1}{1 - \alpha_1} \text{Var}(I_i) > 0 && \text{(Valid instrument)}
 \end{aligned}$$

## 4.2 Motivation: Errors-in-variables

Consider the permanent income hypothesis: “Permanent consumption,”  $C_i^*$  for household  $i$  is (directly) proportional to “permanent income,”  $Y_i^*$ :

$$C_i^* = kY_i^*$$

But measured consumption and measured income are error-ridden measures of permanent

consumption and permanent income:

$$C_i = C_i^* + c_i$$

$$Y_i = Y_i^* + y_i$$

The measurement errors are assumed to be zero mean, uncorrelated with each other, and uncorrelated with both permanent consumption and income. With substitutions, we can rewrite the permanent income hypothesis in terms of measured variables:

$$\begin{aligned} C_i &= kY_i + (c_i - ky_i) \\ &= kY_i + u_i \end{aligned}$$

There's no constant in the specification, so we compute the cross-moment  $\mathbb{E}(Y_i u_i)$ :

$$\begin{aligned} \mathbb{E}(Y_i u_i) &= \mathbb{E}(Y_i(c_i - ky_i)) \\ &= \mathbb{E}[(Y_i^* + y_i)(c_i - ky_i)] \\ &= -k\mathbb{E}(y_i^2) \\ &< 0 \end{aligned}$$

Hence, OLS will not produce consistent estimates of  $k$  due to measurement error. But we

can further show that OLS will produce estimates that are biased downward:

$$\begin{aligned}
 \text{plim } \hat{k}_{\text{OLS}} &= \frac{\mathbb{E}(C_i Y_i)}{\mathbb{E}(Y_i^2)} \\
 &= \frac{\mathbb{E}[(C_i^* + c_i)(Y_i^* + y_i)]}{\mathbb{E}[(Y_i^* + y_i)^2]} \\
 &= \frac{\mathbb{E}[(kY_i^* + c_i)(Y_i^* + y_i)]}{\mathbb{E}[(Y_i^* + y_i)^2]} \\
 &= \frac{k\mathbb{E}[(Y_i^*)^2]}{\mathbb{E}[(Y_i^*)^2] + \mathbb{E}(y_i^2)} \\
 &< k
 \end{aligned}$$

Suppose that we have a valid instrument  $x_i$  such that  $\mathbb{E}(x_i u_i) = 0$  and  $\mathbb{E}(x_i Y_i) \neq 0$ . We derive the instrumental variables estimator of  $k$ :

$$\begin{aligned}
 \mathbb{E}(x_i C_i) &= \mathbb{E}[x_i (kY_i + u_i)] \\
 &= k\mathbb{E}(x_i Y_i) \\
 k &= \frac{\mathbb{E}(x_i C_i)}{\mathbb{E}(x_i Y_i)} \\
 &\Rightarrow \\
 \hat{k} &= \frac{\sum_i x_i C_i}{\sum_i x_i Y_i}
 \end{aligned}$$

Somewhat amazingly, the instrument turns out to be ... 1, a constant.

## 4.3 Motivation: Omitted Variables Bias [Angrist and Pischke, 2008]

### 4.3.1 Wages and education

Suppose that potential outcomes can be written as follows:

$$\begin{aligned} Y_{si} &\equiv f_i(s) \\ &= \alpha + \rho s_i + \eta_i \\ &= \alpha + \rho s_i + (A_i' \gamma + \nu_i) \end{aligned}$$

where  $\gamma$  is again a vector of population regression coefficients, so that  $\nu_i$  and  $A_i$  are uncorrelated by construction.  $A_i$  (i.e. “ability”) is assumed to be the only reason why  $\eta_i$  and  $s_i$  are correlated so that  $\mathbb{E}[s_i \cdot \nu_i] = 0$ . But we cannot observe, or at least it is very difficult to estimate, “ability.” Hence, we cannot estimate  $\rho$  from the following equation due to omitted variables bias:

$$Y_{si} = \alpha + \rho s_i + A_i' \gamma + \nu_i$$

But suppose that there exists a variable  $z_i$ , an instrument, that is correlated with  $s_i$  but not correlated with either  $A_i$  or  $\nu_i$ . Following Hayashi [2000], we can estimate  $\rho$  with this variable:

$$\begin{aligned} Cov(y_i, z_i) &= Cov(\alpha + \rho s_i + A_i' \gamma + \nu_i, z_i) \\ &= \rho Cov(s_i, z_i) \\ \Rightarrow \rho &= \frac{Cov(y_i, z_i)}{Cov(s_i, z_i)} \end{aligned}$$

Thus, the correlation between the causal variable of interest,  $s_i$ , and the instrument,  $z_i$ , cannot be zero for the **instrumental variables** procedure to work. That is, the instrument must have a clear effect on  $s_i$  in the first stage. Moreover, the first stage is the *only* reason

for the relationship between  $y_i$  and  $z_i$  (i.e. the exclusion restriction).

### 4.3.2 Caveat: Are omitted variables always a “problem”?

The following is from John Rust. Suppose the “true” population model is

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

with  $\boldsymbol{\beta}_2 \neq \mathbf{0}$  and  $\mathbb{E}(\boldsymbol{\varepsilon} | \mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$ . Suppose further that due to a belief that  $\boldsymbol{\beta}_2 = \mathbf{0}$ , unobservability of  $\mathbf{X}_2$ , or carelessness, the researcher estimates the following model instead:

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{\xi}$$

The OLS estimator of  $\boldsymbol{\alpha}_1$  is

$$\hat{\boldsymbol{\alpha}}_1 = (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{y}$$

We can show that  $\boldsymbol{\alpha}_1$  is not an unbiased estimator of  $\boldsymbol{\beta}_1$ :

$$\begin{aligned}\mathbb{E}(\boldsymbol{\alpha}_1 | \mathbf{X}_1, \mathbf{X}_2) &= \mathbb{E}((\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{y} | \mathbf{X}_1, \mathbf{X}_2) \\ &= \mathbb{E}((\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}) | \mathbf{X}_1, \mathbf{X}_2) \\ &= \mathbb{E}(\boldsymbol{\beta}_1 + (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2\boldsymbol{\beta}_2 + (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\boldsymbol{\varepsilon} | \mathbf{X}_1, \mathbf{X}_2) \\ &= \boldsymbol{\beta}_1 + (\mathbf{X}'_1\mathbf{X}_1)^{-1}\mathbf{X}'_1\mathbf{X}_2\boldsymbol{\beta}_2 \\ &\neq \boldsymbol{\beta}_1\end{aligned}$$

unless  $\boldsymbol{\beta}_2 = \mathbf{0}$  or  $\mathbf{X}'_1\mathbf{X}_2 = \mathbf{0}$ . But we assumed that  $\boldsymbol{\beta}_2 \neq \mathbf{0}$ , and the second case is extremely special. What about consistency? We can continue assuming that  $\mathbb{E}(\boldsymbol{\varepsilon} | \mathbf{X}_1, \mathbf{X}_2) = \mathbf{0}$  or

weaken strict exogeneity to orthogonality:

$$\begin{aligned}
\hat{\alpha}_1 &= (\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbf{X}'_1 \mathbf{y} \\
&= \left( \sum_{i=1}^n \mathbf{x}_{1i} \mathbf{x}'_{1i} \right)^{-1} \sum_{i=1}^n \mathbf{x}_{1i} y_i \\
&\xrightarrow{p} \mathbb{E}(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbb{E}(\mathbf{X}'_1 \mathbf{y}) \\
&= \mathbb{E}(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbb{E}(\mathbf{X}'_1 (\mathbf{X}_1 \boldsymbol{\beta}_1 + \mathbf{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon})) \\
&= \boldsymbol{\beta}_1 + \mathbb{E}(\mathbf{X}'_1 \mathbf{X}_1)^{-1} \mathbb{E}(\mathbf{X}'_1 \mathbf{X}_2) \boldsymbol{\beta}_2 \\
&\neq \boldsymbol{\beta}_1
\end{aligned}$$

unless  $\boldsymbol{\beta}_2 = \mathbf{0}$  (violates assumption) or  $\mathbb{E}(\mathbf{X}'_1 \mathbf{X}_2) = \mathbf{0}$ . But the second possibility is quite plausible, as it means that  $\hat{\alpha}_1$  is a consistent estimator for  $\boldsymbol{\beta}_1$  if  $\mathbf{X}_1$  and  $\mathbf{X}_2$  are uncorrelated. Recall in the previous example of wages and education that an OLS estimate of  $\rho$  given the unobservability of  $A'_i$  was not even consistent. Thus, an omitted variable is not necessarily “problematic” *asymptotically* if it is uncorrelated with the other regressors, and the only omitted variables we care about in practice are those that are correlated with the other regressors.

## 4.4 Properties

### 4.4.1 Consistency

Strict exogeneity implies orthogonality, therefore non-orthogonality implies non-strict exogeneity. So we lose the unbiasedness property. What about consistency? If the orthogonality assumption is violated, then OLS is not even consistent. What can be done? *If* we can find a predetermined (i.e. orthogonal) variable that is correlated with the endogenous regressor for each endogenous regressor, then through **instrumental variables** we are saved.

Suppose we have a vector of instruments,  $\mathbf{x}_i$ , which are orthogonal to the error term.  $\mathbf{z}_i$

denotes the vector of regressors. The previous assumption of regressor orthogonality is thus modified to *instrument* orthogonality. (Note that an exogenous regressor is an instrument;  $\mathbf{x}_i \cap \mathbf{z}_i$  need not be  $\emptyset$ .)

By modifying accordingly the rank condition assumption, we can derive the instrumental variables estimator through the method of moments and show that it is consistent:

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i &= \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{z}_i' \tilde{\boldsymbol{\beta}}) \\
&= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) - \tilde{\boldsymbol{\beta}} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{z}_i' \right) \\
&= 0 \\
\Rightarrow \tilde{\boldsymbol{\beta}} \equiv \hat{\boldsymbol{\beta}}_{IV} &= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{z}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i y_i \right) \\
&= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{z}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i (\mathbf{z}_i' \boldsymbol{\beta} + \varepsilon_i) \right) \\
&= \boldsymbol{\beta} + \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{z}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i \right) \\
\Rightarrow \hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta} &= \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \mathbf{z}_i' \right)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i \right) \\
&\xrightarrow{p} \boldsymbol{\Sigma}_{xz}^{-1} \mathbf{0} \\
&= \mathbf{0}
\end{aligned}$$

#### 4.4.2 Normality

We again use Slutsky's Theorem to prove asymptotic normality:

$$\begin{aligned}
 \mathbf{S}_{xz}^{-1} &\xrightarrow{p} \boldsymbol{\Sigma}_{xz}^{-1} && \text{(Converges in probability)} \\
 \sqrt{n}\bar{\mathbf{g}} &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{S}) && \Rightarrow \text{(Converges in distribution)} \\
 \sqrt{n}(\hat{\boldsymbol{\beta}}_{IV} - \boldsymbol{\beta}) &= \mathbf{S}_{xz}^{-1} \sqrt{n}\bar{\mathbf{g}} \\
 &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xz}^{-1} \mathbf{S} \boldsymbol{\Sigma}_{xz}^{-1})
 \end{aligned}$$

#### 4.5 Forbidden Regressions [Angrist and Pischke, 2008]

A forbidden regression involves the direct application of 2SLS to a nonlinear model. Consider the following model with  $D_i$  as an endogenous dummy variable:

$$y_i = \boldsymbol{\alpha}' \mathbf{X}_i + \rho D_i + \eta_i$$

Suppose further that we have a vector of instruments,  $\mathbf{Z}_i$ , for  $D_i$ , leading to the first-stage regression:

$$D_i = \boldsymbol{\pi}'_{1,0} \mathbf{X}_i + \boldsymbol{\pi}'_{1,1} \mathbf{Z}_i + \xi_{1,i}$$

But the conditional expectation function (CEF) of the first stage,  $\mathbb{E}(D_i | \mathbf{X}_i, \mathbf{Z}_i)$ , is probably nonlinear because  $D_i$  is dichotomous. Suppose we use a nonlinear first-stage regression, for example, a probit. The fitted values of  $D_i$  are then

$$\hat{D}_i = \Phi[\boldsymbol{\pi}'_{1,0} \mathbf{X}_i + \boldsymbol{\pi}'_{1,1} \mathbf{Z}_i]$$

The “forbidden” regression is the second stage of  $y_i$  on  $\hat{D}_i$ :

$$y_i = \boldsymbol{\alpha}' \mathbf{X}_i + \rho \hat{D}_i + [\eta_i + \rho(D_i - \hat{D}_i)]$$

Only OLS estimation of the first stage is guaranteed to produce residuals that are uncorrelated with fitted values and covariates. To get around this problem, use the  $\widehat{D}_i$ 's as instruments for  $D_i$ .

#### 4.5.1 Example: Founder-CEO Endogeneity [Adams et al., 2009]

Adams et al. [2009] use the following linear specification as a benchmark:

$$y_{i,t} = \alpha + \gamma f_{i,t} + \beta \mathbf{x}_{i,t} + u_{i,t}$$

where  $f_{i,t}$  is a binary random variable that takes the value of 1 if the CEO is one of the founders and zero otherwise. Adams et al. argue that there may be simultaneity between firm performance and founder-CEO status. They therefore use “dead founders” and “number of founders” as instruments for founder-CEO status, which would lead to the following first-stage regression:

$$f_{i,t} = \alpha + \beta \mathbf{x}_{i,t} + \delta_1 z_{i,1,t} + \delta_2 z_{i,2,t} + \xi_{i,t}$$

But since  $f_{i,t}$  is dichotomous, the CEF  $\mathbb{E}(f_{i,t} | \mathbf{x}_{i,t}, \mathbf{z}_{i,t})$  is likely nonlinear. If we use a probit to estimate the first stage, the fitted values will be

$$\widehat{f}_{i,t} = \Phi(\alpha + \beta \mathbf{x}_{i,t} + \delta_1 z_{i,1,t} + \delta_2 z_{i,2,t})$$

As stated before, only OLS is guaranteed to produce first-stage residuals that are uncorrelated with the fitted values and covariates. We apply the same solution: don't use  $\mathbf{z}_{i,t}$  as instruments for  $f_{i,t}$ ; rather, use the  $\widehat{f}_{i,t}$ 's as instruments for  $f_{i,t}$ .

Hence, there are three steps and not just two:

1. Estimate a probit of the determinants of founder-CEO status and obtain the fitted values,  $\widehat{f}_{i,t}$ .

2. Regress  $f_{i,t}$  on  $\widehat{f}_{i,t}$  and  $\mathbf{x}_{i,t}$ , not on  $\mathbf{z}_{i,t}$  and  $\mathbf{x}_{i,t}$ .
3. Regress  $y_{i,t}$  on  $\mathbf{x}_{i,t}$  and the fitted values of the second stage ( $\widehat{y}_{i,t}$ ).

## 5 The Generalized Method of Moments: Single-Equation

What if we had more than one instrument for each endogenous regressor? In that case, there are more orthogonality conditions than parameters,  $K > L$ , and so the system may not have a solution. Thus we cannot always set

$$\begin{aligned} \mathbf{g}_n(\tilde{\boldsymbol{\delta}}) &\equiv \frac{1}{n} \sum_{i=1}^n \mathbf{x}_i (y_i - \mathbf{z}_i' \tilde{\boldsymbol{\delta}}) \\ &= \mathbf{s}_{xy} - \mathbf{S}_{xz} \tilde{\boldsymbol{\delta}} \\ &= 0 \end{aligned}$$

Instead, we choose  $\tilde{\boldsymbol{\delta}}$  so that  $\mathbf{g}_n(\tilde{\boldsymbol{\delta}})$  is as close to  $\mathbf{0}$  as possible:

$$\begin{aligned} \widehat{\boldsymbol{\delta}}(\widehat{\mathbf{W}}) &\equiv \arg \min_{\tilde{\boldsymbol{\delta}}} J(\tilde{\boldsymbol{\delta}}, \widehat{\mathbf{W}}) \\ &= \arg \min_{\tilde{\boldsymbol{\delta}}} n \cdot \mathbf{g}_n(\tilde{\boldsymbol{\delta}})' \widehat{\mathbf{W}} \mathbf{g}_n(\tilde{\boldsymbol{\delta}}) \\ &= \arg \min_{\tilde{\boldsymbol{\delta}}} n \cdot (\mathbf{s}_{xy} - \mathbf{S}_{xz} \tilde{\boldsymbol{\delta}})' \widehat{\mathbf{W}} (\mathbf{s}_{xy} - \mathbf{S}_{xz} \tilde{\boldsymbol{\delta}}) \end{aligned}$$

where  $\widehat{\mathbf{W}}_{K \times K}$  is symmetric and positive definite and converges to  $\mathbf{W}$ . Taking the FOC results in the GMM estimator:

$$\widehat{\boldsymbol{\delta}}(\widehat{\mathbf{W}}) = (\mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{s}_{xy}$$

## 5.1 OLS and IV as GMM estimators

If  $K = L$ , then the  $\mathbf{S}_{xz}$  matrix is square and combined with the rank condition assumption, invertible. It follows that the GMM estimator reduces to  $\mathbf{S}_{xz}^{-1} \mathbf{s}_{xy}$ , which is none other than the instrumental variables estimator. Moreover, if the instruments *are* the regressors, then the instrumental variables estimator reduces to OLS.

## 5.2 Sampling Error and Consistency

Begin with the estimation equation:

$$y_i = \mathbf{z}'_i \boldsymbol{\delta} + \varepsilon_i$$

Multiply both sides from the left by  $\mathbf{x}_i$  and take averages:

$$\begin{aligned} \frac{1}{n} \sum_i \mathbf{x}_i y_i &= \frac{1}{n} \sum_i \mathbf{x}_i \mathbf{z}'_i \boldsymbol{\delta} + \frac{1}{n} \sum_i \mathbf{x}_i \varepsilon_i \\ &\Rightarrow \\ \mathbf{s}_{xy} &= \mathbf{S}_{xz} \boldsymbol{\delta} + \bar{\mathbf{g}} \end{aligned}$$

Plug the last expression into the GMM estimator:

$$\begin{aligned} \hat{\boldsymbol{\delta}}(\widehat{\mathbf{W}}) &= (\mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{W}} (\mathbf{S}_{xz} \boldsymbol{\delta} + \bar{\mathbf{g}}) \\ &= \boldsymbol{\delta} + (\mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{W}} \bar{\mathbf{g}} \\ &\Rightarrow \\ \hat{\boldsymbol{\delta}}(\widehat{\mathbf{W}}) - \boldsymbol{\delta} &= (\mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{W}} \bar{\mathbf{g}} \end{aligned}$$

To see that this converges in probability to zero,

$$\begin{aligned} [\widehat{\boldsymbol{\delta}}(\widehat{\mathbf{W}}) - \boldsymbol{\delta}] &= (\mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{W}}(\widehat{\mathbf{g}}) \\ &\xrightarrow{p} (\boldsymbol{\Sigma}'_{xz} \mathbf{W} \boldsymbol{\Sigma}_{xz})^{-1} \boldsymbol{\Sigma}'_{xz} \mathbf{W} \mathbf{0} \\ &= \mathbf{0} \end{aligned}$$

To show asymptotic normality, use Slutsky's Theorem:

$$\begin{aligned} (\mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{W}} &\xrightarrow{p} (\boldsymbol{\Sigma}'_{xz} \mathbf{W} \boldsymbol{\Sigma}_{xz})^{-1} \boldsymbol{\Sigma}'_{xz} \mathbf{W} && \text{(Converges in probability)} \\ \sqrt{n} \widehat{\mathbf{g}} &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \mathbf{S}) && \text{(Converges in distribution)} \\ \sqrt{n} [\widehat{\boldsymbol{\delta}}(\widehat{\mathbf{W}}) - \boldsymbol{\delta}] &= (\mathbf{S}_{xz}' \widehat{\mathbf{W}} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{W}}(\sqrt{n} \widehat{\mathbf{g}}) \\ &\xrightarrow{D} \mathcal{N}(\mathbf{0}, (\boldsymbol{\Sigma}'_{xz} \mathbf{W} \boldsymbol{\Sigma}_{xz})^{-1} \boldsymbol{\Sigma}'_{xz} \mathbf{W} \mathbf{S} \mathbf{W} \boldsymbol{\Sigma}_{xz} (\boldsymbol{\Sigma}'_{xz} \mathbf{W} \boldsymbol{\Sigma}_{xz})^{-1}) \\ &&& \text{(Slutsky's Theorem)} \end{aligned}$$

### 5.3 Efficient GMM, or “Making the formulas look nicer”

Under the assumption that  $\mathbb{E}[(x_{ik}z_{il})^2]$  exists and is finite for all  $k \in \mathcal{K}$  and  $l \in \mathcal{L}$ , a consistent estimator of  $\mathbf{S}$  is

$$\widehat{\mathbf{S}} = \frac{1}{n} \sum_i \widehat{\varepsilon}_i^2 \mathbf{x}_i \mathbf{x}'_i$$

A GMM estimator satisfying the efficiency condition that  $\text{plim } \widehat{\mathbf{W}} = \mathbf{W} = \mathbf{S}^{-1}$  is the **efficient GMM estimator**:

$$\begin{aligned} \widehat{\boldsymbol{\delta}}(\widehat{\mathbf{S}}^{-1}) &= (\mathbf{S}_{xz}' \widehat{\mathbf{S}}^{-1} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \widehat{\mathbf{S}}^{-1} \mathbf{s}_{xy} \\ \text{Avar}(\widehat{\boldsymbol{\delta}}(\widehat{\mathbf{S}}^{-1})) &= (\boldsymbol{\Sigma}'_{xz} \mathbf{S}^{-1} \boldsymbol{\Sigma}_{xz})^{-1} \boldsymbol{\Sigma}'_{xz} \mathbf{S}^{-1} \mathbf{S} \mathbf{S}^{-1} \boldsymbol{\Sigma}_{xz} (\boldsymbol{\Sigma}'_{xz} \mathbf{S}^{-1} \boldsymbol{\Sigma}_{xz})^{-1} \\ &= (\boldsymbol{\Sigma}'_{xz} \mathbf{S}^{-1} \boldsymbol{\Sigma}_{xz})^{-1} \end{aligned}$$

### 5.3.1 Imposing Conditional Homoskedasticity

If we assume that  $\mathbb{E}(\varepsilon_i^2 | \mathbf{x}_i) = \sigma^2$ , then we no longer need the finite fourth-moment assumption. Instead, the matrix of fourth moments ( $\mathbf{S}$ ) can be written as a product of second moments:

$$\begin{aligned}
 \mathbf{S} &= \mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i') \\
 &= \mathbb{E}[\mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i' | \mathbf{x}_i)] && \text{(Law of Iterated Expectations)} \\
 &= \mathbb{E}[\mathbb{E}(\varepsilon_i^2 | \mathbf{x}_i) \mathbf{x}_i \mathbf{x}_i'] \\
 &= \mathbb{E}[\sigma^2 \mathbf{x}_i \mathbf{x}_i'] \\
 &= \sigma^2 \mathbb{E}(\mathbf{x}_i \mathbf{x}_i') \\
 &= \sigma^2 \boldsymbol{\Sigma}_{xx}
 \end{aligned}$$

$\mathbf{S}$  can be estimated by

$$\begin{aligned}
 \widehat{\mathbf{S}} &= \widehat{\sigma}^2 \frac{1}{n} \sum_i \mathbf{x}_i \mathbf{x}_i' \\
 &= \widehat{\sigma}^2 \mathbf{S}_{xx}
 \end{aligned}$$

As a consequence of imposing conditional homoskedasticity, the formulas reduce to the Two-Stage Least Squares (2SLS) estimator:

$$\begin{aligned}
 \widehat{\boldsymbol{\delta}}(\widehat{\mathbf{S}}^{-1}) &= (\mathbf{S}_{xz}' (\widehat{\sigma}^2 \mathbf{S}_{xx})^{-1} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' (\widehat{\sigma}^2 \mathbf{S}_{xx})^{-1} \mathbf{s}_{xy} \\
 &= (\mathbf{S}_{xz}' \mathbf{S}_{xx}^{-1} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \mathbf{S}_{xx}^{-1} \mathbf{s}_{xy} \\
 &\equiv \widehat{\boldsymbol{\delta}}_{2SLS} \\
 \text{Avar}(\widehat{\boldsymbol{\delta}}_{2SLS}) &= \sigma^2 (\boldsymbol{\Sigma}'_{xz} \boldsymbol{\Sigma}_{xx}^{-1} \boldsymbol{\Sigma}_{xz})^{-1}
 \end{aligned}$$

And so the sample variance of the 2SLS residuals is

$$\hat{\sigma}^2 \equiv \frac{1}{n} \sum_i (y_i - \mathbf{z}'_i \hat{\boldsymbol{\delta}}_{2SLS})^2$$

### 5.3.2 2SLS as an IV estimator or as Two Separate Regressions

Define the following:

$$\mathbf{X}_{n \times K} = \begin{bmatrix} \mathbf{x}'_1 \\ \vdots \\ \mathbf{x}'_n \end{bmatrix}, \mathbf{Z}_{n \times L} = \begin{bmatrix} \mathbf{z}'_1 \\ \vdots \\ \mathbf{z}'_n \end{bmatrix}, \mathbf{y}_{n \times 1} = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}$$

Then the 2SLS estimator can be written as

$$\begin{aligned} \hat{\boldsymbol{\delta}}(\hat{\mathbf{S}}^{-1}) &= (\mathbf{S}_{xz}' \mathbf{S}_{xx}^{-1} \mathbf{S}_{xz})^{-1} \mathbf{S}_{xz}' \mathbf{S}_{xx}^{-1} \mathbf{S}_{xy} \\ &= (\underbrace{\mathbf{Z}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{Z}}_{\text{Projection matrix}})^{-1} \mathbf{Z}' \mathbf{X} (\mathbf{X}' \mathbf{X})^{-1} \mathbf{X}' \mathbf{y} \\ &= (\mathbf{Z}' \mathbf{P} \mathbf{Z})^{-1} \mathbf{Z}' \mathbf{P} \mathbf{y} \end{aligned}$$

2SLS can be written as an IV estimator. Suppose that  $K = L$ , and define

$$\hat{\mathbf{Z}}_{n \times L} = \begin{bmatrix} \hat{\mathbf{z}}'_1 \\ \vdots \\ \hat{\mathbf{z}}'_n \end{bmatrix}$$

The IV estimator with  $\hat{\mathbf{z}}_i$  serving as the instruments instead of  $\mathbf{x}_i$  is

$$\begin{aligned} \hat{\boldsymbol{\delta}}_{IV} &= \left( \frac{1}{n} \sum_i \hat{\mathbf{z}}_i \mathbf{z}'_i \right)^{-1} \frac{1}{n} \sum_i \hat{\mathbf{z}}_i y_i \\ &= (\hat{\mathbf{Z}}' \mathbf{Z})^{-1} \hat{\mathbf{Z}}' \mathbf{y} \end{aligned}$$

The  $l$ -th instrument is the fitted value from regressing the  $l$ -th regressor,  $z_{il}$ , on  $\mathbf{x}_i$  with OLS. The  $n \times L$  data matrix of instruments is thus

$$\begin{aligned}\widehat{\mathbf{Z}} &= \mathbf{X}\mathbf{b} && \text{(Fitted values (cf. } \widehat{\mathbf{y}} = \mathbf{X}\mathbf{b}\text{))} \\ &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} && \text{(Z is y)} \\ &= \mathbf{P}\mathbf{Z}\end{aligned}$$

If we plug this back into  $\widehat{\boldsymbol{\delta}}_{IV}$ , we have the 2SLS estimator again. As the previous discussion suggested, as well as the very name of the estimator, 2SLS can be interpreted as two OLS regressions.

1. Regress the  $L$  regressors on  $\mathbf{x}_i$  with OLS and obtain fitted values  $\widehat{\mathbf{z}}_i$ :

$$\begin{aligned}\widehat{\mathbf{Z}} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Z} \\ &= \mathbf{P}\mathbf{Z}\end{aligned}$$

2. Regress  $y_i$  on those fitted values with OLS. The coefficient estimate is

$$\begin{aligned}\widehat{\boldsymbol{\delta}}_{2SLS} &= (\widehat{\mathbf{Z}}'\widehat{\mathbf{Z}})^{-1}\widehat{\mathbf{Z}}'\mathbf{y} \\ &= (\mathbf{Z}'\mathbf{P}'\mathbf{P}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}'\mathbf{y} \\ &= (\mathbf{Z}'\mathbf{P}\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{P}\mathbf{y} && \text{(P is symmetric and idempotent)}\end{aligned}$$

But be careful! In the second-stage regression, statistical software does not “know” you are doing 2SLS; it only “knows” that you ran one regression with OLS. So the standard errors calculated are based on the residual vector

$$\mathbf{y} - \widehat{\mathbf{Z}}\widehat{\boldsymbol{\delta}}_{2SLS}$$

That's not what you want. You want the standard errors to be calculated based on the residual vector

$$\mathbf{y} - \mathbf{Z}\hat{\boldsymbol{\delta}}_{2SLS}$$

After all, you were estimating the following equation, right?

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\delta} + \boldsymbol{\varepsilon}$$

## 6 Panel Data and Clustering

**Warning:** As is the case with anything that I teach myself, the following reflects my understanding and may contain errors. In particular, Cameron and Trivedi [2005, pg. 832] specifically note that a difference between the panel case and the cluster case is that in the former, the individual unit of analysis is observed more than once while in the latter, the individual unit of analysis is observed *only* once. In Wooldridge [2010, pg. 878], however, it seems like the cluster case allows for repeated observations on the individual unit of analysis.

### 6.1 OLS

Assume that we have independent, identically distributed cross section observations  $\{(\mathbf{X}_i, \mathbf{y}_i) : i \in \mathcal{N}\}$ , where  $\mathbf{X}_i$  is a  $T \times K$  matrix<sup>1</sup>

$$\begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iT} \end{bmatrix}$$

with each  $\mathbf{x}_{ij}, j \in \mathcal{T}$  is a  $1 \times K$  row vector, and  $\mathbf{y}_i$  is a  $T \times 1$  vector [Wooldridge, 2010].

---

<sup>1</sup>That is, we have  $n$  units (e.g. firms), each of which have  $T$  observations (e.g. years), and  $K$  regressors.

### 6.1.1 Clustering

The concern is that some aspects of the population regression model vary by “cluster”  $c, c \in \mathcal{C}$  [Cameron and Trivedi, 2005]. Suppose that the  $i$ th unit in the overall sample is the  $j$ th unit in the  $c$ th cluster <sup>2</sup>. A general model for clustered data is

$$\begin{aligned} y_{j,c} &= \mathbf{x}'_{j,c} \boldsymbol{\beta}_c + u_{j,c} \\ j &= 1, \dots, N_c \\ c &= 1, \dots, C \end{aligned}$$

where  $\text{Cov}[u_{j,c}, u_{k,c}] \neq 0$  but  $\text{Cov}[u_{j,c}, u_{k,d}] = 0$  for  $c \neq d$ . That is, errors *within* clusters may be correlated, but errors *across* clusters are not. The correlation within clusters could arise from a cluster-specific effect,  $\alpha_c$ :

$$y_{j,c} = \mathbf{x}'_{j,c} \boldsymbol{\beta} + (\alpha_c + \varepsilon_{j,c})$$

We can consistently estimate  $\boldsymbol{\beta}$  with OLS if we assume that the cluster-specific effect is uncorrelated with  $\mathbf{x}_{j,c}$  <sup>3</sup>.

*Proof.* Stack the observations within a cluster to yield

$$\mathbf{y}_c = \mathbf{X}_c \boldsymbol{\beta} + \mathbf{u}_c$$

where  $\mathbf{y}_c, \mathbf{u}_c$  are  $N_c T \times 1$  and  $\mathbf{X}_c$  is  $N_c T \times K$  <sup>4</sup>. If the cluster-specific effect is uncorrelated with the (observed) regressors, then we use the method of moments procedure to derive the

---

<sup>2</sup>That is, a cluster  $c$  could contain several different  $i$ 's.

<sup>3</sup>This is equivalent to treating the cluster-specific effect as an unobserved, uncorrelated omitted variable.

<sup>4</sup>We can further stack the clusters to yield

$$\mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{u}$$

where  $\mathbf{y}, \mathbf{u}$  are  $NT \times 1$  and  $\mathbf{X}$  is  $NT \times K$  because  $\sum_{c=1}^C N_c T = NT$ .

OLS estimator, just as before:

$$\begin{aligned}
\mathbb{E}(\mathbf{X}'_c \mathbf{u}_c) &= 0 && \text{(Orthogonality)} \\
\Rightarrow \\
\frac{1}{C} \sum_{c=1}^C \mathbf{X}'_c (\mathbf{y}_c - \mathbf{X}_c \boldsymbol{\beta}) &= 0 && \text{(Sample analogue)} \\
\Rightarrow \\
\widehat{\boldsymbol{\beta}}_{OLS} &= \left( \frac{1}{C} \sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1} \left( \frac{1}{C} \sum_{c=1}^C \mathbf{X}'_c \mathbf{y}_c \right)
\end{aligned}$$

□

From the sampling error, we derive the asymptotic distribution of the OLS estimator:

$$\begin{aligned}
\sqrt{C}(\widehat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1} \mathbf{S} \boldsymbol{\Sigma}_{xx}^{-1}) \\
\boldsymbol{\Sigma}_{xx}^{-1} &= \mathbb{E}(\mathbf{X}'_c \mathbf{X}_c)^{-1} \\
\mathbf{S} &= \mathbb{E}(\mathbf{X}'_c \mathbf{u}_c \mathbf{u}'_c \mathbf{X}_c)
\end{aligned}$$

It follows that a consistent estimator of the asymptotic variance is given by

$$\widehat{\text{Avar}}[\widehat{\boldsymbol{\beta}}_{OLS}] = \left( \sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1} \sum_{c=1}^C \mathbf{X}'_c \widehat{\mathbf{u}}_c \widehat{\mathbf{u}}'_c \mathbf{X}_c \left( \sum_{c=1}^C \mathbf{X}'_c \mathbf{X}_c \right)^{-1}$$

which is **cluster-robust** assuming independence over  $c$  and  $C \rightarrow \infty$ .

### 6.1.2 Special Case: The individual unit of analysis is the cluster

As previously mentioned, one cluster  $c$  could contain several different  $i$ 's. But each  $i$  could be its own cluster, such that cluster  $c \in \mathcal{C}$  is  $i \in \mathcal{N}$ . The multivariate linear model for a

random draw from the population can be expressed as

$$\mathbf{y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{u}_i$$

where as before  $\mathbf{X}_i$  is a  $T \times K$  matrix and  $\mathbf{y}_i$  is a  $T \times 1$  vector [Wooldridge, 2010]. In the notation of the previous section,  $N_c = 1$ , and the cluster-specific effect  $\alpha_c$  is a unit-specific effect  $\alpha_i$ . That is, errors may be correlated *within* the same unit of observation  $i$  but not *across* units of observation,  $i \neq j$ .

To consistently estimate  $\boldsymbol{\beta}$ , we require that

$$\mathbb{E}(\mathbf{X}_i'\mathbf{u}_i) = \mathbf{0}$$

It follows from the method of moments procedure that

$$\hat{\boldsymbol{\beta}} = \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i'\mathbf{X}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i'\mathbf{y}_i \right)$$

Since

$$\begin{aligned} \mathbf{X}_i &= \begin{bmatrix} \mathbf{x}_{i1} \\ \vdots \\ \mathbf{x}_{iT} \end{bmatrix} \\ \mathbf{X}_i' &= \begin{bmatrix} \mathbf{x}'_{i1} & \cdots & \mathbf{x}'_{iT} \end{bmatrix} \\ \mathbf{X}_i'\mathbf{X}_i &= \sum_{t=1}^T \mathbf{x}'_{it}\mathbf{x}_{it} \end{aligned}$$

We can therefore rewrite the expression for the “system ordinary least squares (SLOS) esti-

mator” of  $\beta$  as

$$\begin{aligned}\widehat{\beta} &= \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}'_i \mathbf{y}_i \right) \\ &= \left( \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}'_{it} \mathbf{x}_{it} \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \sum_{t=1}^T \mathbf{x}'_{it} y_{it} \right)\end{aligned}$$

From the sampling error

$$\widehat{\beta} - \beta = \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}'_i \mathbf{u}_i \right)$$

we have

$$\begin{aligned}\sqrt{N} (\widehat{\beta} - \beta) &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \Sigma_{xx}^{-1} \mathbf{S} \Sigma_{xx}^{-1}) \\ \mathbf{S} &\equiv \mathbb{E}(\mathbf{X}'_i \mathbf{u}_i \mathbf{u}'_i \mathbf{X}_i)\end{aligned}$$

A consistent estimator of the asymptotic variance is given by

$$\widehat{\text{Avar}}(\widehat{\beta}) = \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1} \sum_{i=1}^N \mathbf{X}'_i \widehat{\mathbf{u}}_i \widehat{\mathbf{u}}'_i \mathbf{X}_i \left( \sum_{i=1}^N \mathbf{X}'_i \mathbf{X}_i \right)^{-1}$$

which is **panel-robust** with the individual identifier (e.g. GVKEY) as the **cluster variable** assuming independence over  $i$  and  $N \rightarrow \infty$ .

### 6.1.3 A more special case: Plain ol’ heteroskedasticity-robust standard errors

We have presented consistent estimates of the asymptotic variance for the cluster case in general and for the special case when the cluster is the unit of observation. In the former case, we assumed independence over the clusters, and in the latter case, we assumed independence over the units of observation. What if we assume independence over the observations, as in the classic cross-sectional treatment [Cameron and Trivedi, 2005, pg. 76]?

As we derived earlier,

$$\begin{aligned}\sqrt{n}(\mathbf{b} - \boldsymbol{\beta}) &= \mathbf{S}_{xx}^{-1}(\sqrt{n}\bar{\mathbf{g}}) \\ &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{xx}^{-1} \mathbf{S} \boldsymbol{\Sigma}_{xx}^{-1}) \\ \mathbf{S} &\equiv \mathbb{E}(\varepsilon_i^2 \mathbf{x}_i \mathbf{x}_i')\end{aligned}$$

While the consistent estimator of the asymptotic variance is robust to heteroskedasticity, because we assumed independence over the observations, the estimator is **not** robust to serial correlation. I believe this is the “common error” made in pooled regression noted by Cameron and Trivedi [2005, pg. 707] (i.e. to estimate an OLS regression using the standard (i.e. plain ol’) robust standard error option).

## 6.2 Random and Fixed Effects

For a randomly drawn cross section observation  $i$ , the unobserved effects model is

$$\begin{aligned}y_{i,t} &= \mathbf{x}_{i,t} \boldsymbol{\beta} + c_i + u_{i,t} \\ t &\in \mathcal{T}\end{aligned}$$

The key issue is whether  $c_i$ , which may be unobservable, is uncorrelated with the observed explanatory variables,  $\mathbf{x}_{i,t}$ . If we assume that  $c_i$  is mean independent of the observed explanatory variables,

$$\mathbb{E}(c_i | \mathbf{x}_{i,1} \dots \mathbf{x}_{i,T}) = \mathbb{E}(c_i)$$

then we are using the **random effects framework**. If we allow for arbitrary dependence between  $c_i$  and  $\mathbf{x}_{i,t}$ ,

$$\text{Cov}(x_j, c) \neq 0 \text{ for some } j$$

then we are using the **fixed effects framework**.

### 6.2.1 Random Effects

We assume that<sup>5</sup>

$$\begin{aligned}\mathbb{E}(u_{i,t} | \mathbf{x}_i, c_i) &= 0 \\ \mathbb{E}(c_i | \mathbf{x}_i) &= \mathbb{E}(c_i)\end{aligned}$$

We write

$$\begin{aligned}y_{i,t} &= \mathbf{x}_{i,t}\boldsymbol{\beta} + v_{i,t} \\ &= \mathbf{x}_{i,t}\boldsymbol{\beta} + (c_i + u_{i,t}) \\ \mathbb{E}(v_{i,t} | \mathbf{x}_i) &= 0\end{aligned}$$

For all time periods, we have

$$\begin{aligned}\mathbf{y}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{v}_i \\ &= \mathbf{X}_i\boldsymbol{\beta} + (c_i\mathbf{j}_T + \mathbf{u}_i)\end{aligned}$$

We define the unconditional variance matrix of  $\mathbf{v}_i$  as follows and assume that it is equal to the conditional variance matrix:

$$\begin{aligned}\boldsymbol{\Omega} &\equiv \mathbb{E}(\mathbf{v}_i\mathbf{v}_i') \\ &= \mathbb{E}(\mathbf{v}_i\mathbf{v}_i' | \mathbf{x}_i)\end{aligned}$$

---

<sup>5</sup>As we will see later, the second assumption distinguishes random effects from fixed effects.

We add additional assumptions on the idiosyncratic errors,  $u_{i,t}$ , to give  $\mathbf{\Omega}$  a special form:

$$\mathbb{E}(u_{i,t}^2) = \sigma_u^2 \quad (\text{Constant unconditional idiosyncratic variance})$$

$$\mathbb{E}(u_{i,t}u_{i,s}) = 0 \quad (\text{No idiosyncratic serial correlation})$$

With these assumptions, it follows that

$$\mathbb{E}(v_{i,t}^2) = \sigma_c^2 + \sigma_u^2$$

$$\mathbb{E}(v_{i,t}v_{i,s}) = \sigma_c^2$$

$$\mathbf{\Omega} = \sigma_u^2 \mathbf{I}_T + \sigma_c^2 \mathbf{j}_T \mathbf{j}_T'$$

Following Hayashi [2000], we assume that  $\mathbf{\Omega}$  is invertible and positive definite such that there exists a matrix  $\mathbf{C}$  satisfying

$$\mathbf{\Omega}^{-1} = \mathbf{C}'\mathbf{C}$$

We transform the regression specification by left-multiplying  $\mathbf{C}$ :

$$\mathbf{C}\mathbf{y}_i = \mathbf{C}\mathbf{X}_i\boldsymbol{\beta} + \mathbf{C}\mathbf{v}_i$$

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{X}}_i\boldsymbol{\beta} + \tilde{\mathbf{v}}_i$$

The method of moments procedure yields the **random effects** estimator of  $\boldsymbol{\beta}$ , a variant of the **generalized least squares (GLS)** estimator:

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{RE} &= \left( \frac{1}{N} \sum_{i=1}^N \tilde{\mathbf{X}}_i' \tilde{\mathbf{X}}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \tilde{\mathbf{X}}_i' \tilde{\mathbf{y}}_i \right) \\ &= \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{X}_i \right)^{-1} \left( \frac{1}{N} \sum_{i=1}^N \mathbf{X}_i' \mathbf{\Omega}^{-1} \mathbf{y}_i \right) \end{aligned}$$

### 6.2.2 Fixed Effects

Consider a system of  $M$  linear equations [Hayashi, 2000]:

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\delta} + \boldsymbol{\varepsilon}_i$$

with orthogonality conditions

$$\begin{aligned} \mathbb{E}(\boldsymbol{\varepsilon}_i \otimes \mathbf{x}_i) &= \mathbf{0} \\ \mathbf{x}_i &= \bigcup_{j=1}^M \mathbf{z}_{ij} \end{aligned}$$

The **error components** model assumes the unobservable error term  $\varepsilon_{im}$  consists of two components:

$$\varepsilon_{im} = \alpha_i + \eta_{im}$$

The first unobservable component  $\alpha_i$  does not contain the  $m$  subscript and therefore is common to all equations. It is called the **fixed effect**. If we define

$$\boldsymbol{\eta}_i \equiv \begin{bmatrix} \eta_{i1} \\ \vdots \\ \eta_{iM} \end{bmatrix}$$

the matrix representation of the  $M$ -equation system is

$$\mathbf{y}_i = \mathbf{Z}_i \boldsymbol{\delta} + \mathbf{1}_M \cdot \alpha_i + \boldsymbol{\eta}_i \tag{6}$$

The orthogonality conditions are satisfied if the regressors of the system are orthogonal to

both error components:

$$\mathbb{E}(\mathbf{z}_{im} \cdot \alpha_i) = \mathbf{0} \quad (7)$$

$$\mathbb{E}(\mathbf{z}_{im} \cdot \eta_{ih}) = \mathbf{0} \quad (8)$$

and we can use either pooled OLS or a random effects approach to estimate  $\delta$ . But if we do not believe that (7) is reasonable, then we need an approach that is robust to its failure: fixed effects.

The **fixed-effects estimator** is consistent even when the regressors are not orthogonal to the fixed effect  $\alpha_i$ . The estimator is applied to an  $M$ -equation system transformed from the original system (6). The matrix used for the transformation is the annihilator associated with  $\mathbf{1}_M$ :

$$\begin{aligned} \mathbf{Q}_{M \times M} &\equiv \mathbf{I}_M - \mathbf{1}_M(\mathbf{1}'_M \mathbf{1}_M)^{-1} \mathbf{1}'_M \\ &= \mathbf{I}_M - \frac{1}{M} \mathbf{1}_M \mathbf{1}'_M \\ &= \mathbf{I}_M - \frac{1}{M} \mathbf{1}_{M \times M} \end{aligned}$$

The  $\mathbf{Q}$  matrix extracts *deviations from group means*. Multiplying the  $M$ -dimensional vector  $\mathbf{y}_i$  from the left by  $\mathbf{Q}$  results in

$$\tilde{\mathbf{y}}_i \equiv \mathbf{Q} \mathbf{y}_i = \begin{bmatrix} y_{i1} - \bar{y}_i \\ \vdots \\ y_{iM} - \bar{y}_i \end{bmatrix} = \mathbf{y}_i - \mathbf{1}_M \cdot \bar{y}_i$$

where

$$\bar{y}_i = \frac{1}{M} \mathbf{1}'_M \mathbf{y}_i = \frac{1}{M} \sum_{m=1}^M y_{im}$$

is the **group mean** for the dependent variable. This mean is over  $m$ , not  $i$ ; it is specific to

each group. Though fixed effects is robust to the possibility that  $\mathbb{E}(\mathbf{z}_{im} \cdot \alpha_i) \neq \mathbf{0}$ , the robustness comes at a price; transforming the original system with  $\mathbf{Q}$  implies that any variable which is time-invariant can no longer be identified.

The identification condition for fixed-effects estimation is that  $\mathbf{F}_i$  and  $\mathbf{b}_i$  are defined so that

$$\mathbb{E}(\mathbf{Q}\mathbf{F}_i \otimes \mathbf{x}_i) \tag{9}$$

is of full column rank and  $\mathbf{x}_i = \bigcup_{j=1}^M \mathbf{z}_{ij}$ .

The system of  $M$ -equations (6) can be rewritten as<sup>6</sup>

$$\mathbf{y}_i = \mathbf{F}_i\boldsymbol{\beta} + \mathbf{1}_M \cdot \mathbf{b}'_i\boldsymbol{\gamma} + \mathbf{1}_M \cdot \alpha_i + \boldsymbol{\eta}_i \tag{10}$$

To derive the fixed effects estimator, we multiply both sides of (10) from the left by  $\mathbf{Q}$  and exploit  $\mathbf{Q}\mathbf{1}_M = \mathbf{0}$  (annihilator) to obtain

$$\begin{aligned} \mathbf{Q}\mathbf{y}_i &= \mathbf{Q}\mathbf{F}_i\boldsymbol{\beta} + \mathbf{Q}\boldsymbol{\eta}_i \\ \tilde{\mathbf{y}}_i &= \tilde{\mathbf{F}}_i\boldsymbol{\beta} + \tilde{\boldsymbol{\eta}}_i \end{aligned}$$

Thus, the fixed effect  $\alpha_i$  and common regressors drop out in the transformed equations.

---

<sup>6</sup>The **error-components model** consists of the following assumptions:

1. Linearity (10).
2. Random samples:  $\{\mathbf{y}_i, \mathbf{Z}_i\}$  is i.i.d.
3. SUR assumption with two error components (7).
4. Identification:  $\mathbb{E}(\mathbf{Z}_i \otimes \mathbf{x}_i)$  is of full column rank.
5. Conditional homoskedasticity:  $\mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}'_i | \mathbf{x}_i) = \mathbb{E}(\boldsymbol{\varepsilon}_i\boldsymbol{\varepsilon}'_i) \equiv \boldsymbol{\Sigma}$ .
6. Nonsingular  $E(\mathbf{g}_i\mathbf{g}'_i)$ .
7. Fixed-effects identification (9).

Form the pooled sample of transformed  $\mathbf{y}_i$  and  $\mathbf{F}_i$  as

$$\tilde{\mathbf{y}} \equiv \begin{bmatrix} \tilde{\mathbf{y}}_1 \\ \vdots \\ \tilde{\mathbf{y}}_n \end{bmatrix}, \tilde{\mathbf{F}} \equiv \begin{bmatrix} \tilde{\mathbf{F}}_1 \\ \vdots \\ \tilde{\mathbf{F}}_n \end{bmatrix}$$

The **fixed-effects estimator** of  $\boldsymbol{\beta} \equiv \hat{\boldsymbol{\beta}}_{\text{FE}}$ , is the pooled OLS estimator, that is, the OLS estimator applied to the pooled sample  $(\tilde{\mathbf{y}}, \tilde{\mathbf{F}})$  of size  $Mn$ :

$$\begin{aligned} \hat{\boldsymbol{\beta}}_{\text{FE}} &\equiv (\tilde{\mathbf{F}}' \tilde{\mathbf{F}})^{-1} \tilde{\mathbf{F}}' \tilde{\mathbf{y}} \\ &= \left( \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{F}}_i' \tilde{\mathbf{F}}_i \right)^{-1} \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{F}}_i' \tilde{\mathbf{y}}_i \end{aligned}$$

From the sampling error

$$\hat{\boldsymbol{\beta}}_{\text{FE}} - \boldsymbol{\beta} = \left( \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{F}}_i' \tilde{\mathbf{F}}_i \right)^{-1} \frac{1}{n} \sum_{i=1}^n \tilde{\mathbf{F}}_i' \tilde{\boldsymbol{\eta}}_i$$

we derive the asymptotic distribution of the fixed effects estimator

$$\begin{aligned} \sqrt{n}(\hat{\boldsymbol{\beta}}_{\text{FE}} - \boldsymbol{\beta}) &\xrightarrow{D} \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{ff}^{-1} \mathbf{S} \boldsymbol{\Sigma}_{ff}^{-1}) \\ \mathbf{S} &= \mathbb{E}(\tilde{\mathbf{F}}_i' \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i' \tilde{\mathbf{F}}_i) \end{aligned}$$

A consistent estimator of the asymptotic variance is (cf. Wooldridge [2010, pg. 868])

$$\widehat{\text{Avar}}(\hat{\boldsymbol{\beta}}_{\text{FE}}) = \left( \sum_{i=1}^n \tilde{\mathbf{F}}_i' \tilde{\mathbf{F}}_i \right)^{-1} \sum_{i=1}^n \tilde{\mathbf{F}}_i' \tilde{\boldsymbol{\eta}}_i \tilde{\boldsymbol{\eta}}_i' \tilde{\mathbf{F}}_i \left( \sum_{i=1}^n \tilde{\mathbf{F}}_i' \tilde{\mathbf{F}}_i \right)^{-1}$$

**6.2.2.1 Interlude: Fixed Effects and Least-Squares Dummy Variables** The system of  $M$ -equations is

$$\mathbf{y}_i = \mathbf{F}_i \boldsymbol{\beta} + \mathbf{1}_M \cdot \mathbf{b}'_i \boldsymbol{\gamma} + \mathbf{1}_M \cdot \alpha_i + \boldsymbol{\eta}_i$$

$(M \times 1) \quad (M \times \#\beta)(\#\beta \times 1) \quad (M \times 1) \quad (1 \times \#\gamma)(\#\gamma \times 1) \quad (M \times 1) \quad (M \times 1)$

With an abuse of notation, redefine  $\alpha_i \equiv \underset{\text{"Old"}}{\alpha_i} + \mathbf{b}'_i \boldsymbol{\gamma}$ :

$$\mathbf{y}_i = \mathbf{F}_i \boldsymbol{\beta} + \mathbf{1}_M \cdot \alpha_i + \boldsymbol{\eta}_i$$

$(M \times 1) \quad (M \times \#\beta)(\#\beta \times 1) \quad (M \times 1) \quad (1 \times \#\gamma)(\#\gamma \times 1) + 1 \times 1 \quad (M \times 1)$

Define the corresponding stacked vectors and matrices:

$$\begin{aligned} \mathbf{y} &= \mathbf{F} \boldsymbol{\beta} + \mathbf{D} \boldsymbol{\alpha} + \boldsymbol{\eta} \\ (Mn \times 1) & \quad (Mn \times \#\beta) \quad (Mn \times n)(n \times 1) \quad (Mn \times 1) \\ &= \mathbf{W} \boldsymbol{\theta} + \boldsymbol{\eta} \\ & \quad (Mn \times (n + \#\beta))((n + \#\beta) \times 1) \end{aligned}$$

where

$$\mathbf{D} = \mathbf{I}_n \otimes \mathbf{1}_M \quad (\text{An identity matrix with } \mathbf{1}_M \text{ along the diagonal})$$

$$\mathbf{W} = (\mathbf{D} \ ; \ \mathbf{F})$$

$$\boldsymbol{\theta} = [\boldsymbol{\alpha}' \ \boldsymbol{\beta}']'$$

It follows that the OLS estimator of  $\boldsymbol{\theta}$  on a pooled sample of size  $Mn$ ,  $(\mathbf{y}, \mathbf{W})$  is

$$\underbrace{(\mathbf{W}'\mathbf{W})^{-1}}_{(n+\#\beta)(n+\#\beta)} \underbrace{\mathbf{W}' \mathbf{y}}_{(n+\#\beta) \times 1}$$

$\{(n+\#\beta) \times Mn\} (Mn \times 1)$

(a) Show that the FE estimator is the last  $\#\boldsymbol{\beta}$  elements of  $\widehat{\boldsymbol{\theta}}$ .

*Proof.* Following the second hint, if  $\mathbf{a}$  and  $\mathbf{b}$  are the OLS estimators of  $\boldsymbol{\alpha}$  and  $\boldsymbol{\beta}$ , then

$(\mathbf{a}, \mathbf{b})$  satisfies the system of  $n + \#\beta$  linear simultaneous equations:

$$\begin{bmatrix} \mathbf{D}'\mathbf{D} & \mathbf{D}'\mathbf{F} \\ \mathbf{F}'\mathbf{D} & \mathbf{F}'\mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{a} \\ \mathbf{b} \end{bmatrix} = \begin{bmatrix} \mathbf{D}'\mathbf{y} \\ \mathbf{F}'\mathbf{y} \end{bmatrix}$$

Starting with  $\mathbf{a} = (\mathbf{D}'\mathbf{D})^{-1}(\mathbf{D}'\mathbf{y} - \mathbf{D}'\mathbf{F}\mathbf{b})$ , we have

$$\mathbf{a} = (\mathbf{D}'\mathbf{D})^{-1}(\mathbf{D}'\mathbf{y} - \mathbf{D}'\mathbf{F}\mathbf{b})$$

Substituting  $\mathbf{a}$  into the  $\#\beta$  equations yields

$$\mathbf{F}'\mathbf{D} \left( (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{y} - (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{F}\mathbf{b} \right) + \mathbf{F}'\mathbf{F}\mathbf{b} = \mathbf{F}'\mathbf{y}$$

$$\left( \mathbf{F}'\mathbf{F} - \mathbf{F}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{F} \right) \mathbf{b} = \mathbf{F}'\mathbf{y} - \mathbf{F}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{y}$$

$$\left( \mathbf{F}' - \mathbf{F}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{F} \right) \mathbf{b} = \mathbf{F}'\mathbf{y} - \mathbf{F}'\mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}'\mathbf{y}$$

$$\mathbf{F}' \left( \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' \right) \mathbf{b} = \mathbf{F}' \left( \mathbf{I} - \mathbf{D}(\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}' \right) \mathbf{y}$$

$$\mathbf{F}' \underset{(Mn \times Mn)}{\mathbf{M}_D} \mathbf{b} = \mathbf{F}'\mathbf{M}_D\mathbf{y}$$

$$\begin{aligned} \mathbf{b} &= (\mathbf{F}'\mathbf{M}_D\mathbf{F})^{-1}\mathbf{F}'\mathbf{M}_D\mathbf{y} \\ &= (\mathbf{F}'\mathbf{M}_D\mathbf{M}_D\mathbf{F})^{-1}\mathbf{F}'\mathbf{M}_D\mathbf{M}_D\mathbf{y} \\ &= (\tilde{\mathbf{F}}'\tilde{\mathbf{F}})^{-1}\tilde{\mathbf{F}}'\tilde{\mathbf{y}} \end{aligned}$$

□

Therefore, the fixed effects estimator is equivalent to the LSDV (least-squares dummy variables) estimator since

$$\begin{aligned}
\underset{(Mn \times 1)}{\tilde{\mathbf{y}}} &= \begin{bmatrix} \tilde{y}_1 \\ \vdots \\ \tilde{y}_n \end{bmatrix} = \begin{bmatrix} \mathbf{Q}\mathbf{y}_1 \\ \vdots \\ \mathbf{Q}\mathbf{y}_n \end{bmatrix} = (\mathbf{I}_n \otimes \mathbf{Q})\mathbf{y} = \mathbf{M}_D\mathbf{y} \\
\underset{(Mn \times \#\beta)}{\tilde{\mathbf{F}}} &= \begin{bmatrix} \tilde{\mathbf{F}}_1 \\ \vdots \\ \tilde{\mathbf{F}}_n \end{bmatrix} = \begin{bmatrix} \mathbf{Q}\mathbf{F}_1 \\ \vdots \\ \mathbf{Q}\mathbf{F}_n \end{bmatrix} = (\mathbf{I}_n \otimes \mathbf{Q})\mathbf{F} = \mathbf{M}_D\mathbf{F}
\end{aligned}$$

Note that I used, but did not prove, that  $\mathbf{M}_D = \mathbf{I}_n \otimes \mathbf{Q}$ . Intuitively, the two are equivalent because the within-estimator removes all time invariant regressors, including the fixed effect, while LSDV attributes all “time invariant-ness” to the dummy variables.

### 6.3 Differences-in-Differences

Panel data requires repeated observations on the same units. But what if the regressor of interest varies only at a more aggregate level (e.g. a state)? Omitted variables bias would then come from unobserved variables at the aggregate and year level [Angrist and Pischke, 2008].

#### 6.3.1 Minimum wage on unemployment [Angrist and Pischke, 2008]

Differences-in-differences (DD) is a version of fixed effects estimation using aggregate data. Consider the effect of a minimum wage on unemployment. Let  $Y_{1ist}$  be fast food employment at restaurant  $i$  in state  $s$  at period  $t$  if there is a high (state) minimum wage and  $Y_{0ist}$  be fast food employment at restaurant  $i$  in state  $s$  at period  $t$  if there is a low (state) minimum wage. The DD setup relies on an additive structure for potential outcomes in the **no**-treatment state:

$$\mathbb{E}(Y_{0ist} \mid s, t) = \gamma_s + \lambda_t$$

That is, in the absence of a minimum wage change, employment is determined by the sum of a time-invariant state (fixed) effect and a year effect that is common across states (unit-invariant). Let  $D_{st}$  be a dummy for high-minimum-wage states, where states are indexed by  $s$  and observed in period  $t$ . Assuming that  $\mathbb{E}(Y_{1ist} - Y_{0ist} | s, t) = \beta$ , we have

$$Y_{ist} = \gamma_s + \lambda_t + \beta D_{st} + \varepsilon_{ist}$$

where  $\mathbb{E}(\varepsilon_{ist} | s, t) = 0$ . Consider two time periods, February and November, and two states, New Jersey and Pennsylvania, where New Jersey is the high minimum wage state in November.

$$\begin{aligned} \mathbb{E}[Y_{ist} | s = PA, t = 11] - \mathbb{E}[Y_{ist} | s = PA, t = 2] &= (\gamma_{PA} + \lambda_{11}) - (\gamma_{PA} + \lambda_2) \\ &= \lambda_{11} - \lambda_2 \\ \mathbb{E}[Y_{ist} | s = NJ, t = 11] - \mathbb{E}[Y_{ist} | s = NJ, t = 2] &= (\gamma_{NJ} + \lambda_{11} + \beta) - (\gamma_{NJ} + \lambda_2) \\ &= \lambda_{11} - \lambda_2 + \beta \end{aligned}$$

We have two differences. As the name suggest, we take the difference *of* these differences:

$$\begin{aligned} &(\mathbb{E}[Y_{ist} | s = NJ, t = 11] - \mathbb{E}[Y_{ist} | s = NJ, t = 2]) - (\mathbb{E}[Y_{ist} | s = PA, t = 11] - \mathbb{E}[Y_{ist} | s = PA, t = 2]) \\ &= (\lambda_{11} - \lambda_2 + \beta) - (\lambda_{11} - \lambda_2) \\ &= \beta \end{aligned}$$

which gives us the causal effect of interest.

### 6.3.2 Empirical Corporate Finance: Collateral on Lending [Assunção et al., 2011]

Consider an example from Assunção et al. [2011], who attempt to demonstrate the impact of a law enhancing the repossession of collateral on lending. They initially estimate

$$y_{i,t} = \alpha + \beta_1 \times \text{Law}_{i,t} + \mathbf{T}_{i,t}\boldsymbol{\lambda} + \mathbf{b}_i\boldsymbol{\psi} + \mathbf{c}_{i,t}\boldsymbol{\theta} + \mathbf{e}_{i,t}\boldsymbol{\gamma} + \varepsilon_{i,t}$$

$\beta_1$  is the coefficient of interest and purports to measure the effect of the law on financial contracting, as  $\text{Law}_{i,t}$  is a dummy indicating that the loan was initiated after the implementation of the law. But the specification assumes that after controlling for various characteristics, only changes in the law over time affect the outcomes of interest; that is, only the variable ‘law’ captures the effects of the credit reform. But isn’t it possible that something else coincides with the time-series measure of the legal reform (i.e. an unobserved contemporaneous shock affects lending through channels other than  $\text{Law}_{i,t}$ )?

That is why we use differences-in-differences.

### 6.3.3 Empirical Corporate Finance: Antitakeover legislation on managers’ preferences [Bertrand and Mullainathan, 2003]

Bertrand and Mullainathan study managers’ preferences by asking what goals managers would pursue if they were not closely monitored. States’ passing of takeover legislation constitutes treatment, and they estimate

$$y_{jkl} = \alpha_t + \alpha_j + \gamma X_{jkl} + \delta BC_{kt} + \varepsilon_{jkl}$$

where  $j$  indexes firms,  $k$  indexes state of incorporation,  $l$  indexes state of location,  $t$  indexes time,  $y_{jkl}$  is the dependent variable of interest,  $\alpha_t$  and  $\alpha_j$  are year and firm fixed effects,

$X_{jkl t}$  are control variables,  $BC_{kt}$  is a dummy variable that equals one if an antitakeover law has been passed by time  $t$  in state  $k$ , and  $\varepsilon_{jkl t}$  is an error term.  $\delta$  is the estimate of the law's effect and is the primary coefficient of interest.

The specification implicitly takes as the control group all firms incorporated in states not passing a law at time  $t$ , even if they have already passed a law or will pass one later on.

## 7 The Fama-MacBeth Approach [Campbell et al., 1997]

Recall that *the* CAPM is

$$\mathbb{E}[R_i] = R_f + \beta_{im} (\mathbb{E}[R_m] - R_f)$$

Letting  $Z_i \equiv R_i - R_f$ , we have

$$\mathbb{E}[Z_i] = \beta_{im} \mathbb{E}[Z_m]$$

An interpretation of the CAPM is that expected returns and market betas are linearly related, and the relationship completely explains the cross section of expected returns. To test this using a cross-sectional regression methodology, for each cross section,

1. project the returns on the betas, and
2. aggregate the estimates in the time dimension.

Assuming that the betas are known, the regression model for the  $t$ th cross section of  $N$  assets is

$$Z_t = \gamma_{0t} \mathbf{1}_N + \gamma_{1t} \boldsymbol{\beta}_m + \boldsymbol{\eta}_t \tag{11}$$

Implementation of the Fama-MacBeth approach involves two steps:

1. Given  $T$  periods of data, (11) is estimated using OLS for each  $t \in \mathcal{T}$ , giving the  $T$  estimates of  $\gamma_{0t}$  and  $\gamma_{1t}$ .
2. The time series of  $\hat{\gamma}_{0t}$ 's and  $\hat{\gamma}_{1t}$ 's are analyzed.

Defining  $w(\hat{\gamma}_j)$  as the  $t$ -statistic, we have

$$w(\hat{\gamma}_j) = \frac{\hat{\gamma}_j}{\hat{\sigma}_{\gamma_j}}$$

where

$$\hat{\gamma}_j = \frac{1}{T} \sum_{t=1}^T \hat{\gamma}_{jt}$$

and

$$\hat{\sigma}_{\gamma_j}^2 = \frac{1}{T(T-1)} \sum_{t=1}^T (\hat{\gamma}_{jt} - \hat{\gamma}_j)^2$$

The distribution of  $w(\hat{\gamma}_j)$  is Student  $t$  with  $(T-1)$  degrees of freedom and asymptotically is standard normal.

## References

- R. Adams, H. Almeida, and D. Ferreira. Understanding the relationship between founder-ceos and firm performance. *Journal of Empirical Finance*, 16(1):136–150, 2009.
- J. Angrist and J.-S. Pischke. *Mostly Harmless Econometrics*. Princeton University Press, Princeton, New Jersey, 2008.
- J. J. Assunção, E. Benmelech, and F. S. S. Silva. Judicial efficiency and financial contracts: Evidence from a natural experiment. *Working paper*, 2011.
- M. Bertrand and S. Mullainathan. Enjoying the quiet life? corporate governance and managerial preferences. *Journal of Political Economy*, 111(5):21–39, 2003.
- A. C. Cameron and P. K. Trivedi. *Microeconometrics: Methods and Applications*. Cambridge University Press, New York, NY, 2005.
- J. Y. Campbell, A. W. Lo, and A. C. MacKinlay. *The Econometrics of Financial Markets*. Princeton University Press, Princeton, New Jersey, 1997.

F. Hayashi. *Econometrics*. Princeton University Press, 2000.

J. H. Stock and M. W. Watson. *Introduction to Econometrics*. Pearson Education, Boston, Massachusetts, 2007.

J. M. Wooldridge. *Econometric Analysis of Cross Section and Panel Data*. The MIT Press, Cambridge, MA, 2010.