Notes Tutorial 1 - Advanced Econometrics II - Tinbergen Institute

Gabriela M. Miyazato Szini (g.m.m.szini@uva.nl)

January 13, 2022

1 Introduction to Treatment Effects

Aim: evaluate effect of exposure of set of units to a program or treatment on some outcome.

Ideally: comparison of two potential outcomes for the same unit → however, we only observe one of them.

To estimate: impose assumptions / estimate counterfactuals.

2 Notation of the Rubin Causal Model

- $y_i(1)$ or $y_{i,1}$ if treated
- $y_i(0)$ or $y_{i,0}$ if not treated
- Causal effect: $\Delta_i = y_i(1) - y_i(0)$
- Observed data: $y_i = D_i y_i(1) + (1 - D_i) y_i(0)$, where D_i is a binary variable related to treatment assignment.

Be aware: $\mathbb{E}\left(y_i(1)\right)$ is not necessarily equal to $\mathbb{E}\left(y_i \mid D_i = 1\right)$, since the later considers potential outcomes only for treated individuals ⇒ we can have self selection into treatment.

→ in an RCT treatment is allocated randomly: D_i is independent of potential outcomes and x_i, therefore the equality between the two terms above holds.

→ however, in most cases assignment is randomized only conditional on some confounder $x_i : y_i(1), y_i(0) \perp D_i \mid x_i$. Meaning that only after holding some characteristics constant, assignment to treatment is random.

3 Definition of Average Treatment Effect (ATE) and Assumptions

$$\text{ATE} = \Delta = \mathbb{E}(yi(1) - yi(0))$$

where, its natural estimate would be given by:

$$\hat{\Delta} = \frac{1}{N} \sum_{i=1}^{N} [y_i(1) - y_i(0)]$$

However, we cannot compute this value, since we do not observe both potential outcomes - we only observe $y_i(1)$ for treated individuals and $y_i(0)$ for untreated - we impose some assumptions such that we can estimate the ATE.

ASSUMPTION 1: SUTVA
The treatment of one individual does not affect any other individuals (no general equilibrium effect or spillovers).

ASSUMPTION 2: Unconfounded assignment
Also known as: exogeneity (conditional on x_i, treatment is exogenous), conditional independence, or selection on observables.

Formally: $y_i(1), y_i(0) \perp D_i \mid x_i$

The intuition: x_i might affect both potential outcomes and treatment at the same time, therefore both are orthogonal only after conditioning on x_i.

By adjusting treatment and control groups for differences in observed covariates, we remove biases between treatment and control units. Then, the unconfoundedness assumption is that there is a rich set of predictors for the treatment indicator, such that adjusting for them leads to valid estimates of the causal effect (the basis for the idea of regression analysis or propensity score).

Note: $y_i(1), y_i(0) \perp D_i$ is stronger than the previous expression. If instead this assumption is valid can calculate the ATE based only on the average of observed outcomes of treated individuals minus the average for untreated.

ASSUMPTION 3: Conditional mean assumption (special case of Assumption 2)
The potential outcome of receiving treatment does not change whether received treatment or not, conditioning on covariates x_i.

$$\mathbb{E}(y_i(1) \mid D_i = 1, x_i) = \mathbb{E}(y_i(1) \mid D_i = 0, x_i) = \mathbb{E}(y_i(1) \mid x_i)$$

$$\mathbb{E}(y_i(0) \mid D_i = 1, x_i) = \mathbb{E}(y_i(0) \mid D_i = 0, x_i) = \mathbb{E}(y_i(0) \mid x_i)$$

ASSUMPTION 4: Overlap/ matching assumption
Probability of receiving treatment anywhere along the spectrum of x is strictly positive and < 1.

$$0 < \Pr(Di \mid xi) < 1$$

4 Estimation of ATE with regressions

Defining the conditional averages:

$$\mu_n\left(x_i\right)=\mathbb{E}\left(y_{1,i} \mid x_i\right)$$
$$\mu_0\left(x_i\right)=\mathbb{E}\left(y_{0,i} \mid x_i\right)$$

From the lecture, given the unconfoundedness assumption, the ATE can be defined and estimated as:

$$\text{ATE } \Rightarrow \Delta\left(x_i\right)=\mu_1\left(x_i\right)-\mu_0\left(x_i\right)$$
$$\hat{\Delta}=\frac{1}{N} \sum_{i=1}^N\left[\hat{\mu}_1\left(x_i\right)-\hat{\mu}_0\left(x_i\right)\right]$$

Moreover, in the case of a simple linear regression, we have that:

$$\mu_1\left(x_i\right)=\alpha_1+\beta_1^{\prime}\left(x_i-\mu_x\right)$$
$$\mu_0\left(x_i\right)=\alpha_0+\beta_0^{\prime}\left(x_i-\mu_x\right)$$

Therefore, the ATE is:

$$\begin{aligned}
\Delta(x_i) &= \mathbb{E}\left(\mu_1\left(x_i\right)-\mu_0\left(x_i\right)\right) \\
&= \mathbb{E}\left(\alpha_1+\beta_1^{\prime} \quad \underbrace{(xi-\mu_x)}_{\mathbb{E}(x_i)=\mu_x \rightarrow \text{ cancels!}} \quad -\alpha_0-\beta_0^{\prime}\left(x_i-\bar{x}\right)\right) \\
&= \alpha_1-\alpha_0
\end{aligned}$$

And, its estimated version is:

$$\hat{\Delta}=\hat{\alpha}_1-\hat{\alpha}_0$$

Also, its asymptotic distribution is given by:

$$\sqrt{N}(\hat{\Delta}-\Delta) \xrightarrow{d} N\left(0, \sigma_{\alpha, 1}^2+\sigma_{\alpha, 0}^2\right)$$

where the asymptotic variance is obtained through the formula below, and given the fact that the covariance between $\hat{\alpha}_1$ and $\hat{\alpha}_0$ is zero since those are the estimated intercepts of two separate regressions - one for the treated individuals and one for the untreated.

$$\operatorname{Var}(\hat{\Delta})=\operatorname{Var}\left(\hat{\alpha}_1\right)+\operatorname{Var}\left(\hat{\alpha}_0\right)-2 \cdot \operatorname{Cov}\left(\hat{\alpha}_1, \hat{\alpha}_0\right)$$

Given this result for the estimate of the ATE our **aim now is to show why this is different from the estimates of an RCT** - where only the differences of the averages of the treated and untreated outcomes are taken into account - and how it takes into account a correction for the differences in inputs as mentioned in slides.

From the regression results we would have that:

$$\hat{\alpha}_1=\bar{y}_1-\hat{\beta}_1^{\prime}\left(\bar{x}_1-\bar{x}\right)$$
$$\hat{\alpha}_0=\bar{y}_0-\hat{\beta}_0^{\prime}\left(\bar{x}_0-\bar{x}\right)$$

Moreover, by rewriting $\bar{x}=\frac{N_T \bar{x}_1+N_C \bar{x}_0}{N_T+N_C}$, where N_T is the number of treated individuals and N_C is the number of untreated, and substituting in the above expressions, and further substituting the above expressions in the estimated ATE, we have:

$$\begin{aligned}
\hat{\Delta} &= \bar{y}_1 - \bar{y}_0 - \hat{\beta}_1' \left(\bar{x}_1 - \frac{N_T \overline{x}_1 + N_C \overline{x_0}}{N_T + N_C} \right) + \hat{\beta}_0{}' \left(\overline{x}_0 - \frac{N_T \overline{x}_1 + N_C \overline{x}_0}{N_T + N_C} \right) \\
&= \overline{y}_1 - \overline{y}_0 - \hat{\beta}_1{}' \left(\frac{N_C \overline{x}_1 - N_C \overline{x}_0}{N_T + N_C} \right) + \hat{\beta}_0{}' \left(\frac{N_T \overline{x}_0 - N_T \overline{x}_1}{N_T + N_C} \right) \\
&= \overline{y}_1 - \overline{y}_0 - \frac{N_C}{N_T + N_C} \hat{\beta}_1' \left(\bar{x}_1 - \bar{x}_0 \right) - \frac{N_T}{N_T + N_C} \hat{\beta}_0{}' \left(\overline{x}_1 - \bar{x}_0 \right)
\end{aligned}$$

where $\overline{y}_1$ and $\overline{x}_1$ are averages over the treated individuals and $\overline{y}_0$ and $\overline{x}_0$ are averages over the untreated individuals. The first difference in the last line of the previous expression, $\overline{y}_1 - \overline{y}_0$ would be the estimated ATE if we had an RCT and complete randomization of treatment, therefore, the remaining terms in this line are the correction for differences in inputs between treated and untreated groups.

However, as mentioned in the lecture slides, the study conducted by Lalonde showed that those estimates are very sensitive to the specification of the regression model, and even when considering several different specifications, the results can still be biased. An identified problem that remains in this approach is that depending on the values of the covariates the probability of receiving treatment could vary a lot, therefore there is too much imbalance in the values of the covariates between the treated and untreated groups.

5 Derivations of doubly-robust estimator

Given the problem of imbalance mentioned in the last paragraph, one idea is to take into account how similar individuals are through the propensity score defined by $p(x_i) = \Pr(D_i = 1 \mid x_i)$. For instance, one way to take into account for the propensity score is by using the inverse propensity weighting estimator:

$$\Delta_{ipw} = \mathbb{E}\left[\frac{D_i y_i}{p(x_i)} - \frac{(1 - p_i)\, y_i}{(1 - p(x_i))} \right]$$

that could be estimated as:

$$\hat{\Delta}_{ipw} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{D_i y_i}{\hat{p}(x_i)} - \frac{(1 - D_i) y_i}{1 - \hat{p}(x_i)} \right]$$

where $\hat{p}(x_i)$ is the estimated propensity score, that can be estimated, for instance, with a non-parametric model.

⇒ the IPW is great when the number of observations is large, but in small samples, we may need a parametric model that may be misspecified.
⇒ the idea of the doubly-robust estimator is to combine IPW and the regression framework from before, such that the estimate is unbiased even when only the propensity score, or only the regression model is correctly specified (we do not need both to be correctly specified).

Our aim in this section is to show the definition of the doubly-robust estimator and to derive the later property that we described above.

To obtain the estimator, one should run the following steps:

1) Estimate the parametric model for the propensity score → obtaining $\hat{p}(x_i)$
2) Estimate the regression coefficients weighting the observations by the propensity score (estimated):

$$\min_{\alpha_k,\beta_k} \sum_{i:Di=k} \frac{y_i - \alpha_k - \beta_k' \left(x_i - \bar{x}\right)}{\hat{p}\left(x_i\right)}$$

Then, the estimated ATE is given by $\hat{\Delta}_{dobrob} = \hat{\alpha}_1 - \hat{\alpha}_0$.

To derive the property that only the propensity score or the regression model need to be correctly specified, we first rewrite the expression for the ATE using the doubly-robust specification as:

$$\Delta_{dobrob} = \mathbb{E}\left[\frac{D_i y_i}{p\left(x_i\right)} + \left(1 - \frac{D_i}{p\left(x_i\right)}\right)\mu_1(x) - \left(\frac{y_i\left(1-D_i\right)}{1-p\left(x_i\right)} + \left(1 - \frac{\left(1-D_i\right)}{1-p\left(x_i\right)}\right)\mu_0(x)\right)\right]$$

This expression is equivalent to what we would obtain considering the two steps mentioned above. We can further rewrite:

$$\frac{D_i y_i}{p\left(x_i\right)} = y_i + \left(\frac{D_i - p\left(x_i\right)}{p\left(x_i\right)}\right) y_i$$

And substituting it in the expression for Δ_{dobrob}:

$$\begin{aligned}
\Delta_{dobrob} =& \mathbb{E}\left[y_i(1) + \frac{D_i - p\left(x_i\right)}{p\left(x_i\right)}\left(y_i(1) - \mu_1(x)\right) - \left(y_i(0) - \frac{D_i - p\left(x_i\right)}{1-p\left(x_i\right)}\left(y_i(0) - \mu_0(x)\right)\right)\right] \\
=& \mathbb{E}\left(y_i(1) - y_i(0)\right) + \mathbb{E}\left[\frac{D_i - p\left(x_i\right)}{p\left(x_i\right)}\left[y_i(1) - \mu_1\left(x_i\right)\right] + \frac{D_i - p\left(x_i\right)}{1-p\left(x_i\right)}\left[y_i(0) - \mu_0\left(x_i\right)\right]\right] \\
=& \text{ATE} + \mathbb{E}\left[\frac{D_i - p\left(x_i\right)}{p\left(x_i\right)}\left[y_i(1) - \mu_1\left(x_i\right)\right] + \frac{D_i - p\left(x_i\right)}{1-p\left(x_i\right)}\left[y_i(0) - \mu_0\left(x_i\right)\right]\right]
\end{aligned}$$

Therefore, to show that the expression for Δ_{dobrob} is unbiased for the ATE, we need to show that the second term above goes to zero. We consider then 2 scenarios, where in each either the propensity score or the regression model is correctly specified.

Scenario 1: propensity score is correctly specified but regression model is not
Meaning $p(x_i) = \mathbb{E}(D_i \mid x_i)$ but $\mu_j(x_i) \neq \mathbb{E}(y_i \mid D_i = j, x_i)$. Then, we can rewrite:

$$\begin{aligned}
\Delta_{dobrob} =& \text{ATE} + \mathbb{E}\left[\mathbb{E}\left[\frac{D_i - p(x_i)}{p(x_i)}\left[y_i(1) - \mu_1\left(x_i\right)\right] \mid y_i(1), x_i\right] + \mathbb{E}\left[\frac{D_i - p\left(x_i\right)}{1-p\left(x_i\right)}\left[y_i(0) - \mu_0\left(x_i\right)\right] \mid y_i(0), x_i\right]\right] \\
=& \text{ATE} + \mathbb{E}\left[\frac{\mathbb{E}\left[D_i \mid y_i(1), x_i\right] - p\left(x_i\right)}{p\left(x_i\right)}\left[y_i(1) - \mu_1\left(x_i\right)\right] + \frac{\mathbb{E}\left[D_i \mid y_i(0), x_i\right] - p\left(x_i\right)}{1-p\left(x_i\right)}\left[y_i(0) - \mu_0\left(x_i\right)\right]\right]
\end{aligned}$$

therefore, if the propensity score is estimated correctly, and $\mathbb{E}\left[D_i \mid y_i(0), x_i\right] = p(x_i)$ and $\mathbb{E}\left[D_i \mid y_i(1), x_i\right] = p(x_i)$, then the last term in the last line is equal to zero and Δ_{dobrob} is unbiased.

Scenario 2: regression model is correctly specified but propensity score is not
Meaning $\mu_j(x_i) = \mathbb{E}(y_i \mid D_i = j, x_i)$ but $p(x_i) \neq \mathbb{E}(D_i \mid x_i)$. Then, we can rewrite:

$$\begin{aligned}
\Delta_{dobrob} =& \text{ ATE} + \mathbb{E}\left[\mathbb{E}\left[\frac{D_i - p\left(x_i\right)}{p\left(x_i\right)}\left(y_i(1) - \mu_1\left(x_i\right)\right) \mid D_i = 1, x_i\right] + \mathbb{E}\left[\frac{D_i - p(x)}{1-p\left(x_i\right)}\left(y_i(0) - \mu_0\left(x_i\right)\right) \mid D_i = 0, x_i\right]\right] \\
=& \text{ ATE} + \mathbb{E}\left[\frac{1-p\left(x_i\right)}{p\left(x_i\right)}\left[\mathbb{E}\left(y_i(1) \mid D_i = 1, x_i\right) - \mu_1(x)\right] + \left(\frac{-p\left(x_i\right)}{1-p\left(x_i\right)}\right)\left[\left(\mathbb{E}\left(y_i(0) \mid D_i = 0, x_i\right) - \mu_0(x)\right]\right]\right.
\end{aligned}$$

therefore, if the regression model is specified correctly, and $\mathbb{E}\left(y_i(1) \mid D_i = 1, x_i\right) = \mu_1(x)$ and $\mathbb{E}\left(y_i(0) \mid D_i = 0, x_i\right) = \mu_0(x)$, then the last term in the last line is equal to zero and Δ_{dobrob} is unbiased.

Notes Tutorial 2 - Advanced Econometrics II - Tinbergen Institute

Gabriela M. Miyazato Szini (g.m.m.szini@uva.nl)

January 27, 2022

1 Treatment Effects: When the unconfoundedness assumption does not hold

In the previous TA, when we talked about propensity scores, we assumed that treatment participation was not randomly assigned, But depended on a vector of observable variables x . For example, treatment being targeted to some subpopulation defined by some observable characteristics , such as age, etc.

In this scenario, the assumption of conditional independence, or unconfoundedness was necessary, such that:

$$y_i(0), y_i(1) \perp D_i \mid x_i$$

However, there are cases where the treatment assignment D_i is dependent on unobserved characteristics that are also affecting the potential outcomes such that the assumption above does not hold.

Without the unconfoundedness assumption, there is no general approach to estimate treatment effects. When this self-selection problem arises, one can then estimate treatment effects using for instance, the LATE (IV) estimator, or the difference-in-differences estimator, among others. More specifically:

- IV (LATE): relies on the presence of instruments.
- Difference-in-differences: relies on the presence of panel data and the common trends assumption.

1.1 The difference-in-differences estimator

When one has data on treated and control groups both before and after the treatment period, we can use the difference-in-differences estimator, subject to certain assumptions.

Simplest setting: two groups (treated and non-treated/control) and two time periods, before and after treatment. Outcomes are observed for units that are in one of two groups in one of two time periods. Only units in one of the two groups in the second period are exposed to treatment.

We could think of comparing the average of outcomes only for this group in the two different time periods to obtain the estimated ATE. However, a part of the difference can be due to trends in the (potential) outcomes, and not the treatment itself, therefore, we need to correct for that.

Model: individual i, $i = 1...N$, belongs to a group $G_i \in \{0, 1\}$, where group 1 is the treatment group and outcomes are observed in periods $T_i \in \{0, 1\}$.
In the standard DiD model we can write the outcome for individual i in the absence of reatmint, $Y(0)$:

$$y_i(0) = \alpha + \beta T_i + \gamma G_i + \varepsilon_i$$

Where: β is time component common to both groups, γ is a group-specific and time-invariant component, ε_i are unobserved characteristics. It is assumed that $\varepsilon_i \perp (G_i, T_i)$.

We can also write the model in terms of a time-invariant individual specific fixed effect, potentially correlated with G_i:

$$y_i(0) = \alpha + \beta T_i + \gamma_i + \varepsilon_i$$

The equation for the outcome without treatment is then combined with an equation in the outcome given the treatment: $y_i(1) = y_i(0) + \tau_{did}$.

The standard DiD estimand is then:

$$\begin{aligned}\tau_{did} &= \mathbb{E}\left[y_i(1)\right] - \mathbb{E}\left[y_i(0)\right] \\ &= \left(\mathbb{E}\left[y_i|G_i = 1, T_i = 1\right] - \mathbb{E}\left[y_i|G_i = 1, T_i = 0\right]\right) - \left(\mathbb{E}\left[y_i|G_i = 0, T_i = 1\right] - \mathbb{E}\left[y_i|G_i = 0, T_i = 0\right]\right)\end{aligned}$$

The first term of the expression above, $(\mathbb{E}\left[y_i|G_i = 1, T_i = 1\right] - \mathbb{E}\left[y_i|G_i = 1, T_i = 0\right])$, measures the difference between outcomes in the treated group before and after the treatment, therefore, this measure removes the fixed-effects given by the group $G_i = 1$, however, it is still biased due to the possible effect of time trends in outcomes. On the other hand, taking into account simply the quantity $\mathbb{E}\left[y_i|G_i = 1, T_i = 1\right] - \mathbb{E}\left[y_i|G_i = 0, T_i = 1\right]$, which is the difference between the outcomes of treated and control groups after treatment is also biased, even if not suffering from the possible effect of time trends, due to the possible intrinsic difference in potential outcomes between the two groups both in the presence or in the absence of treatment.

The differences of differences then subtracts the average gain over time in the outcome of the control group (which, if the common trends assumption is satisfied measures the time trends for both groups) from the average gain over time in the outcome of the treated group (which captures both the treatment effect and the time trend), removing both biases mentioned before.

We can estimate the parameter τ_{did} using least squares on the observed outcome:

$$Y_i = \alpha + \beta_1 \cdot T_i + \gamma_i + \tau_{did} \cdot D_i + \varepsilon_i$$

where the treatment indicator D_i is equal to the interaction of group and time indicators : $D_i = T_i \cdot G_i$

2 When should you adjust standard errors for clustering?

Based on the paper Abadie et al. (2017).

2.1 Some fallacies on the topic

- ”The clustering problem is caused by the presence of a common unobserved random shock at the group level that will lead to correlation between all observations within each group”.
  → this motivation makes it difficult to justify clustering by some partitions of the population but not others (age, state). Also why would cluster when including FE?
- ”The consensus is to be conservative and avoid bias and use bigger and more aggregate clusters”.
  → there is actually harm in clustering at too aggregate level.
- ”If there is difference between clustered standard errors and others, then use clustered”.

2.2 Clustering as a design problem

Sampling design: sampling follows a 2-stage process, where in the first stage a subset of clusters is sampled and in the second stage units are sampled randomly from sampled clusters.

Experimental design: clusters of units rather than units are assigned to treatment.

In previous studies, clustering was based on experimental design, now this paper focuses on the sampling design. Some conclusions in this change of perspective are then reached:

- Correlations between residuals within clusters are not necessary or sufficient for cluster adjustment to matter. Clustering can matter also even when both residuals and regressors are uncorrelated within clusters.
- Data is only partially informative about whether should cluster. It matters:
  - How units in the sample were selected and if all clusters were sampled.
  - If units were assigned to treatment clustered.
- If sampling process and assignment mechanism are both not clustered, then one should not adjust standard errors, even when adjustment changes standard errors.

2.3 Example and misconceptions

As mentioned earlier, before we had a model-based approach and not a design-based approach. The model considered before takes into account outcomes Y_i, with covariates $W_i \in \{0,1\}$, and each unit i belongs to a cluster $C_i \in \{1,...,C\}$. It is based on a linear model:

$$Y_i = \alpha + \tau W_i + \varepsilon_i = \beta' X_i + \varepsilon_i$$

where ε_i is stochastic, and X_i, C_i are non-stochastic. Moreover, repeated sampling relies on only redrawing ε_i.

It is imposed that: $\mathbb{E}[\varepsilon \mid X, C] = 0$ and $\mathbb{E}\left[\varepsilon\varepsilon^\top \mid X, C\right] = \Omega$. Which leads to the variance:

$$\mathbb{V}(\hat{\beta}) = \left(X^\top X\right)^{-1}\left(X^\top \Omega X\right)\left(X^\top X\right)^{-1}$$

without clustering Ω is diagonal. We can also assume homoskedasticity, such that $\sigma^2 = \Omega_{ii} = \mathbb{V}(\varepsilon_i)$, leading to:

$$\mathbb{V}_{\text{OLS}} = \sigma^2 \left(X^\top X\right)^{-1}$$

Or, we can often allow for general heteroskedasticity:

$$\mathbb{V}_{\text{EHW}}(\hat{\beta}) = (X^\top X)^{-1} \left(\sum_{i=1}^{N} \Omega_{ii} X_i X_i^\top\right) \left(X^\top X\right)^{-1}$$

To incorporate clusters, Kloek (1981) use the homoskedastic structure:

$$\Omega_{ij} = \begin{cases} 0 & \text{if } C_i \neq C_j \\ \rho\sigma^2 & \text{if } C_i = C_j, i \neq j \\ \sigma_2 & \text{if } i = j \end{cases}$$

Then, we obtain the variance:

$$\mathbb{V}_{\text{KLOEK}}(\hat{\tau}) = \mathbb{V}_{\text{OLS}} \times \left(1 + \rho_\varepsilon \rho_W \frac{N}{C}\right)$$

where ρ_ε is the within cluster correlation of errors, and ρ_W is the within cluster correlation of regressors.

Liang and Zeger: further relax this expression such that Ω_{ij} is unrestricted for pairs (i, j) with $C_i = C_j$:

$$\mathbb{V}_{\text{LZ}}(\hat{\beta}) = \left(X^\top X\right)^{-1} \left(\sum_{c=1}^{C} X_c^\top \Omega_c X_c\right) \left(X^\top X\right)^{-1}$$

2.4 Intuition of main results of the paper

The expressions above for the variance estimators have been widely used, and from the equation for $\mathbb{V}_{\text{KLOEK}}(\hat{\tau})$ it is clear that clustering was considered for cases where there were correlations within clusters for errors and regressors. Therefore, according to this, clustered standard errors should not differ when we have a randomized experiment with random treatment assignment.

The main idea for this section in the paper is to disprove this statement with some simulation results. It was considered $N \approx 100000$, and $C = 100$ clusters with approximately 1000 units each, having therefore only modest variation in cluster sizes.

The model to be considered was:

$$Y_i = \alpha + \tau W_i + \varepsilon_i$$

where the within cluster correlations are approximately zero. Therefore, the adjustment in the expression for $\mathbb{V}_{\text{KLOEK}}(\hat{\tau})$, given by the term $\rho_\varepsilon \rho_W \frac{N}{C}$ would be close to zero, indicating that there should be no need for clustering.

However, in the simulations, it was found that $\hat{\mathbb{V}}_{\text{EHW}} \neq \hat{\mathbb{V}}_{\text{LZ}}$. Then, the authors claimed that instead, what should matter for the clustered standard errors to differs is whether, instead, the correlation between residuals and regressors $\rho_{\hat{\varepsilon}W}$ to be different than zero. Moreover, they reinforced the idea that even if the clustered standard errors differ, it does not necessarily mean that one should cluster standard errors.

Then, the authors introduce the sampling design, taking into account in the model how this sample was obtained. They consider a population of 10000000 units with $C = 100$ clusters, from which the $N \approx 100000$ sampled units were sampled randomly across all clusters. Moreover, the treatment assignment variable, $W_i \in \{0, 1\}$ assumes each value with independent probability $p = \frac{1}{2}$. Based on the obtained samples, they estimate the variances $\hat{\mathbb{V}}_{\text{EHW}}$ and $\hat{\mathbb{V}}_{\text{LZ}}$. At the end, they show that with this sampling design, the estimated variance $\hat{\mathbb{V}}_{\text{EHW}}$ leads to an appropriate coverage rate, while the estimated variance $\hat{\mathbb{V}}_{\text{LZ}}$ does not, once it is a higher estimate, which indicates that even if the clustered standard errors differ, it does not necessarily mean that one should cluster the standard errors. This conclusion also follows from the fact that the $\mathbb{V}_{\text{LZ}}$ is based on the assumption that there are clusters in the population beyond the 100 clusters which units were drawn from, which is not the case in this design.

2.5 Formal results

The aim of this section is to derive the exact variance to an approximation of the least squares estimators, that takes into account the sampling variation and the variation from the assignment mechanism. Once this expression is obtained, it is then compared to the previously proposed variance estimators.

Definitions

It is considered a sequence of populations n, which for each population there is a number of units, for instance, for the nth population we have M_n units $i = 1, .., M_n$, which is strictly increasing in n.

Each population is also partitioned into C_n clusters, which is weakly increasing in n. It is also defined:

- $C_{in} \in \{1, ..., C_n\}$, which is the stratum to which unit i belongs.
- $C_{inc} = \nVdash_{C_{in}=c}$, which is an indicator function for whether unit i belongs to cluster $C = c$.
- $M_{cn} = \sum_{i=1}^{n} C_{inc}$, which is the number of units in each cluster.

We are interested in the population average of treatment effect:

$$\tau_n = \frac{1}{M_n} \sum_{i=1}^{M_n} (Y_{in}(1) - Y_{in}(0)) = \bar{Y}_n(1) - \bar{Y}_n(0)$$

where each average in the last term is given by $\bar{Y}_n(W) = \frac{1}{M_n} \sum_{in}^{M_n} Y_{in}(W)$. We can also define treatment-specific residuals:

$$\varepsilon_{in}(W) = Y_{in}(W) - \frac{1}{M_n} \sum_{j=1}^{M_n} Y_{j_n}(W)$$

Then $Y_{in}(W)$ and $\varepsilon_{in}(W)$ are not stochastic.

Sampling and Assignment Mechanisms

We do not observe both Y_{in} (0) and Y_{in} (1), which one is observed is then depending on the value of the stochastic treatment $W_{in} \in \{0,1\}$. The realized outcomes and residuals are given by:

$$Y_{in} = Y_{in}\left(W_{in}\right)$$
$$\varepsilon_{in} = \varepsilon_{in}\left(W_{in}\right)$$

Moreover, we observe a subset of the M_n units in the population, that is indicated by the stochastic variable R_{in}, which is equal to 1 if observed and 0 otherwise. Therefore, if $R_{in} = 1$, we observe the triple (Y_{in}, W_{in}, C_{in}).

In a nutshell, what is observed depends on the stochastic variables W_{in} and R_{in}, and therefore, given an estimand of τ_n, the standard errors need to capture both variations.

Sampling process determines R_{in} and is independent of potential outcomes and assignment. Steps:

1. Clusters are sampled with probability P_{C_n}
2. Sample units from those with probability P_{U_n}

Thus, if $P_{C_n} \approx 0$, then it is the case covered by the LZ estimator.

Assignment process determines W_{in}. Steps:

1. For a cluster c in population n there is an assignment probability $q_{cn} \in [0,1]$ that is drawn from a distribution $f(\cdot)$ with mean μ_n and variance σ_n^2 (note: if the variance is equal to 0, then we have random assignment, if the variance is positive then there is correlated assignment in clusters).
2. Each unit in c is assigned to treatment independently with a cluster-specific probability q_{cn}.

Therefore, as the random part of the estimator comes from the variables R_{in} and W_{in}, then the mean and the variance of it depend on the first and second cross-moments of those variables:

Variable	Expected Value	Variance	Within Cluster Covariance
R_{in}	$P_{Cn}P_{Un}$	$P_{Cn}P_{Un}\left(1-P_{Cn}P_{Un}\right)$	$P_{Cn}\left(1-P_{Cn}\right)P_{Un}^2$
W_{in}	$1/2$	$1/4$	σ_n^2
$R_{\text{in}}\, W_{\text{in}}$	$P_{Cn}P_{Un}/2$	$P_{Cn}P_{Un}\left(2-P_{Cn}P_{Un}\right)/4$	$P_{Cn}P_{Un}^2\left(1-P_{Cn}\right)/4+\sigma_n^2 P_{Cn}P_{Un}^2$

Note that the within cluster covariance of R_{in} is zero if $P_{C_n} = 0$ or $P_{C_n} = 10$, which indicates that either all clusters are sampled or only a vanishing fraction. The within cluster covariance of W_{in} is zero if the assignment probability is constant across clusters ($\sigma_n^2 = 0$).

The estimator

We are interested in the least squares estimator for τ in the regression

$$Y_{in} = \alpha + \tau W_{in} + \varepsilon_{in}.$$

Define the averages

$$\bar{R}_n = \tfrac{1}{M_n}\textstyle\sum_{i=1}^{M_n} R_{in}, \quad \bar{W}_n = \tfrac{1}{N_n}\textstyle\sum_{i=1}^{M_n} R_{in}W_{in},$$
$$\bar{Y}_n = \tfrac{1}{N_n}\textstyle\sum_{i=1}^{M_n} R_{in}Y_{in}.$$

Note that except for $\bar{R}_n$ these averages are defined over the units in the sample, not the units in the population. Now we can write the least squares estimator $\hat{\tau}$ as

$$\hat{\tau} = \frac{\sum_{i=1}^{n} R_{in}\left(W_{in} - \bar{W}_n\right) Y_{in}}{\sum_{i=1}^{n} R_{in}\left(W_{in} - \bar{W}_n\right)^2}$$

Note further that the estimator $\hat{\tau}$ depends on the random variables R_{in}, W_{in} and Y_{in}.

The main formal result of the paper is then to compute the exact variance of this expression for the estimator and then to compare it to the previous proposed estimated variances, as shown in the following proposition:

Proposition 1. Suppose Assumptions 1-5 hold. Then (i), the exact variance of η_n is

$$\begin{aligned}\mathbb{V}\left(\eta_n\right) = \frac{1}{M_n} \sum_{i=1}^{M_n} &\left\{2\left(\varepsilon_{in}(1)^2 + \varepsilon_{in}(0)^2\right) - P_{Un}\left(\varepsilon_{in}(1) - \varepsilon_{in}(0)\right)^2 + 4P_{Un}\sigma_n^2\left(\varepsilon_{in}(1) - \varepsilon_{in}(0)\right)^2\right\} \\ &+ \frac{P_{Un}}{M_n} \sum_{c=1}^{C_n} M_{cn}^2 \left\{\left(1 - P_{Cn}\right)\left(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0)\right)^2 + 4\sigma_n^2\left(\bar{\varepsilon}_{cn}(1) + \bar{\varepsilon}_{cn}(0)\right)^2\right\}\end{aligned}$$

(ii) the difference between the limit of the normalized LZ variance estimator and the correct variance is

$$\mathbb{V}_{\mathrm{LZ}} - \mathbb{V}\left(\eta_n\right) = \frac{P_{Cn}P_{Un}}{M_n} \sum_{c=1}^{C_n} M_{cn}^2 \left(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0)\right)^2 \geq 0,$$

and (iii), the difference between the limit of the normalized LZ and EHW variance estimators is

$$\begin{aligned}\mathbb{V}_{\mathrm{LZ}} - \mathbb{V}_{\mathrm{EHW}} = -\frac{2P_{Un}}{M_n} \sum_{i=1}^{M_n} &\left\{\left(\varepsilon_{in}(1) - \varepsilon_{in}(0)\right)^2 + 4\sigma^2\left(\varepsilon_{in}(1) + \varepsilon_{in}(0)\right)^2\right\} \\ &+ \frac{P_{Un}}{M_n} \sum_{c=1}^{C_n} M_{cn}^2 \left\{\left(\bar{\varepsilon}_{cn}(1) - \bar{\varepsilon}_{cn}(0)\right)^2 + 4\sigma^2\left(\bar{\varepsilon}_{cn}(1) + \bar{\varepsilon}_{cn}(0)\right)^2\right\}.\end{aligned}$$

Main intuitions of Proposition 1

Part (i): The first sum is approximately the EHW variance, while the second sum then reflects the effects of the the sampling and assignment mechanisms of clusters. The second sum disappears if: $P_{C_n} = 1$, in which case we have a non clustered sample, and $\sigma_n^2 = 0$, meaning that we have no clustering in treatment assignment. → Indicates that one should cluster only when we have a clustered sample (not all clusters are sampled) and there is clustering in the treatment assignment (some clusters are more likely to receive treatment than others).

Part (ii): There is a difference between the LZ variance and the exact variance if P_{C_n} is not close to zero. Therefore, the LZ variance captures the clustered assignment to treatment, but not the sampling assignment, unless it is the case that only a small proportion of clusters are sampled.→ Indicates that the LZ variance estimator only correctly takes into account clustering if a small proportion of clusters are sampled.

Part (iii): The difference between the EHW and LZ variances remain if M_{cn} is constant and large relative to C, which is when the second term dominates.

3 References

Abadie, A., Athey, S., Imbens, G. W., Wooldridge, J. (2017). When should you adjust standard errors for clustering? (No. w24003). National Bureau of Economic Research.

Notes Tutorial 3 - Advanced Econometrics II - Tinbergen Institute

Gabriela M. Miyazato Szini (g.m.m.szini@uva.nl)

January 27, 2022

1 Recap of IV/2SLS Estimators

1.1 IV Estimator

Consider the following model:

$$Y = X\beta + u$$

with k regressors X and k instruments Z.
Consider also the following two assumptions:

- Instrument validity: $\text{Cov}(Z, u) = \mathbb{E}(Z'u) = 0$, provided that $\mathbb{E}(u) = 0$
- Relevance: $\mathbb{E}\left(Z'X\right)$ is non singular or: $\text{Cov}(Z, X) \neq 0$

Then, from validity, $\mathbb{E}(zu) = 0$, it follows that:

$$\mathbb{E}(Z'Y) = \mathbb{E}(Z'X)\beta$$

$$\beta = (\mathbb{E}(Z'X))^{-1}\mathbb{E}(Z'Y)$$

which is a valid expression provided that $\mathbb{E}(Z'X)$ is non-singular (relevance condition). An estimator for this quantity is:

$$\hat{\beta}_{IV} = (Z'X)^{-1}(Z'Y)$$

which is the known IV estimator. However, the number of instruments r may differ from k, leading to the 2SLS estimator in the case that $r > k$. We can consider 3 cases:

- $r < k$: $\mathbb{E}(Z'Y) = \mathbb{E}(Z'X)\beta$ has no solution
- $r = k$: β is just-identified
- $r > k$: β is over-identified, that is, there are more equations than unknowns. The idea of the 2SLS in this case is to use linear combinations of instruments.

1.2 2SLS estimator

Consider now the following model:

$$Y = X\beta + u$$

$$X = \Pi' Z + \nu$$

where X is a matrix. Moreover, the following assumptions hold:

$$\mathbb{E}[Z'u] = 0 \qquad \mathbb{E}[Z'\nu] = 0$$

Then, one can estimate β by 2SLS with the following steps:

1. Estimate Π by OLS: $\hat{\Pi} = (Z'Z)^{-1}Z'Y$
2. Define $\hat{X} = Z\hat{\Pi}$ and estimate the OLS for the model:

$$Y = \hat{X}\beta + \varepsilon$$

leading to:

$$\hat{\beta}_{2SLS} = (\hat{X}'\hat{X})^{-1}\hat{X}'Y$$

When working out the expression for the $\hat{\beta}_{2SLS}$, we can see that it resembles the previous expression for the IV estimator, $\hat{\beta}_{IV}$ where instead of taking into account Z as instruments, we take into account $\hat{X} = Z\hat{\Pi}$ as instruments, which essentially are a linear combination of the original instruments. To do so, first, remember that:

$$\hat{X} = Z\hat{\Pi} = Z(Z'Z)^{-1}Z'X \quad \text{ and } \quad Z(Z'Z)^{-1}Z' = P_Z$$

where P_Z is a projection matrix. Then:

$$\begin{aligned}
\hat{\beta}_{2SLS} &= \left(\hat{X}'\hat{X}\right)^{-1}\hat{X}'Y \\
&= \left[X'Z\left(Z'Z\right)^{-1}Z'X\right]^{-1}X'Z\left(Z'Z\right)^{-1}Z'Y \\
&= \left(X'P_ZX\right)^{-1}X'P_ZY \\
&= \left(\hat{X}'X\right)^{-1}\hat{X}'Y
\end{aligned}$$

Note that the expression in the last line is very similar to that of the β_{IV}, what differs is that instead of having directly the instruments Z, we have the fitted values $\hat{X}$, which are essentially a linear combination of the instruments.

2 Testing for instrument validity - OIR test (exogeneity)

For more details on this section, please refer to sections 6.3.8 and 8.4.4 in Cameron & Trivedi.

Consistency of the IV/2SLS estimators requires that $\mathbb{E}(Z'u) = 0$, that is, that the instruments Z are exogenous. The idea of the OIR/Sargan test is to test for this condition.

- In case of just identification ($r = k$) this condition cannot be tested, as $Z'\hat{u} = 0$ was already imposed for the estimation.
- In case of overidentification ($r > k$) (case where we have more equations from FOC then unknown parameters) we can use k instruments given by $\hat{X} = Z\hat{\Pi}$ for estimation and test the validity of the remaining $r - k$ instruments.

- The Sargan test then rejects $H_0 : \mathbb{E}[Z'u] = 0$ when:

  $$\textbf{OIR} = \frac{\hat{u}'Z\left(Z'Z\right)^{-1}Z'\hat{u}}{s^2} = \frac{\hat{u}'P_Z\hat{u}}{s^2}$$

  exceeds the $\chi^2(r-k)$ critical value. The idea is that when the exogeneity condition is met, the projection of $\hat{u}$ spanned into the space of Z should be 0, as they should be orthogonal.

The aim of this section is then to explain why the test has this functional form, and why we have that it follows this distribution with (r-k) degrees of freedom.

Consider the model:

$$\begin{aligned} Y = X\beta + u \quad & X \text{ is } N \times k \\ & u_i \sim N\left(0, \sigma^2\right) \end{aligned}$$

Further assume that X is endogenous and that we have a set of instruments Z that is $N \times r$, with $r > k$, that is, more instruments than endogenous regressors.

We can partition the set of instruments such that:

$$Z = [Z_1 \quad Z_2]$$

with Z_1 being $(N \times k)$ and Z_2 being $(N \times (r-k))$. Then, we can also rewrite the first stage model as:

$$\begin{aligned} X &= Z_1\Pi_1 + Z_2\Pi_2 + v \\ &= Z\Pi + v \end{aligned}$$

where Π_1 is a matrix of parameters $(k \times k)$ and Π_2 is $(r-k \times k)$.

We can then plug-in the model for X in the expression for Y:

$$\begin{aligned} Y &= X\beta + u \\ &= (Z_1\Pi_1 + Z_2\Pi_2 + v)\beta + u \\ &= Z_1\Pi_1\beta + Z_2\Pi_2\beta + \varepsilon \end{aligned}$$

where $\varepsilon = u + v\beta$ is a new defined error term. The validity of instruments is still given by $\mathbb{E}(u|Z) = 0$, or $\mathbb{E}(Z'u) = 0$.

Criterion function

The form of the OIR test originates from the criterion function of the 2SLS estimator (from the GMM literature), which ideally you want it to be as close to zero as possible (reflecting the moment condition given by the exclusion restriction of instruments). Therefore, we will first define which is this criterion function, and then show that for the just-identified case it will always be equal to zero, while in the over-identified case it is not necessarily. The criterion function is defined as:

$$Q(\beta) = \left(Z'(Y - X\beta)\right)'\left(Z'Z\right)^{-1}Z'(Y - X\beta)$$

This expression comes from the moment $\mathbb{E}(Z'u) = 0$, we are minimizing the square of this expression weighted by $(Z'Z)^{-1}$. This weighting matrix is not optimal in the sense that it does not lead to the smallest variance for the GMM estimator, but it is the weight used to obtain the 2SLS estimator in a GMM framework.

We can further rewrite the criterion function:

$$\begin{aligned} Q(\beta) &= (Y - X\beta)' \underbrace{Z\left(Z'Z\right)^{-1} Z'}_{P_Z}(Y - X\beta) \\ &= (Y - X\beta)' P_Z (Y - X\beta) \end{aligned}$$

Now, we separate into two cases, the just-identified and the over-identified, and we evaluate the criterion function at the IV and the 2SLS estimators, respectively.

Just-identified case: it is equivalent to taking into account only Z_1 as instruments. The estimator is given by $\hat{\beta}_{IV} = (Z_1'X)^{-1}Z_1'Y$. Evaluating the criterion function at this estimator:

$$\begin{aligned} Q\left(\hat{\beta}_{IV}\right) &= \left(Y - X\left(Z_1'X\right)^{-1} Z_1'Y\right)' P_{Z_1} \left(Y - X\left(Z_1'X\right)^{-1} Z_1'Y\right) \\ &= \left(P_{Z_1}Y - P_{Z_1}X\left(Z_1'X\right)^{-1} Z_1'Y\right)' \left(P_{Z_1}Y - P_{Z_1}X\left(Z_1'X\right)^{-1} Z_1'Y\right) \end{aligned}$$

where in the second line we used the fact that $P_{Z_1} = P_{Z_1}'P_{Z_1}$. Considering then only the first part of the term in the second line:

$$\begin{aligned} P_{Z_1}Y - P_{Z_1}X\left(Z_1'X\right)^{-1} Z_1'Y &= P_{Z_1}Y - Z_1\left(Z_1'Z_1\right)^{-1} Z_1X\left(Z_1'X\right)^{-1} Z_1'Y \\ &= P_{Z_1}Y - Z_1\left(Z_1'Z_1\right)^{-1} Z_1'Y \\ &= P_{Z_1}Y - P_{Z_1}Y = 0 \end{aligned}$$

Therefore, evaluating it at the estimator β_{IV} will always lead 0.

Over-identified case: we now take into account the entire set of Z instruments. The estimator is given by $\hat{\beta}_{GIV} = (X'P_ZX)^{-1}X'P_ZY$. Evaluation the criterion function at this estimator:

$$\begin{aligned} Q\left(\hat{\beta}_{\text{GIV}}\right) &= \left(Y - X\left(X'P_ZX\right)^{-1} X'P_ZY\right)' P_Z \left(Y - X\left(X'P_ZX\right)^{-1} X'P_ZY\right) \\ &= \left(P_ZY - P_ZX\left(X'P_ZX\right)^{-1} X'P_ZY\right)' \left(P_ZY - P_ZX\left(X'P_ZX\right)^{-1} X'P_ZY\right) \\ &= (P_ZY - P_{P_ZX}Y)'(P_ZY - P_{P_ZX}Y) \\ &= Y'(P_Z - P_{P_ZX})'(P_Z - P_{P_ZX})Y \\ &= Y'(P_Z - P_{P_ZX})Y \end{aligned}$$

where in the second line we used the fact that $P_Z = P_Z'P_Z$, in the third line, we used the fact that $P_ZX\left(X'P_ZX\right)^{-1} X'P_Z$ has the form of a projection matrix that we denote by P_{P_ZX}, and finally in the last line we use the fact that as both P_Z and P_{P_ZX} are projection matrices, the term $(P_Z - P_{P_ZX})$ is symmetric and idempotent, so the square is itself. We can further open up the expression for Y to obtain:

$$Q\left(\hat{\beta}_{\text{GIV}}\right) = (X\beta + u)'(P_Z - P_{P_Z X})(X\beta + u)$$

We can further simplify this expression, given that:

$$\begin{aligned}(X\beta)'(P_Z - P_{P_Z X}) &= \beta'(X'P_Z - X'P_{P_Z X}) \\ &= \beta'(X'P_Z - X'P_Z X\left(X'P_Z X\right)^{-1} X'P_Z) \\ &= \beta'(X'P_Z - X'P_Z) = 0\end{aligned}$$

Therefore:

$$Q\left(\hat{\beta}_{\text{GIV}}\right) = u'(P_Z - P_{P_Z X})u$$

Our goal is to compare this expression to the numerator of the OIR test, namely, $\hat{u}'Z(Z'Z)^{-1}Z'\hat{u}$, given that the denominator is only a normalization to obtain the distribution of the test. It is clear that the two expressions differ for now in two aspects: the test considers the residuals instead of the error term, and the test does not contain the term $P_{P_Z X}$. However, intuitively it is clear that $u'P_{P_Z X} = 0$, once those terms should be orthogonal for the OLS estimation in the second step: this terms reflect the projection of the errors into the space spanned by the part of X that is explained by Z, meaning $\hat{X}$, and by OLS assumptions the errors and $\hat{X}$ should be orthogonal. To formally show the equivalence between the expressions, we take 4 additional steps.

Step 1. Note that $(P_Z - P_{P_Z X}) = P_Z(I - P_{P_Z X})$ This follows since $P_Z P_{P_Z X} = P_{P_Z X}$:

$$\begin{aligned}&P_Z P_{P_Z X} = Z\left(Z'Z\right)^{-1} Z'P_Z X\left(X'P_Z'P_Z X\right)^{-1} X'P_Z \\ &= Z\left(Z'Z\right)^{-1} Z'Z\left(Z'Z\right)^{-1} Z'X\left(X'P_Z'P_Z X\right)^{-1} X'P_Z \\ &= P_Z X\left(X'P_Z'P_Z X\right)^{-1} X'P_Z = P_{P_Z X}\end{aligned}$$

Step 2. We also have that as P_Z is a projection matrix and $(I - P_{P_Z X})$ is a residual maker matrix, that we can denote by $M_{P_Z X}$:

$$P_Z(I - P_{P_Z X}) = (I - P_{P_Z X})'P_Z P_Z(I - P_{P_Z X})$$

Then:

$$\begin{aligned}Q\left(\hat{\beta}_{\text{GIV}}\right) &= u'(I - P_{P_Z X})'P_Z P_Z(I - P_{P_Z X})u \\ &= u'M_{P_Z X} P_Z M_{P_Z X} u\end{aligned}$$

Step 3. We also know that $Z\hat{\Pi} = Z(Z'Z)^{-1}Z'X = P_Z X$, and therefore $M_{P_Z X} = M_{Z\hat{\Pi}}$. Therefore:

$$Q\left(\hat{\beta}_{\text{GIV}}\right) = u'M_{Z\hat{\Pi}} P_Z M_{Z\hat{\Pi}} u$$

Substituting $u = Y - X\beta$:

$$Q\left(\hat{\beta}_{\text{GIV}}\right) = (Y - X\beta)'M_{Z\hat{\Pi}} P_Z M_{Z\hat{\Pi}}(Y - X\beta)$$

Step 4: As we know that $\hat{u} = M_{Z\hat{\Pi}}Y$, to complete the proof it suffices to show that $(X\beta)'M_{Z\hat{\Pi}}P_Z = 0$:

$$\begin{aligned}(X\beta)'M_{Z\hat{\Pi}}P_Z &= (X\beta)'M_{P_Z X}P_Z \\ &= (X\beta)'(I - P_Z X(X'P_Z'P_Z X)^{-1}X'P_Z)P_Z \\ &= (X\beta)'P_Z - \beta'X'P_Z X(X'P_Z X)^{-1}X'P_Z P_Z \\ &= (X\beta)'P_Z - \beta'X'P_Z = 0\end{aligned}$$

Therefore, the proof is complete.

Distribution of the OIR test

In general, for a random variable $\varepsilon \sim N(0,1)$, we have that:

$$\varepsilon' P_X \varepsilon \sim \chi^2(k)$$

where k is the rank of the projection matrix P_X:

$$\begin{aligned}k = \text{rank}(P_X) = \text{tr}(P_X) &= \text{tr}(X(X'X)^{-1}X') \\ &= \text{tr}((X'X)^{-1}X'X) \\ &= \text{tr}(I_k) = k\end{aligned}$$

where in the second line we used that $\text{tr}(AB) = \text{tr}(BA)$. Now we can apply the above to:

$$\frac{1}{s^2}u'(P_Z - P_{P_Z X})u \sim \chi^2(q)$$

where we scale for $\frac{1}{s^2}$ such that $\frac{1}{s}u \sim N(0,1)$. We only need to find now that $q = r - k$:

$$\begin{aligned}q &= \text{rank}(P_Z - P_{P_Z X}) \\ &= \text{tr}(P_Z) - \text{tr}(P_{P_Z X}) = \text{tr}(I_r) - \text{tr}(I_k) = r - k\end{aligned}$$

3 Handy projection matrix properties

$P_X = X\left(X'X\right)^{-1}X$
$P_X X = X$
$P_X' P_X = P_X$
$P_X' = P_X$
$M_X = I - P_X$
$P_X + M_X = I$
$P_X M_X = 0$
$M_{X'} = M_X$
$M_X' M_X = M_X$
$M_X X = 0$

Notes Tutorial 4 - Advanced Econometrics II - Tinbergen Institute

Gabriela M. Miyazato Szini (g.m.m.szini@uva.nl)

February 8, 2022

1 Method of Moments and Generalized Method of Moments

Method of Moments: number of moment conditions is equal to the number of unknown parameters to be estimated.
Generalized Method of Moments (GMM): number of moment conditions is bigger than the number of estimated parameters.

1.1 Two stage estimation in nonlinear models

We start with a brief review of GMM. We have seen in the lecture that the GMM estimation of models based on non-linear moment conditions such as,

$$\mathbb{E}[r(y_i, x_i, \beta)|z_i] = 0$$

that can be given through some economic theory that implies a conditional moment, leads to, according to the notation of the lecture slides,

$$h_i(\beta) = z_i r(y_i, x_i, \beta)$$

which, in turn will imply in an unconditional moment. This follows since:

$$\mathbb{E}[h_i(\beta)] = \mathbb{E}[z_i r(y_i, x_i, \beta)] = \mathbb{E}\left[\mathbb{E}[z_i r(y_i, x_i, \beta)|z_i]\right] = \mathbb{E}\left[z_i \mathbb{E}[r(y_i, x_i, \beta)|z_i]\right] = 0$$

We denote that $h_i(\beta)$ is a vector of size r, and, therefore, r denotes the number of moment conditions, while β has dimension q. If $r = q$, one can simply obtain the estimates of the parameters by taking the sample analogue of the expression $\mathbb{E}[h_i(\beta)] = 0$. However, when $r > q$, one instead look at the set of parameters that minimizes the weighted euclidean length of the vector defined by $g_N(\beta) := \frac{1}{N}\sum_{i=1}^{N} h_i(\beta)$, the sample analogue of the moment condition. Therefore, it minimizes the following criteria function:

$$Q_N(\beta) = g_N(\beta)' W_N g_N(\beta)$$

Moreover, denoting the following vectors of derivatives:

$$G_N(\beta) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial h_i(\beta)}{\partial \beta'} = \frac{1}{N}\sum_{i=1}^{N} z_i d_i(\beta)' = \frac{1}{N} Z' D\left(\beta\right)$$

where $D(\beta)$ has rows $d_i(\beta)' = \frac{\partial r(y_i,x_i,\beta)}{\partial \beta'}$. Therefore, the FOC of $Q_N(\beta)$ is given by:

$$D(\beta)'ZW_N Z'r(\beta) = 0$$

and the general expression for the asymptotic variance of the resulting estimator applies with $\hat{G} = N^{-1}Z'D(\hat{\beta})$.

However, while the GMM method translates easily to a 2SLS approach in linear models, it does not generalize straightforwardly to non-linear models. That is, taking the OLS fitted values $\hat{x_i} = z_i'\hat{\Pi}$ from a first stage estimation and minimizing $\frac{1}{N}\sum_{i=1}^{N} r(y_i, \hat{x_i}, \beta)^2$ is not equivalent to GMM and leads to inconsistent estimators.

To illustrate this problem, let's consider the following model, considering the variables to be scalars, and where the errors may or may not be correlated:

$$\begin{aligned} y_i &= \beta x_i^2 + u_i \\ x_i &= \pi z_i + v_i \end{aligned}$$

The 2SLS procedure would suggest then to regress first x on z and obtain the fitted values $\hat{x}$, and then regress y on $\hat{x}^2$. However, this procedure leads to wrong standard errors in the second stage and it breaks down if you have a nonlinear model, that is, if we take $y = g(\hat{x}, \beta) + u$, with $g(\cdot)$ a nonlinear function either in the parameters or in the variables, then it will lead to inconsistent estimates of β.

Note that in the **case of IV (just identified)**, this is not particularly a problem, since one does not rely on the 2SLS procedure, and the IV estimator can simply be given by:

$$\hat{\beta}_{IV} = (z'x^2)^{-1}z'y = \left(\sum_{i=1}^{N} z_i x_i^2\right)^{-1} \sum_{i=1}^{N} z_i y_i$$

which is implemented by a regular IV regression of y on x^2, with instruments z. Also, the IV estimator can be shown to be equal to the nonlinear IV defined by the sample analogue moment conditions:

$$\frac{1}{N}\sum_{i=1}^{N} z_i u_i = 0$$

In the **case of two stage least squares**, in the second stage we would regress y on $(\hat{x})^2$, that is, the square of the fitted values, delivering:

$$\hat{\beta}_{2SLS} = \left(\sum_{i=1}^{N} (\hat{x_i})^2 (\hat{x_i})^2\right)^{-1} \left(\sum_{i=1}^{N} (\hat{x_i})^2 y_i\right)$$

However, this is an inconsistent estimate. To prove that, we start by writing:

$$\begin{aligned} y_i &= \beta {x_i}^2 + u_i \\ &= \beta(\hat{x_i})^2 + w_i \end{aligned}$$

where w_i is a new error term defined by $w_i = \beta(x_i^2 - (\hat{x}_i)^2) + u_i$. Taking into account, from the model, that $x_i = \pi z_i + v_i$, we can further open up the expression $(x_i^2 - (\hat{x}_i)^2)$:

$$\begin{aligned} x_i^2 - (\hat{x}_i)^2 &= (\pi z_i + v_i)^2 - (\hat{\pi} z_i)^2 \\ &= \pi^2 {z_i}^2 + 2\pi z_i v_i + v_i^2 - \hat{\pi}^2 {z_i}^2 \end{aligned}$$

With some abuse of notation, as long as $\hat{\pi}$ is a consistent estimator for π, the first and the last terms of the latter expression cancels out:

$$x_i^2 - (\hat{x}_i)^2 = 2\pi z_i v_i + v_i^2$$

Given this result, it is easy to see that the regressor $(\hat{x}_i)^2$ is correlated with the error, leading to inconsistency of the estimates:

$$(\hat{x}_i)^2 \left({x_i}^2 - (\hat{x}_i)^2\right) = (\hat{x}_i)^2 \left(2\pi z_i v_i + {v_i}^2\right) = \hat{\pi}^2 {z_i}^2 \left(2\pi z_i v_i + {v_i}^2\right)$$

Taking the probability limit of this term:

$$\begin{aligned} \text{plim}\frac{1}{N}\sum_{i=1}^{N}(\hat{x}_i)^2 \left({x_i}^2 - (\hat{x}_i)^2\right) &= \text{plim}\frac{1}{N}\sum_{i=1}^{N}\left(2\pi^3 z_i^3 v_i + \pi^2 z_i^2 v_i^2\right) \\ &= \text{plim}\frac{1}{N}\sum_{i=1}^{N}\left(\pi^2 z_i^2 v_i^2\right) \neq 0 \end{aligned}$$

where, to reach to the final expression we used the fact that by assumption z_i and v_i are independent, and that v_i has expected value zero. Moreover, even if they are independent, we do not assume that the variance of v_i for instance, is zero, leading to the final result.

Therefore, $\text{plim}N^{-1}\sum_{i=1}^{N}(\hat{x}_i)^2 w_i \neq 0$, and the regressor is asymptotically correlated with the composite error term w_i, leading to inconsistency. Note that this problem mostly arises as:

$$\text{plim}N^{-1}\sum_{i=1}^{N}(\hat{x}_i)^2 w_i = \text{plim}N^{-1}\sum_{i=1}^{N}(\hat{x}_i)^2 \left(\beta(x_i^2 - (\hat{x}_i)^2) + u_i\right) \neq \text{plim}N^{-1}\sum_{i=1}^{N}(\hat{x}_i)^2 \left(\beta(x_i - \hat{x}_i)^2 + u_i\right) = 0$$

Given this latter observation, there is a **variation that is consistent**, in which in the first stage we regress x^2 directly on z, rather than x on z, the we use the prediction $\widehat{x^2}$ ($\neq \hat{x}^2$) in the second stage. In the scalar case, we can show that this equals β_{IV}.

$$
\begin{aligned}
\hat{\beta}_{2SLS} &= \left(\sum_{i=1}^{N}\widehat{x_i^2}\widehat{x_i^2}\right)^{-1}\sum_{i=1}^{N}\widehat{x_i^2}y_i \\
&= \left(\widehat{x^2}'\widehat{x^2}\right)^{-1}\widehat{x^2}'y \\
&= \left(z'\hat{\pi}\hat{\pi}z\right)^{-1}z'\hat{\pi}y \\
&= \left(z'\hat{\pi}z\right)^{-1}z'y \\
&= \left(z'z\hat{\pi}\right)^{-1}z'y \\
&= \left(z'z(z'z)^{-1}z'x^2\right)^{-1}z'y \\
&= (z'x^2)^{-1}z'y = \hat{\beta}_{IV}
\end{aligned}
$$

In the result above we use the fact that $\hat{\pi}$ is a scalar, and that $\hat{\pi} = (z'z)^{-1}z'x^2$. Moreover, this example generalizes to other non-linear models where the nonlinearity is only in the regressors, such that $y = g(x)'\beta + u$. Then, we use $\widehat{g(x)}$ rather than $g(\hat{x})$ as instruments for $g(x)$.

1.2 Optimal moment conditions/Optimal instruments

Once again, we start with essentially the same recap of GMM estimators for the overidentified case. We estimate the parameters θ (change of notation!) by the value that makes the weighted squared Euclidean length the smallest:

$$Q_N(\theta) = g_N(\theta)'W_N g_N(\theta)$$

where $g_N(\theta) = \frac{1}{N}\sum_{i=1}^{N}h(w_i,\theta)$ is the sample analogue of the moment conditions $\mathbb{E}[h(w;\theta_0)] = 0$, w includes the vectors y, x and z. Moreover, the weight matrix satisfies $W_N \xrightarrow{p} W_0 > 0$, intuitively, it gives different weights to different moments.

The FOC is given by:

$$\left.\frac{\partial g_N(\theta)'}{\partial\theta}\right|_{\theta=\hat{\theta}} \times W_N \times g_N(\hat{\theta}) = 0$$

Defining the derivative matrix: $G_N(\theta) = \frac{\partial g_N(\theta)}{\partial\theta'} = \frac{1}{N}\sum_{i=1}^{N}\frac{\partial h_i(\theta)}{\partial\theta'}$, with $h_i(\theta) = h(w_i,\theta)$, the FOC is then:

$$G_N(\hat{\theta})'W_N g_N(\hat{\theta}) = 0$$

And, by mean-value expansion, we can then obtain the asymptotic distribution of the estimator given by:

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left[0,\left(G_0'W_0G_0\right)^{-1}G_0'W_0S_0W_0G_0\left(G_0{}'W_0G_0\right)^{-1}\right]$$

The term S_0 is essentially the variance of the moment conditions, given by:

$$\sqrt{N}g_N\left(\theta_0\right) = \frac{1}{\sqrt{N}}\sum_{i=1}^{N}h_i\left(\theta_0\right) \xrightarrow{d} N\left[0,S_0\right]$$

with $S_0 = \mathbb{E}[h(\theta)h(\theta)']^{-1}$. The idea behind the optimal weighting matrix is that, if we choose $W_0 = S_0^{-1}$, then the asymptotic variance is reduced, leading to efficiency:

$$\sqrt{N}\left(\hat{\theta}-\theta_0\right) \xrightarrow{d} N\left[0, \left(G_0^1 S_0^{-1} G_0\right)^{-1}\right]$$

This efficiency is only for a given initial moment condition. **However, we can also choose optimal moment conditions that leads to efficiency.**

Suppose you have the conditional moment restriction:

$$\mathbb{E}[r(y_i, x_i, \theta_0)|Z] = 0$$

We showed before that we can rewrite this moment condition as $\mathbb{E}[Zr(y_i, x_i, \theta_0)] = 0$. However, in this expression, instead of Z, we could also have any function of Z, say $D(Z)$, leading to:

$$\mathbb{E}[D(Z)'r(y, x, \theta_0)] = 0$$

From this moment condition, we could proceed normally to the GMM approach, what matters is which choice of $D(Z)$ should be made. To do so, we look at the terms of the asymptotic variance of the GMM for this particular moment condition given by $(G_0' S_0^{-1} G_0)^{-1}$. The terms are the following:

$$\begin{aligned} G_0 &= \mathbb{E}[D(Z)'J] \\ &= \mathbb{E}[\mathbb{E}[D(Z)'J|Z]] \\ &= \mathbb{E}[D(Z)'\mathbb{E}[J|Z]] \end{aligned}$$

where $J_i = \frac{\partial r(y_i, x_i, \theta_0)}{\partial \theta'}$.

$$\begin{aligned} S_0 &= \mathbb{E}[D(Z)'r(y, x, \theta_0)r(y, x, \theta_0)'D(Z)] \\ &= \mathbb{E}[D(Z)'\mathbb{E}[r(y, x, \theta_0)r(y, x, \theta_0)'|Z]D(Z)] \\ &= \mathbb{E}[D(Z)'\Omega D(Z)] \end{aligned}$$

where, we denote $\mathbb{E}[r(y, x, \theta_0)r(y, x, \theta_0)'|Z]$ by Ω. Given the above expressions for G_0 and S_0, we can now get the expression for the asymptotic variance.

$$(G_0' S_0^{-1} G_0)^{-1} = \mathbb{E}\left[\mathbb{E}[J'|Z]D(Z)\left(D'(Z)\Omega D(Z)\right)^{-1} D'(Z)\mathbb{E}[J|Z]\right]^{-1}$$

Denoting $\mathbb{E}[J|Z] = A$, and proposing $D(Z) = \Omega^{-1}A$, which resembles a FGLS approach, and plugging into the expression, we obtain:

$$(G_0' S_0^{-1} G_0)^{-1} = \mathbb{E}\left[A'\Omega^{-1}A\left(A'\Omega^{-1}\Omega\Omega^{-1}A\right)^{-1} A'\Omega^{-1}A\right]^{-1} =$$

Which looks like a smaller asymptotic variance. To be formal about this result, we can compare this variance with the standard one, for which no proposal was plugged in, and look at precision instead of variance:

$$
\begin{aligned}
A'\Omega^{-1}A - A'D(Z)\left(D'(Z)\Omega D(Z)\right)^{-1}D'(Z)A &= A'\Omega^{-1/2}\left(I - \Omega^{1/2}D(Z)\left(D'(Z)\Omega D(Z)\right)^{-1}D(Z)'\Omega^{1/2}\right)\Omega^{-1/2}A \\
&= A'\Omega^{-1/2}M_{\Omega^{1/2}D(Z)}\Omega^{-1/2}A \geq 0
\end{aligned}
$$

where we use that, by Choleski decomposition, $\Omega^{-1} = (\Omega^{-1/2})'\Omega^{-1/2}$, and we reach to the conclusion of it being positive since the final term is a quadratic term.

Therefore, the optimal moment is given by the function:

$$
\begin{aligned}
D(Z) &= \Omega^{-1}A \\
D(Z_i) &= \frac{1}{\operatorname{Var}(r(y_i, x_i, \theta_0)|Z_i)}\mathbb{E}\left[\frac{\partial r\left(y_i, x_i, \theta_0\right)}{\partial \theta_0'} \mid Z_i\right]
\end{aligned}
$$

As noticed before, there are some similarities with the GLS approach. For the GLS we would get the following moment condition:

$$
\mathbb{E}[X'\Omega^{-1}(y - X\beta)] = 0
$$

which leads to the estimator $\hat{\beta} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y$, which can be written in terms of the transformed data $X^* = \Omega^{-1/2}X$ as: $\hat{\beta} = (X^{*\prime}X^*)^{-1}X^{*\prime}y^*$.

This result of optimal moment conditions also has implications for choosing **optimal instruments**. For instruments Z for endogenous regressors X, the optimal moment condition would be:

$$
\mathbb{E}[Z'\Omega^{-1}(y - X\beta)] = 0
$$

Then, the asymptotic variance would be given by:

$$
\operatorname{plim}\frac{1}{N}\left(Z'\Omega^{-1}(y - X\beta)(y - X\beta)'\Omega^{-1}Z\right) = \operatorname{plim}\frac{1}{N}Z'\Omega^{-1}Z
$$

Moreover, the criteria function of the GMM and its associated FOC would be:

$$
Q_N(\beta) = \frac{1}{N}(y - X\beta)'\Omega^{-1}Z(Z'\Omega^{-1}Z)^{-1}Z'\Omega^{-1}(y - X\beta)
$$

$$
FOC: -\frac{2}{N}X'\Omega^{-1}Z\left(Z'\Omega^{-1}Z\right)^{-1}Z'\Omega^{-1}(y - X\beta) = 0
$$

Then, thinking in terms of transformed data, we can rewrite:

$$
X^{*\prime}Z^*\left(Z^{*\prime}Z^*\right)^{-1}Z^{*\prime}\left(y^* - X^*\beta\right) = 0
$$

Finally, the estimator is:

$$
\hat{\beta} = (X^{*\prime}P_{Z^*}X^*)^{-1}X^{*\prime}P_{Z^*}y^*
$$

Which is essentially the expression for the 2SLS estimator considering transformed data. That is, using optimal instruments is the same as using transformed data.

Notes Tutorial 5 - Advanced Econometrics II - Tinbergen Institute

Gabriela M. Miyazato Szini (g.m.m.szini@uva.nl)

February 8, 2022

1 Fixed and Random Effects

1.1 A Recap of the Lecture

Consider the following panel data model:

$$\begin{aligned} y_{it} &= \alpha_i + x_{it}'\beta + \varepsilon_{it} \\ \alpha_i &= \gamma_0 + z_i'\gamma_1 \end{aligned}$$

Related effects: α_i and x_{it} are correlated, so α_i is treated as a fixed nuisance parameter (FE).
Unrelated effects: α_i and x_{it} are independently distributed, so α_i is treated as a part of the error term (RE).

Random effects (RE) are unrealistic for observational data, since, typically, omitted and included regressors show multicollinearity.

Considering the model above, when we suppose that N ($i = 1, ..., N$) is small and T ($t = 1, ..., T$)is large one can consistently estimate the fixed effects α_i and the slope parameter β by including individual-specific dummies. The steps would then be to first create N dummy variables $d_{i,it}, ..., d_{N,it}$ with $d_{j,it} = 1$ if $j = i$ and 0 otherwise. Note also that the dummies are time invariant, $d_{j,it} = d_{j,i}$. Then, the model can be expressed as:

$$y_{it} = \alpha_1 d_{1,i} + ... + \alpha_N d_{N,i} + x_{it}'\beta + \varepsilon_{it}$$

The model is then estimated with OLS:

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} (y_{it} - \hat{y}_{it})^2$$

$$\min_{\alpha,\beta} \sum_{i=1}^{N} \sum_{t=1}^{T} \left(y_{it} - \hat{\alpha}_1 d_{1,i} - \cdots - \hat{\alpha}_N d_{N,i} - x_{it}'\hat{\beta} \right)^2$$

Note that we will then have $N + K$ parameters to be estimated, when N is large (tending to infinity), we have the **incidental parameter problem**, which is caused by the fact that we will then have many parameters to be estimated, while having only a fixed number of T observations to estimate each of then. Note however, that this causes inconsistency only for the fixed effects α, and not for the slope parameter β in linear models, since we will show later that the dummy approach leads to the same estimate of the within transformation algebraically.

However, for non-linear models such as probit and logit, the incidental parameter problems may lead to (asymptotically) biased estimates of the slope parameters as well.

Possible solutions for estimating a linear model with FE:

- **The within transformation**
  We first take the time average of the linear model above:

  $$\bar{y_i} = \alpha_i + \bar{x_i}'\beta + \bar{\varepsilon_i} \quad \bar{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

  Then we subtract this averages from the original model, resulting in:

  $$y_{it} - \bar{y_i} = (x_{it} - \bar{x_i})'\,\beta + \varepsilon_{it} - \bar{\varepsilon_i}$$

  Where clearly the fixed effects where differenced out, and one can simply carry an OLS estimation of the latter model.

- **First differences transformation**
  Essentially the same idea behind the previous transformation, but instead of subtracting the time average for each individual i, we subtract the model evaluated at the previous period $t-1$:

  $$\begin{aligned} y_{it} - y_{it-1} &= (x_{it} - x_{it-1})'\beta + (\varepsilon_{it} - \varepsilon_{it-1}) \\ \Delta y_{it} &= \Delta x'_{it}\beta + \Delta\varepsilon_{it} \end{aligned}$$

However, the solutions above do not work in some cases:

→ if the regressors do not change over time $x_{it} = x_i$
→ lagged variables cannot be a part of the original regressors, because otherwise the regressors of the transformed model are correlated with the error term, and an IV-GMM approach would be needed:

$$y_{it} - y_{it-1} = (y_{it-1} - y_{it-2})'\beta + \varepsilon_{it} - \varepsilon_{it-1}$$

2 Non-linear panel data

Examples: panel logit, poisson regression, etc.
We consider here a way of differencing out the fixed effects in a panel logit case, with $T = 2$ (but the method is easily extendable for $T > 2$.

Consider the binary dependent variable model:

$$y_{it} = \mathbb{1}\left\{x'_{it}\beta + \alpha_i + \varepsilon_{it} \geqslant 0\right\}, \quad t = 1, 2$$

If the error terms follow a logistic distribution, then we can write:

$$\mathbb{P}\left(y_{it} = 1 \mid x_{i1}, x_{i2}, \alpha_i, \beta\right) = \frac{\exp\left(x'_{it}\beta + \alpha_i\right)}{1 + \exp\left(x'_{it}\beta + \alpha_i\right)}$$

The ”trick” is then to consider 2 possible sets of outcomes in order to difference out the fixed effects:

$$A = \{y_{i1} = 0, y_{i2} = 1\}, \quad B = \{y_{i1} = 1, y_{i2} = 0\}$$

Then, using the expression above and from Bayes rule, we can derive the probability of one of the two sets occurring given that we observe only observations in one of the two sets:

$$\begin{aligned} \mathbb{P}\left(y_{i1} = 0, y_{i2} = 1 \mid y_{i1} + y_{i2} = 1, x_{i2}, x_{i2}, \alpha_i, \beta\right) &= \mathbb{P}\left(A \mid A \cup B, x_{i1}, x_{i2}, \alpha_i, \beta\right) \\ &= \frac{\mathbb{P}\left(A \mid x_{i1}, x_{i2}, \alpha_i, \beta\right)}{\mathbb{P}\left(A \mid x_{i1}, x_{i2}, \alpha_i, \beta\right) + \mathbb{P}\left(B \mid x_{i1}, x_{i2}, \alpha_i, \beta\right)} \end{aligned}$$

We also know, from the logistic distribution that:

$$\begin{aligned} \mathbb{P}\left(A \mid x_{i1}, x_{i2}, \alpha_i, \beta\right) &= \tfrac{1}{1+\exp\left(x_{i1}'\beta+\alpha_i\right)} \cdot \tfrac{\exp\left(x_{i2}'\beta+\alpha i\right)}{1+\exp\left(x_{i2}'\beta+\alpha_i\right)} \\ \mathbb{P}\left(B \mid x_{i1}, x_{i2}, \alpha_i, \beta\right) &= \tfrac{1}{1+\exp\left(x_{i2}'\beta+\alpha_i\right)} \cdot \tfrac{\exp\left(x_{i1}'\beta+\alpha_i\right)}{1+\exp\left(x_{i1}'\beta+\alpha_i\right)} \end{aligned}$$

By substituting these expressions in the previous equation we have:

$$\begin{aligned} \mathbb{P}\left(y_{i1} = 0, y_{i2} = 1 \mid y_{i1} + y_{i2} = 1, x_{i1}, x_{i2,}\alpha_i, \beta\right) &= \frac{\exp\left(x_{i2}'\beta + \alpha_i\right)}{\exp\left(x_{i2}'\beta + \alpha_i\right) + \exp\left(x_{i1}'\beta + \alpha_i\right)} \\ &= \frac{\exp\left(\Delta x_{i2}'\beta\right)}{1 + \exp\left(\Delta x_{i2}'\beta\right)} \end{aligned}$$

The final expression is then free of fixed effects, and it still follows a logistic function, and thus a logit model can be employed by redefining the dependent variable as:

$$w_i = \begin{cases} 1 & \text{if } (y_{i1}, y_{i2}) \in A \\ 0 & \text{if } (y_{i1}, y_{i2}) \in B \end{cases}$$

3 The dummy approach is algebraically equal to the within transformation

First, we stack the original panel data model over time:

$$y_i = e\alpha_i + x_i\beta + \varepsilon_i$$

where x_i is a $(T \times K)$ matrix of covariates, and e is a $(T \times 1)$ vector of ones. Then, we further stack over individuals:

$$y = \delta\alpha + x\beta + \varepsilon$$

where y and ε are now $(NT \times 1)$ vectors, x is a $(NT \times K)$ matrix of covariates, α is the vector of fixed effects, $\alpha = (\alpha_1, ..., \alpha_N)'$ and δ is a $(NT \times N)$ matrix defined as $\delta = I_N \otimes e$.

Our goal is to show that the estimated β when considering the dummy variables and an OLS regression is the same as the one obtained by the within transformation. To show that, we first use the Frisch-Waugh-Lovell theorem, that states that the OLS estimation of the model above is equivalent to the estimation of the following model:

$$M_\delta y = M_\delta \delta \alpha + M_\delta x \beta + M_\delta \varepsilon$$

where $M_\delta = I - \delta(\delta'\delta)^{-1}\delta'$ is the residual maker matrix of δ, that is, it projects into the orthogonal space spanned by δ.

Remembering that the within transformation is given by premultiplying the original model (stacked over time only) by the matrix $Q = I_T - \frac{1}{T}ee'$, we then aim to show that the projection matrix M_δ is equivalent to $I_N \otimes Q$ (we have it multiplied by I_N, as we are working with the model stacked over time and units, and we want to apply the within transformation for all individuals).

To start with, a little reminder of Kronecker-product rules:

$$\begin{aligned}
&(A \otimes B)' = A' \otimes B' \\
&(A \otimes B)(C \otimes D) = AC \otimes BD \\
&(A \otimes B)^{-1} = A^{-1} \otimes B^{-1} \\
&A \otimes (B + C) = A \otimes B + A \otimes C
\end{aligned}$$

Then, we also obtain the following expressions, by definition, and using that $\delta = I_N \otimes e$:

$$\begin{aligned}
M_\delta &= I_{NT} - P_\delta \\
P_\delta &= \delta(\delta'\delta)^{-1}\delta^{-1}\delta' \\
&= I_N \otimes e \left((I_N \otimes e)'(I_N \otimes e)\right)^{-1} (I_N \otimes e)'
\end{aligned}$$

We can further work out the expression of P_δ using (i) the Kronecker product rules, (ii) the fact that $I_N = I_N'$, and (iii) $e'e$ is simply equal to T:

$$\begin{aligned}
P_\delta &= (I_N \otimes e)\left((I_N \otimes e')(I_N \otimes e)\right)^{-1} (I_N \otimes e') \\
&= (I_N \otimes e)\left((I_N \otimes e'e)\right)^{-1} (I_N \otimes e') \\
&= \frac{1}{T}(I_N \otimes e)(I_N \otimes e') \\
&= \frac{1}{T}(I_N \otimes ee') \\
&= I_N \otimes \frac{1}{T}ee'
\end{aligned}$$

Then we can plug this in for $M_\delta = I_{NT} - P_\delta$, and using the fact that $I_{NT} = I_N \otimes I_T$:

$$\begin{aligned}
M_\delta &= I_{NT} - I_N \otimes \frac{1}{T}ee' \\
&= (I_N \otimes I_T) - (I_N \otimes \frac{1}{T}ee') \\
&= I_N \otimes \left(I_T - \frac{1}{T}ee'\right) \\
&= I_N \otimes Q
\end{aligned}$$

Which concludes the proof.

4 GLS First Differences Estimator

The aim of this section is to show that when estimating the FD model through a GLS procedure (since the first-differencing leads to correlation between the transformed error terms), the obtained estimator is the same as the one obtained with the within transformation.

Consider the model stacked over periods:

$$y_i = e\alpha_i + x_i'\beta + \varepsilon_i$$

Then, we take the first difference of this model by premultiplying it by a given matrix D_T that produces this first difference:

$$D_T y_i = D_T e\alpha_i + D_T x_i'\beta + D_T\varepsilon_i$$

Note that D_T is a $(T-1\times T)$ matrix, since one observation is lost due to the first differencing. Moreover, we have necessarily that the fixed-effects are differenced out, that is, $D_T e\alpha_i = 0$. Therefore:

$$D_T y_i = D_T x_i'\beta + D_T\varepsilon_i$$

We can further stack over individuals, and denote $D = I_N \otimes D_T$, which is equivalent to premultiply all individuals by the matrix D_T:

$$Dy = Dx\beta + D\varepsilon$$

Then, it is easy to obtain the variance of the new error terms for both the model stacked only over time periods, or over both dimensions, by assuming homoskedasticity:

$$\begin{aligned}\text{Var}(D_T\varepsilon_i) &= \mathbb{E}[D_T\varepsilon_i\varepsilon_i' D_T'] \\ &= D_T\sigma_\varepsilon^2 I_T D_T' \\ &= \sigma_\varepsilon^2 D_T D_T'\end{aligned}$$

$$\text{Var}(D\varepsilon) = \sigma_\varepsilon^2 DD'$$

Taking $\Omega = \sigma_\varepsilon^2 DD'$, and performing GLS with the transformed model, i.e., running an OLS of $y^* = Dy\Omega^{-1/2}$ on $x^* = Dx\Omega^{-1/2}$ we obtain:

$$\begin{aligned}\hat{\beta}_{GLS} &= ((Dx)'(DD')^{-1}Dx)^{-1}(Dx)'(DD')^{-1}Dy \\ &= (x'D'(DD')^{-1}Dx)^{-1}x'D'(DD')^{-1}Dy\end{aligned}$$

We can then look at what $D'(DD')^{-1}D$ looks like:

$$\begin{aligned} D'\left(DD'\right)^{-1}D &= \left(I_N \otimes D_T'\right)\left(\left(I_N \otimes D_T\right)\left(I_N \otimes D_T'\right)\right)^{-1}\left(I_N \otimes D_T\right) \\ &= \left(I_N \otimes D_T'\right)\left(I_N \otimes \left(D_T D_T'\right)^{-1}\right)\left(I_N \otimes D_T\right) \\ &= I_N \otimes \left(D_T'\left(D_T D_T'\right)^{-1} D_T\right) \\ &= I_N \otimes \left(I_T - \frac{1}{T}ee'\right) \end{aligned}$$

which yields then exactly the same result as the within transformation, since the final result is given by $D'\left(DD'\right)^{-1}D = I_N \otimes Q$, which is an idempotent matrix.