

ECMA 31100, IEA2 Homework

Homework 3

03/2022

JOHN RUF

1 The Torg Problem

Consider the heterogeneous IV setting from the Week 5/6 notes. Consider the IV estimand produced by estimating the linear regression model (1) using 2SLS and first stage (2):

$$\begin{aligned} Y &= \beta_0 + \beta_1 D + X' \beta_2 + U \\ D &= \pi_0 + \pi_1 Z + X' \pi_2 + V; \quad E([1, Z, X']' V) = 0 \end{aligned}$$

Suppose that $D \in \{0, 1\}$, $Z \in \{0, 1\}$, and X is a vector of regressors containing functions of the covariates W , e.g. $X = W$ or $X = (\mathbf{1}(W = w_2), \dots, \mathbf{1}(W = w_k))$, where k is the number of distinct values W can take, and $\mathbf{1}(W = w_1)$ is omitted to avoid perfect collinearity. The instrument Z is exogenous conditional on W and the monotonicity assumption conditional on W holds. In this question we will consider various specifications of X that alter the interpretation of β_1 .

1(a)

Write a formula for the IV estimand β_1^{IV} using the specification above. Decompose it into a weighted average of complier average treatment effects (conditional on W) and always-taker treatment effects (conditional on W), assuming $E(y_0 | W)$ is linear in W . Note that in general the term $E(E(\tilde{Z} | W)E(y_0 | W))$ will not disappear like it did in the notes, except, for example, when W is included linearly or in a saturated fashion (see parts c and d). Leave this term in your formula (separated from the two terms weighting complier and always-taker treatment effects) and add in an explanation for why it's there.

Solution. Following the notes we arrive The formula for the two stage least squares estimator is the following:

$$\begin{aligned} \beta_{IV} &= \frac{E(Y\tilde{Z})}{E(D\tilde{Z})} \\ &= \frac{E[\text{cov}(Y, \tilde{Z}|w) + E(Y|w)E(\tilde{Z}|w)]}{E(D\tilde{Z})} \end{aligned}$$

Where $\tilde{Z} = Z - BLP(Z|X)$ and now we just need to grind apart these terms, going one at a time we arrive at:

$$E(\tilde{Z}|w) = E(Z|w) - BLP(Z|w)$$

$$E(Y|w) = E(y_0|w) + E(D(y_1 - y_0)|w)$$

$$\begin{aligned} E[cov(Y, \tilde{Z}|w)] &= E[cov(Y, Z - BLP(Z|w)|w)] \\ &= E[cov(Y, Z|w)] \\ &= E[LATE(w)cov(D, Z)|w] \end{aligned}$$

So we obtain:

$$\beta_{IV} = \frac{E[LATE(w)cov(D, Z)|w] + E[E(Y|w)E(\tilde{Z}|w)]}{E(D\tilde{Z})}$$

Now decomposing the term:

$$\beta_{IV} = \frac{E[LATE(w)cov(D, Z)|w] + E[(E(y_0|w) + E(D(y_1 - y_0)|w))E(\tilde{Z}|w)]}{E(D\tilde{Z})}$$

Simplifying and seperating terms using linearity yields:

$$\beta_{IV} = \frac{E[LATE(w)cov(D, Z)|w]}{E(D\tilde{Z})} + \frac{E[E(D(y_1 - y_0)|w))E(\tilde{Z}|w)]}{E(D\tilde{Z})} + \frac{E[(E(y_0|w)E(\tilde{Z}|w)]}{E(D\tilde{Z})}$$

Noting that the first two terms are identical to the problem given in the notes, we can apply the same analysis and arrive at the following weighted average, plus the additional term:

$$\begin{aligned} \beta_{IV} &= \frac{E[LATE(w)[cov(D, Z|w) + E(Z|w)E(\tilde{Z}|w)P(cp(w))]}{E(D\tilde{Z})} + \frac{E[ATE(at, w)P(at|w) * E(\tilde{Z}|w)]}{E(D\tilde{Z})} \\ &+ \frac{E[(E(y_0|w)E(\tilde{Z}|w)]}{E(D\tilde{Z})} \end{aligned}$$

Note that LATE's are just conditional complier treatment effects, this seems obvious but it took me a minute to realize. However, since W is no longer included linearly in y_0 we cannot assume that it's orthogonal to W. This is because in reality we only can go off of X which is a function of W that may or may not be a linear transformation of W. If it is, then it is orthogonal to X and the additional term may be eliminated. ■

1(b)

1(b).1

Let $BLP(Z | X)$ denote the best linear predictor of Z given X (including a constant). Show that the numerator of weights on the complier ATEs are negative iff $BLP(Z | X) > 1$. Give an example of Z and X that produces $BLP(Z | X) > 1$ for some realization of X .

Solution. Therefore this question is entirely about the value of:

$$cov(D, Z|w) + E(Z|w)E(\tilde{Z}|w)P(cp(w))$$

Which so, lets factor out $P(cp(w))$ from which we get for the first term

$$cov(D, Z|w)/P(cp(w)) = \frac{var(Z|w) * p(cp(w))}{p(cp(w))}$$

Thus it comes out to:

$$\begin{aligned} cov(D, Z|w) + E(Z|w)E(\tilde{Z}|w)P(cp(w)) = \\ cov(D, Z|w)/P(cp(w)) * [E(Z|w) - E(Z|w)^2 + E(\tilde{Z}|w)] \end{aligned}$$

Factoring the bracketed term and noting the definition of \tilde{Z} yields the following.

$$\begin{aligned} E(Z|w)[1 - E(Z|w) + E(\tilde{Z}|w)] \\ = E(Z|w)[1 - E(Z|w) + E(Z|w) - BLP(Z|X)] \\ = E(Z|w)[1 - BLP(Z|X)] \end{aligned}$$

However, since probabilities can only be positive, and the expectation of Z given w can only be positive (as Z is binary) thus this term is negative iff $BLP(Z|X) > 1$.

For the example, we can just use the example given in 1D, which obviously fulfills the requirements since

$$1 - BLP(Z|X = 2) = 1 - (1 + 2\epsilon) = -\epsilon$$

■

1(b).2

Show that the weights on the always-taker ATEs must take positive and negative values unless $E(Z | W) = BLP(Z | X)$.

Solution. This comes out of the fact that a best linear predictor must "over" and "under" predict if it's not perfect. The weights on the always taker ATE's are directly equivalent to:

$$P(at|w) * E(\tilde{Z}|w) = P(at|w) * (E(Z|W) - BLP(Z|X))$$

So this rearrangement alone proves the "unless" portion works, and so we now need to show that if $BLP(Z|X) \neq E(Z|W)$ that there cannot be only positive or negative values.

So for contradiction, let us suppose that $BLP(Z|X) \neq E(Z|W)$ and only takes positive values. Thus by supposition there exists a w_j such that

$$E(\tilde{Z}|w = w_j) > 0$$

However, looking at $E(\tilde{Z})$ we obtain from supposition:

$$\begin{aligned} E(\tilde{Z}) &= \sum_{i=1}^K E(\tilde{Z}|w = w_i)P(w = w_i) \\ &> E(\tilde{Z}|w = w_j) > 0 \end{aligned}$$

However, by construction $E(\tilde{Z}) = 0$ as we subtracted the best linear predictor of Z given X , so this cannot be the case. A similar proof works for the case where $E(\tilde{Z}|w = w_j) < 0$. ■

1(c)

Suppose $X = (\mathbf{1}(W = w_2), \dots, \mathbf{1}(W = w_k))$. Use part a) to argue that β^{IV} is a positive weighted average of complier average treatment effects (the average is over the distribution of W). Do this in 2 steps: - Argue that the numerator of the weights equals $\text{Cov}(D, Z | X)$, which is non-negative. What must be true about the proportion of compliers and proportion of those receiving $Z = 1$ and $Z = 0$ at each X for this quantity to be strictly positive at each X ? - Argue that the denominator equals $E(\text{Cov}(D, Z | X))$, which is non-negative. Under what condition are the weights well-defined?

Solution. From the supposition given we know that:

$$\begin{aligned} E(\tilde{Z}|w) &= E(Z|w) - BLP(Z|X) \\ &= E(Z|w) - E(Z|w) = 0 \end{aligned}$$

Where the second line follows because Z is binary, so the BLP is by construction the conditional expectation. This implies that the second two terms in a are zero and thus:

$$\beta_{IV} = \frac{E[LATE(w)[\text{cov}(D, Z|w) + 0]}{E(D\tilde{Z})} + 0 + 0$$

Which is the first part of the argument. We also know that:

$$E(D\tilde{Z}) = E(p(cp|w)\text{var}(Z|w)) = E(\text{cov}(D, z|w))$$

Which is the second part of the argument. Thus the weights are well defined if the relevance condition holds: IE: the covariance is not 0 for some value of w . ■

1(d)

Now suppose $X = W$. Show by example that the denominator of the weights can be strictly negative even though monotonicity conditional on W holds. Hint: Write $E(D\tilde{Z}) = E(E(D\tilde{Z} | W))$, where $\tilde{Z} = Z - \text{BLP}(Z | W)$ and decompose $E(D\tilde{Z} | W)$ as you did for $E(Y\tilde{Z} | W)$ in part a). Now let $W \in \{0, 1, 2\}$, where each value occurs with probability $1/3$, and, for some small $\epsilon > 0$, let

$$E(Z | W) = \begin{cases} \epsilon & W = 0 \\ 1 - \epsilon & W \in \{1, 2\} \end{cases}$$

which ensures $1 - \text{BLP}(Z | W) < 0$ for $W = 2$. Now let the proportion of never takers vary with W .

Solution. Exploiting the hint,

$$\begin{aligned} E(D\tilde{Z}) &= E(D\tilde{Z}|w) \\ &= E(\text{cov}(D, \tilde{Z}|w) + E(D|w)E(\tilde{Z}|w)) \\ &= E(P(cp|w)\text{var}(z|w) + E(D|w)E(\tilde{Z}|w)) \end{aligned}$$

Going through these terms one by one, starting with $E(D|w)$ and plugging in $D = D_0 + Z(D_1 - D_0)$:

$$\begin{aligned} E(D|w) &= p(at|w) + E(Z(D_1 - D_0)|w) \\ &= p(at|w) + E(Z|cp, w)p(cp|w) \end{aligned}$$

However, turning our attention to $E(\tilde{Z}|w)$, applying our normal transformation of this term gets us:

$$\begin{aligned} E(D\tilde{Z}) &= E(P(cp|w)\text{var}(z|w) + E(D|w)E(\tilde{Z}|w)) \\ &= p(cp|w)E(Z|w)[1 - E(Z|w) + E(\tilde{Z}|w)] + p(at|w)E(\tilde{Z}|w) \\ &= p(cp|w)E(Z|w)[1 - E(Z|w) + (E(Z|w) - E(\text{BLP}(Z|w)|w))] + p(at|w)E(\tilde{Z}|w) \\ &= p(cp|w)E(Z|w)[1 - \text{BLP}(Z|w)] + p(at|w)[E(Z|w) - \text{BLP}(Z|w)] \end{aligned}$$

However, we have a restriction based off simple probability theory that $p(at|w) + p(cp|w) + p(nt|w) = 1$. Ergo, we can set the probability of being a never taker condition on w , $p(nt|w)$, such that $p(at|w) = 1 - p(cp|w) - p(nt|w)$. However, monotonicity requires only that $P(D_1 = 1|W = w) \geq P(D_0 = 1|W = w)$, thus a set of probabilities that guarantees that this is negative is: $p(cp|w = 2) = 1$, $p(nt|w = 2) = p(at|w = 2) = 0$. Which is clearly consistent with the monotonicity condition given in the notes. ■

1(e)

Now suppose there are no covariates (X is empty) but instead of a binary instrument we have a multi-valued instrument Z which satisfies the monotonicity condition. Suppose the first stage regression is estimated using dummy variables for each possible value of Z (without the first value, so we can keep the constant). Show that the IV estimand is a weighted average of LATEs, and interpret the weights. Hint: Imbens Angrist 1994.

Solution. Taking note of the analysis given by Imbens and Angrist 1994, let $g(z) = p(Z = z)$, which will use as an instrument in this case. Since we have n values of the instrument we can define one LATE per pair of points in Z . Furthermore, we can relate the following local average treatment effects for elements k, m, l such that $k \neq l \neq m$:

$$\alpha_{z_m, z_k} = \frac{P(z_l) - P(z_k)}{P(z_m) - P(z_k)} \cdot \alpha_{z_l, z_k} + \frac{P(z_m) - P(z_l)}{P(z_m) - P(z_k)} \cdot \alpha_{z_m, z_l}$$

We know that the IV procedure using this function estimates:

$$\alpha_g^{IV} = \frac{\text{Cov}(Y, g(Z))}{\text{Cov}(D, g(Z))} = \frac{E[Y \cdot (g(Z) - E[g(Z)])]}{E[D \cdot (g(Z) - E[g(Z)])]}$$

Where the numerator of this expression can be analyzed as:

$$\begin{aligned} E[Y \cdot (g(Z) - E[g(Z)])] &= \sum_{l=0}^K \pi_l \cdot E[Y \mid Z = z_l] \cdot (g(z_l) - E[g(Z)]) \\ &= \sum_{l=0}^K \pi_l \cdot E[Y \mid Z = z_0] \cdot (g(z_l) - E[g(Z)]) \\ &\quad + \sum_{l=1}^K \pi_l \sum_{k=1}^l \alpha_{z_k, z_{k-1}} \cdot (P(z_k) - P(z_{k-1})) \cdot (g(z_l) - E[g(Z)]) \\ &= \sum_{k=1}^K \alpha_{z_k, z_{k-1}} \cdot (P(z_k) - P(z_{k-1})) \sum_{l=k}^K \pi_l \cdot (g(z_l) - E[g(Z)]) \end{aligned}$$

And by a similar computation the denominator comes out to:

$$\begin{aligned} E[D \cdot (g(Z) - E[g(Z)])] &= \sum_{m=1}^K (P(z_m) - P(z_{m-1})) \cdot \sum_{l=m}^K \pi_l \cdot (g(z_l) - E[g(Z)]) \end{aligned}$$

Thus the weights sum to 1 and everything works out hunky dory. The weights here are combined between the differences in the values of the instrument (such that a major change in the instrument value gives a greater weight), and that a greater change in the probability between this instrument value and the next is more highly weighted. ■

1(f)

Now suppose there are no covariates (X is empty) but instead of a binary treatment we have a multi-valued treatment $D \in \{0, 1, \dots, K\}$ and binary instrument. Note that there are now $K + 1$ potential outcomes y_0, \dots, y_K . Show that

$$\beta_{IV} = \sum_{k=1}^K \left[\frac{P(D_1 \geq k > D_0)}{\sum_{m=1}^K P(D_1 \geq m > D_0)} \right] E(Y_k - Y_{k-1} \mid D_1 \geq k > D_0).$$

How do you interpret the weights, expectations and estimand as a whole?

Solution. Starting off with our traditional starting point we have:

$$\begin{aligned} \beta_{IV}^1 &= \frac{\text{cov}(Y, Z)}{\text{cov}(D, Z)} \\ &= \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D|Z=1) - E(D|Z=0)} \\ &= \frac{E(Y|Z=1) - E(Y|Z=0)}{E(D_1) - E(D_0)} \end{aligned}$$

Where we can rewrite:

$$E(D_0) = \sum_i^J P(D_0 = j) = \sum_i^J P(D_0 \geq j)$$

However monotonicity guarantees that $\{D_0 \geq j\} \subset \{D_1 \geq j\}$. Thus we can say that:

$$\begin{aligned} E(D_1) - E(D_0) &= \sum_i^J [P(D_1 \geq j) - P(D_0 \geq j)] \\ &= \sum_i^J P(D_1 \geq j > D_0) \end{aligned}$$

We also know that because Z is a valid instrument that it must be orthogonal to y except through D . Thus we can decompose y as the following:

$$y = y_0 + \sum_{j=1}^J [(y_j - y_{j-1}) \mathbf{1}(D \geq j)]$$

Where we can take expectations and obtain:

$$E(y|z=1) = E(y_0) + \sum_{j=1}^J [E(y_j - y_{j-1})] + \sum_{j=1}^J (E[(y_j - y_{j-1}) \mathbf{1}(D \geq j)])$$

Where $E(y|z=0)$ has a similar structure. So putting this together we can obtain:

$$E(Y|z = 1) - E(y|z = 0) = \sum_{j=1}^J E(y_j - y_{j-1}) \mathbf{1}(D_1 \geq j \geq D_0)$$

So putting this all together we have:

$$\begin{aligned} \beta_{IV}^1 &= \frac{\sum_{j=1}^J E(y_j - y_{j-1}) \mathbf{1}(D_1 \geq j \geq D_0)}{\sum_{m=1}^K P(D_1 \geq m > D_0)} \\ &= \frac{\sum_{j=1}^J [(p(D_1 \geq j \geq D_0)) E(y_j - y_{j-1} | D_1 \geq j \geq D_0)]}{\sum_{m=1}^K P(D_1 \geq m > D_0)} \end{aligned}$$

Which is equivalent to the expression given in the problem statement by some basic algebra. ■

2 Abadie: The Legend of Kappa

Read Abadie (2003) ([here](#)). The data used in Section 6 is on canvas. Reproduce columns 1-3 of Table 2. Explain how you would use Abadie's Kappa with a linear local average response function (defined in section 4.2.1 of that paper) to reproduce column 4. Implement this using logistic regression to estimate $P(Z = 1 \mid X)$. Are your parameter estimates much different? Note: You don't need to write your own code for the logistic regression or produce standard errors for the estimates constructed using Abadie's Kappa.

Solution. Using the code attached I have reproduced the following columns:

Table 1: First 3 Results

	<i>Dependent variable:</i>		
	nettfa (1)	p401k (2)	nettfa (3)
p401k	13,527.050*** (1,394.046)		
p401khat			9,418.828*** (1,864.539)
inc	976.931*** (28.255)	0.001*** (0.0001)	997.190*** (29.003)
age2	−376.165* (217.622)	−0.002** (0.001)	−345.955 (218.612)
agesq2	38.699*** (5.774)	0.0001* (0.00003)	37.852*** (5.801)
marr	−8,369.471*** (1,639.216)	−0.0005 (0.008)	−8,355.871*** (1,645.263)
fsize	−785.650 (497.072)	0.0001 (0.002)	−818.962 (499.005)
e401k		0.688*** (0.006)	
Constant	−23,549.000*** (1,993.603)	−0.031*** (0.010)	−23,298.740*** (2,002.360)
Observations	9,275	9,275	9,275
R ²	0.187	0.596	0.181
Adjusted R ²	0.186	0.596	0.180
Residual Std. Error (df = 9268)	57,705.920	0.284	57,918.620
F Statistic (df = 6; 9268)	354.417***	2,281.604***	340.494***

Note:

*p<0.1; **p<0.05; ***p<0.01

I would use Abadie's Kappa as a weighting to the standard OLS regression, so I would solve the following problem:

$$\beta = (X' \kappa X)^{-1} X' \kappa y$$

Which is the standard least squares problem with κ as weights in the approach. However, to estimate the probabilistic portions of κ we use the route given in the problem definition and take a logistic regression of the X variables on the instrument and use the fitted values as the probability $p(z = 1|x)$ and then using 1 minus that for the probability z is zero. Using this I get the following coefficients:

Table 2: Kappa weighted Least Squares Coefficients

p401k	9556.437
Const	-25396.87
inc	1011.52
age	56.47
agesq	30.83397
marr	-7477.74
fszise	-1056.82



3 MOAR



Figure 1: Me to my Differences estimator when I have an extra control group and suspect spillover effects

3(a)

a) What does a triple difference in difference argument allow for that an ordinary difference in difference argument doesn't? Give an example where you might use a triple diff in diff rather than a double difference, and an example where a triple difference would actually lead to a biased ATT estimate but a double difference wouldn't.

Solution. One example where triple differences would be superior to single differences is if there are significant spillovers in one are but not the other: It would be great if we could just compare groups A and B within the treatment state and be done with it, but there may be spillovers, formally speaking this means

$$Y_T^A = c + \delta Y_T^B$$

Where the partial derivative of f wrt to the observed outcome of group B Y^B is positive. Thus let D reflect the treatment status of the group, and let the true treatment effect on the treated be 1, such that $ATT = D$. Thus taking the naive differences-in-differences estimator yields:

$$\begin{aligned} DiD &= [E(Y|D = 1, T = 1) - E(Y|G = 1, T = 0)] - \\ &\quad [E(Y|D = 0, T = 1) - E(Y|D = 0, T = 0)] \\ &= 1 - [c - (c + \delta)] \\ &= 1 - \delta \neq 1 \end{aligned}$$

Which is not the true average treatment effect on the treated. However, let us suppose that groups A and B have parallel pretrends in the treatment state, but not in the control state. Therefore we obtain:

$$Y_C^A = c + \delta Y_C^B$$

$$Y_T^A = \delta Y_T^B = c$$

Plugging this into the DiDiD estimator we get:

$$DiDiD = ((1 - 0) - [c - c]) - ((0 - 0) - [c - (c + \delta)])$$

$$= 1 + \delta \neq 1$$

■

3(b)

b) Suppose you use a double difference argument and have one treatment and one control group ($g = 0, 1$), one post-treatment date t_0 and several pre-treatment dates $t = 0, \dots, t_0 - 1$. You don't believe the common trends assumption, but still want to identify the ATT for the treated group in the post-period. You make a parametric trends assumption:

$$E(y(0) | G, T) = \alpha + \sum_{t=1}^{t_0} \beta_t \mathbf{1}(T = t) + G(\gamma + \delta_1 T + \delta_2 T^2)$$

Show that this assumption relaxes common trends, but then specify the regression which identifies ATT_{t_0} . Is there an issue with collinearity here? (Think about how large t_0 must be to avoid it).

Solution. The common trends assumption is given by:

$$E(y(0)|g = 1, T = 1) - E(y(0)|g = 1, T = 0) = E(y(0)|G = 0, T = 1) - E(y(0)|G = 0, T = 0)$$

Which indicates that the differences between the untreated outcomes of between the pre and post periods are equal for both the treated and untreated group.

However, the parametric trends assumptions says that the treated and untreated groups may not be equal, but that those differences parameterizations are constant in each time period. Thus the differences between the untreated outcome trends has a constant function over time. Thus the differences between them would be:

$$E(\Delta y(0)|T) = (\gamma + \delta_1 T + \delta_2 T^2)$$

Which is only a function of constants and T. Thus the parametrization of this in a regression would simply be the following:

$$Y = \beta_0 + \sum_{t=1}^{t_0} \beta_t + \gamma_1 G + \gamma_2 G * T + \gamma_3 G * T^2 + \beta_4 * D * t_0 + U$$

Where D is a dummy variable indicating that $G = 1$ and $T = t_0$. We are also assuming that $E(u|g, T) = 0$ with this assumption.

However, this indicates that we have two more parameters to estimate. What we're effectively doing is assuming that the difference in the treated and not treated groups is a function of only time and estimating that difference and removing it with the terms β_2 and β_4 in the regression and computing our DiD estimator taking that difference into account.

However, we now have the number of pre-periods+5 parameters with $2t_0 + 1$ conditional means we need to estimate, giving us the following inequality:

$$\begin{aligned} 2t_0 + 1 &\geq t_0 + 5 \\ \implies t_0 &\geq 4 \end{aligned}$$

So we need at least 3 pre-periods to avoid collinearity issues. ■

4 Panel Data Equivalence

Question 4 Suppose we have a panel $\{(y_{it}, x_{it}) : t = 1, \dots, T\}_{i=1}^N$. Show that the OLS estimates of β are numerically equivalent in the following specifications:

$$y_{it} = x'_{it}\beta + \gamma_i + u_{it}$$

$$y_{it} - \bar{y}_i = (x_{it} - \bar{x}_i)' \beta + \epsilon_{it},$$

where γ_i is an individual fixed effect (adds a dummy for each i into the specification) and $\bar{y}_i = \frac{1}{T} \sum_{t=1}^T y_{it}$ and $\bar{x}_i = \frac{1}{T} \sum_{t=1}^T x_{it}$ does not contain a constant to avoid collinearity.

Solution. Note that by stacking, the differenced regression is equivalent to:

$$Y = \begin{bmatrix} \frac{\sum_{t=1}^T (y_{1,t})}{T} & 1 \\ \frac{\sum_{t=1}^T (y_{1,t})}{T} & 2 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (y_{1,t})}{T} & N \\ \frac{\sum_{t=1}^T (y_{2,t})}{T} & 1 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (y_{N,t})}{T} & N \end{bmatrix}^{NT \times 1} = X' \beta - \begin{bmatrix} \frac{\sum_{t=1}^T (x_{1,t})}{T} & 1 \\ \frac{\sum_{t=1}^T (x_{1,t})}{T} & 2 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (x_{1,t})}{T} & N \\ \frac{\sum_{t=1}^T (x_{2,t})}{T} & 1 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (x_{N,t})}{T} & N \end{bmatrix}^{NT \times K} \beta + \epsilon_{it}$$

So let's define a dummy matrix $D \in \mathbb{R}^{NT \times N}$ where each entry is 1 for individual i , regardless of time, and zero elsewhere. So from this we know that the dummy model is in actuality:

$$Y = X' \beta + D' \Gamma + U$$

Where the capital vectors and matrices are produced by stacking (including turning γ to Γ). Now apply FWL with respect to the dummy and we obtain:

$$M_D Y = M_D X' \beta + M_D D' \Gamma + M_D U$$

Where M is regular residual mater matrix, thus evaluating we get

$$Y - D(D'D)^{-1}D'Y = X' \beta - D(D'D)^{-1}D'X' \beta$$

Now, all that's left to do is to argue that $D(D'D)^{-1}D'Y = \bar{Y}_i$, where $D(D'D)^{-1}D'X' \beta = \beta \bar{X}$ will follow by symmetry.

WLOG of generality, let the observations be ordered as the following:

$$Y = \begin{bmatrix} y_{1,1} \\ y_{1,2} \\ \dots \\ y_{1,T} \\ y_{2,1} \\ \dots \\ y_{N,T} \end{bmatrix}^{NT \times 1}$$

Then, the dummy variable matrix would look like:

$$D' = \begin{bmatrix} 1_{1,1} & 0_{1,1} & \dots & 0_{1,1} \\ 1_{1,2} & 0_{1,2} & \dots & 0_{1,2} \\ \dots & \dots & \dots & \dots \\ 1_{1,T} & 0_{1,T} & \dots & 0_{1,T} \\ 0_{2,T} & 1_{2,T} & \dots & 0_{2,T} \\ \dots & \dots & \dots & \dots \\ 0_{N,T} & 0_{N,T} & \dots & 1_{N,T} \end{bmatrix}^{N \times NT}$$

However, by standard matrix multiplication we can see

$$(D'D) = \begin{bmatrix} T & 0 & \dots & 0 \\ 0 & T & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & T \end{bmatrix}^{N \times N}$$

And:

$$(D'Y) = \begin{bmatrix} \sum_{t=1}^T (y_{1,t}) \\ \sum_{t=1}^T (y_{2,t}) \\ \dots \\ \sum_{t=1}^T (y_{N,t}) \end{bmatrix}^{N \times 1}$$

Thus we can see:

$$(D'D)^{-1}(D'Y) = \begin{bmatrix} \frac{\sum_{t=1}^T (y_{1,t})}{T} \\ \frac{\sum_{t=1}^T (y_{2,t})}{T} \\ \dots \\ \frac{\sum_{t=1}^T (y_{N,t})}{T} \end{bmatrix}^{N \times 1}$$

Finally yielding:

$$D(D'D)^{-1}(D'Y) = \begin{bmatrix} \frac{\sum_{t=1}^T (y_{1,t})}{T} & 1 \\ \frac{\sum_{t=1}^T (y_{1,t})}{T} & 2 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (y_{1,t})}{T} & N \\ \frac{\sum_{t=1}^T (y_{2,t})}{T} & 1 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (y_{N,t})}{T} & N \end{bmatrix}^{NT \times 1}$$

Where the the outer subscript is just denoting where in the vector one is at, not denoting any different values. Now note that:

$$Y - D(D'D)^{-1}(D'Y) = \begin{bmatrix} y_{1,1} \\ y_{1,2} \\ \dots \\ y_{1,T} \\ y_{2,1} \\ \dots \\ y_{N,T} \end{bmatrix}^{NT \times 1} - \begin{bmatrix} \frac{\sum_{t=1}^T (y_{1,t})}{T} & 1 \\ \frac{\sum_{t=1}^T (y_{1,t})}{T} & 2 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (y_{1,t})}{T} & N \\ \frac{\sum_{t=1}^T (y_{2,t})}{T} & 1 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (y_{N,t})}{T} & N \end{bmatrix}^{NT \times 1}$$

Which is equivalent to the LHS of the differenced regression. By symmetrical logic, we can note that

$$M_D X' \beta = \begin{bmatrix} \frac{\sum_{t=1}^T (x_{1,t})}{T} & 1 \\ \frac{\sum_{t=1}^T (x_{1,t})}{T} & 2 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (x_{1,t})}{T} & N \\ \frac{\sum_{t=1}^T (x_{2,t})}{T} & 1 \\ \dots & \dots \\ \frac{\sum_{t=1}^T (x_{N,t})}{T} & N \end{bmatrix}^{NT \times K} \beta$$

Ergo, the right hand side is also equivalent. Thus since all we did was right multiply both sides of the equation by a constant matrix, the two regression equations are equivalent under the same set of moment conditions. ■