Problem set 2: Panel Data Models

1. *You have a sample of N individuals for T years. Suppose you estimate by OLS the annual income equation:*

$$y_{it} \quad = \quad \alpha_0 + \alpha_1 ed_i + \alpha_2 age_{it} + \alpha_3 (ed_i \times age_{it}) + \gamma y_{it-1} + u_{it}$$

where ed_i represents the years of education of the ith individual, age_{it} represents the age of the individual i in period t and u_{it} represents all unobservables.

(a) *Suppose you estimate γ as 0.82 with the standard error of 0.12. State a set of sufficient assumptions for the consistency of the OLS estimator in this context.*

A dynamic relationship is illustrated in this example by the presence of a lagged dependent variable among the regressors.

Assumptions for OLS to be consistent:

- Regularity condition (no asymptotic multi-collinearity)

$$\operatorname*{plim}_{N \to \infty} \left(\frac{X'X}{N} \right) = M$$

where M is positive definite $\forall_T$.

- $E(u_{it} x_{it}) = 0$: no contemporaneous correlation between the regressors and the unobservable. Since X includes the lag dependent variable, this implies also no feedback of the error into past values of Y. It then implies that u is not serially correlated.

(b) *Describe an alternative estimation technique that you could use to evaluate the validity of some of your assumptions. Justify your choice and explain carefully the conditions under which your alternative estimator is consistent.*

Assume $u_{it} = f_i + v_{it}$ where f is a fixed effect. The model is,

$$y_{it} = \alpha_0 + \alpha_1 \text{ed}_i + \alpha_2 \text{age}_{it} + \alpha_3 \left(\text{ed}_i * \text{age}_{it} \right) + \gamma y_{it-1} + f_i + v_{it}$$

Since the model includes the lagged dependent variable as an explanatory variable, OLS is biased and inconsistent.

To estimate the model consistently we use the first-difference model,

$$\Delta y_{it} = \alpha_2 + \alpha_3 \Delta \mathrm{ed}_i + \gamma \Delta y_{it-1} + \Delta v_{it}$$

where $\mathrm{age}_{it} - \mathrm{age}_{it-1} = 1$, $\Delta y_{it-1} = y_{it-1} - y_{it-2}$ and $\Delta v_{it} = v_{it} - v_{it-1}$.

We got rid of the fixed effect and the time invariant regressors. However, there is now correlation between Δy_{it-1} and Δv_{it} since y_{it-1} is correlated with v_{it-1}.

Thus, we need to use IV (instrumental variables).

- The instrument for Δy_{it-1} could be Δy_{it-2} or y_{it-2}.
- These instruments are not correlated with Δv_{it} (as long as the v_{it} is not serially correlated) but should be correlated with Δy_{it-1}.
- In order to use the instrument Δy_{it-2}, T must be larger than 3: the minimum number of periods is 4.
- If we use as instrument y_{it-2}, the minimum number of periods is 3.

We need to check the rank and order conditions:

- $E\left(\Delta y_{it-2} \Delta v_{it}\right) = 0$: the order condition is satisfied if this condition holds.
- $E\left(z_{it} \Delta x'_{it}\right)$ has a rank equal to the number of regressors. In this case,

$$\begin{aligned} \Delta x &= \begin{bmatrix} 1 & \Delta \mathrm{ed} & \Delta y_{(-1)} \end{bmatrix} \\ z &= \begin{bmatrix} 1 & \Delta \mathrm{ed} & y_{(-2)} \end{bmatrix} \end{aligned}$$

  We can check the rank condition by estimating $B = Cov(\Delta y_{it-1} y_{it-2})$. When $B = 0$ the instrument is useless.

To check the validity of OLS we can apply the Hausman test:

- H_0: no fixed effect
- H_1: fixed effect

Under H_0, both OLS and IV are consistent. OLS is efficient if there are no random effects.

Under H_1, OLS is inconsistent and IV is consistent.

The test asks if the estimates ($\widehat{\gamma}_{IV}$ and $\widehat{\gamma}_{OLS}$) are significantly different.

The test statistic is

$$m = \left(\widehat{\gamma}_{IV} - \widehat{\gamma}_{OLS}\right)' \left[\mathrm{var}\left(\widehat{\gamma}_{IV}\right) - \mathrm{var}\left(\widehat{\gamma}_{OLS}\right)\right]^{-1} \left(\widehat{\gamma}_{IV} - \widehat{\gamma}_{OLS}\right) \overset{a}{\sim} \chi^2_K$$

2. *Consider the model with a single regressor x_{it}*

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

where α_i represents an unobserved effect fixed over time and u_{it} is a homoskedastic error term which is independent over time (t) and individuals (i). There are N randomly sampled individuals, each observed for $T = 4$ time periods. Assume that $E(u_{it} \mid X) = 0$ for all i and that $E(u_{it}u_{is} \mid X,\ \text{any } t \text{ and } s : t \neq s) = 0$, where X represents the $NT \times 1$ data matrix.

(a) *Derive the covariance matrix for the Within Groups estimator and for the random effects estimator.*

`Within Groups estimator`

The assumptions are:

- $E(u_{it} \mid X) = 0, \forall i$ (strict exogeneity)
- $E(u_{it}u_{is} \mid X, t, s = 1, \ldots, T, t \neq s) = 0$ (no serial correlation)

The within groups estimator is:

$$\begin{aligned} \beta_1^{WG} &= (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y} \\ &= (X'QX)^{-1}X'QY \end{aligned}$$

where $Q = I - P$ is idempotent.

The variance of the WG estimator is,

$$\text{var}(\beta_1^{WG}) = E\left[(\beta_1^{WG} - \beta_1)(\beta_1^{WG} - \beta_1)'\right]$$

where

$$\begin{aligned} \beta_1^{WG} &= (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'\widetilde{Y} \\ &= (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'(\widetilde{X}\beta_1 + u) \\ &= \beta_1 + (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'u \end{aligned}$$

and so,

$$\begin{aligned}
\text{avar}(\beta_1^{WG}) &= E\left[(\beta_1 + (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'u - \beta_1)(\beta_1 + (\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'u - \beta_1)'\right] \\
&= E\left[(\widetilde{X}'\widetilde{X})^{-1}\widetilde{X}'uu'\widetilde{X}(\widetilde{X}'\widetilde{X})^{-1}\right] \\
&= M^{-1}_{\widetilde{X}'\widetilde{X}} E\left[\widetilde{X}'E\left(uu'|X\right)\widetilde{X}\right] M^{-1}_{\widetilde{X}'\widetilde{X}} \\
&= M^{-1}_{\widetilde{X}'\widetilde{X}} E\left[\widetilde{X}'\left(\sigma_u^2 I\right)\widetilde{X}\right] M^{-1}_{\widetilde{X}'\widetilde{X}} \\
&= \sigma_u^2 M^{-1}_{\widetilde{X}'\widetilde{X}} \\
&= \sigma_u^2 M^{-1}_{X'QX}
\end{aligned}$$

`Random effects estimator`

The random effects estimator is:

$$\begin{aligned}
\beta^{GLS} &= \left(\sum_{i=1}^{N} X_i'V_i^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'V_i^{-1}Y_i\right) \\
&= \left(X'V^{-1}X\right)^{-1}\left(X'V^{-1}Y\right)
\end{aligned}$$

where $V_i^{-1} = \frac{1}{\sigma_u^2}[Q + \Psi P]$, $\Psi = \frac{\sigma_u^2}{T\sigma_f^2 + \sigma_u^2}$ and $V = \text{diag}\left(V_1, ..., V_N\right)$.

Note that to obtain the GLS estimator we need V^{-1} and the assumption of random effects.

The variance of the GLS estimator is,

$$\text{var}\left(\beta^{GLS}\right) = E\left[(\beta^{GLS} - \beta)(\beta^{GLS} - \beta)'\right]$$

where

$$\begin{aligned}
\beta^{GLS} &= \beta + \left(\sum_{i=1}^{N} X_i'V_i^{-1}X_i\right)^{-1}\left(\sum_{i=1}^{N} X_i'V_i^{-1}v_i\right) \\
&= \beta + \left(X'V^{-1}X\right)^{-1} X'V^{-1}v
\end{aligned}$$

where $v_{it} = u_{it} + \alpha_i$

The variance of GLS can now be computed,

$$\begin{aligned}
\text{avar}(\beta^{GLS}) &= E\left[\left(\beta_1 + \left(X'V^{-1}X\right)^{-1} X'V^{-1}v - \beta_1\right)\left(\beta_1 + \left(X'V^{-1}X\right)^{-1} X'V^{-1}v - \beta_1\right)'\right] \\
&= E\left[\left(X'V^{-1}X\right)^{-1} X'V^{-1}vv'V^{-1}X\left(X'V^{-1}X\right)^{-1}\right] \\
&= M^{-1}_{XVX} E\left[X'V^{-1}E\left(vv|X\right)'V^{-1}X\right] M^{-1}_{X'VX} \\
&= M^{-1}_{XVX}
\end{aligned}$$

where $E(v_i v_i' | X) = V_i$.

(b) *Explain how you could test the assumption that* $E(\alpha_i \mid x_{it}) = 0$

Apply the Hausman test to check the random effect assumption $E(\alpha_i \mid x_{it}) = 0$.

H0: $E(\alpha_i \mid x_{it}) = 0$ (GLS is most efficient)

H1: $E(\alpha_i \mid x_{it}) \neq 0$ (GLS is inconsistent and biased)

If H0 is rejected, there is evidence of unobserved individual effects that are correlated with the regressors.

3. *You wish to study the effects of unionisation on wages using a panel of N individuals and T time periods. You wish to allow for the following phenomena: a) unionised firms select the higher ability workers and b) workers with bad productivity shocks join the union sector.*

(a) *Set up a suitable model and explain how these phenomena are reflected in your specification.*

The model is

$$w_{it} = \beta_1 \text{union}_{it} + x_{it}\gamma + f_i + v_{it}$$

where x includes other possible explanatory variables like education and age.
This specification allows for

- unionised firms selecting the higher ability workers:

$$E(f_i | \text{union}_{it}) > 0$$

- workers with bad productivity shocks being more likely to join the union sector:

$$E(v_{it} \mid \text{union}_{it+j}) < 0 \text{ for } j \geqslant 1$$

(b) *Explain how you would estimate this model and present the estimator. Carefully state any assumptions you make.*

Under these conditions, this is a fixed effect model with weak exogeneity.

We use the first differences model in the estimation:

$$\Delta w_{it} = \beta_1 \Delta \text{union}_{it} + \Delta x_{it}\gamma + \Delta v_{it}$$

Although the fixed effect has been eliminated, we have created another problem with the first differences: Δunion is endogenous,

$$E\left(\Delta v_{it} | \Delta \text{union}_{it}\right) \neq 0$$

because

$$E\left(v_{it-1} | \text{union}_{it}\right) \neq 0$$

Therefore, we need to use IV in the estimation.

Suitable choice of instrument: under the stated assumptions,

$$E\left(\Delta v_{it} \Delta \text{union}_{it-1}\right) = 0$$

thus $\Delta \text{union}_{it-1}$ is a suitable instrument.

The IV estimator will be,

$$\begin{bmatrix} \beta \\ \gamma \end{bmatrix} = (Z'\widetilde{X})^{-1} Z' \Delta W$$

where $\widetilde{X}_{it} = [\Delta \text{union}_{it} \quad \Delta X_{it}]$ and $Z_{it} = [\Delta \text{union}_{it-1} \quad \Delta X_{it}]$.

Note of caution:

IV can be used if the rank condition holds: $E\left(\Delta \text{union}_{it-1} \Delta \text{union}_{it}\right) \neq 0$. But this assumption is unlikely to hold given that union_{it} is a dummy variable. Choose instead union_{it-1} for instrument.

4. *Suppose you decide to estimate the single β parameter in*

$$y_{it} = x_{it}\beta + f_i + u_{it}$$

by OLS on the first differences model when x_{it} is strictly exogenous and there are $T > 2$ time periods of data available for N individuals. Assume f_i is unobserved and var$(\Delta x_{it}) > 0$ where $\Delta x_{it} = x_{it} - x_{it-1}$.

(a) *Show that this estimator is consistent.*

The first differences model is,

$$\Delta y_{it} = \Delta x_{it}\beta + \Delta u_{it}$$

The OLS estimator of the first differences model is unbiased and consistent if,

$$E(\Delta x_{it}\Delta u_{it}) \quad = \quad 0 \tag{1}$$

Condition (1) follows from x being strictly exogenous, which means that

$$E(\Delta x_{it}|\Delta u_{ij},\, j=1,...T) \quad = \quad 0$$

To establish consistency, we also need to assume that $E(\Delta x_{it}'\Delta x_{it})$ is full rank. In general, if this does not hold, we will remove the explanatory variables that lead to multi-collinearity. Under these conditions, we can write,

$$\Delta x_{it}'\Delta y_{it} = \Delta x_{it}'\Delta x_{it}\beta + \Delta x_{it}'\Delta u_{it}$$

and taking expectations under condition (1),

$$E(\Delta x_{it}'\Delta y_{it}) = E(\Delta x_{it}'\Delta x_{it})\beta$$

For a full-rank matrix $E(\Delta x_{it}'\Delta x_{it})$ we can solve for β,

$$\beta = E(\Delta x_{it}'\Delta x_{it})^{-1}E(\Delta x_{it}'\Delta y_{it})$$

which proves the consistency of OLS since,

$$\widehat{\beta}^{OLS} \quad = \quad (\Delta X'\Delta X)^{-1}\Delta X'\Delta Y$$

and, by the WLLN,

$$\begin{aligned}
\text{plim}_{N\to\infty}\frac{\Delta X'\Delta X}{N} \quad &= \quad E(\Delta x_i'\Delta x_i) \\
\text{plim}_{N\to\infty}\frac{\Delta X'\Delta Y}{N} \quad &= \quad E(\Delta x_i'\Delta y_i)
\end{aligned}$$

(b) *Derive its variance assuming that u_{it} is serially uncorrelated and homoskedastic.*

Let

$$v_{it} = \Delta u_{it}$$

Then,

$$\begin{aligned} \mathrm{var}(v_{it}) &= E(v_{it}^2) \\ &= E((u_{it} - u_{it-1})(u_{it} - u_{it-1})) \\ &= E(u_{it}^2) + E(u_{it-1}^2) \\ &= 2\sigma_u^2 \end{aligned}$$

and

$$\begin{aligned} \mathrm{cov}(v_{it}, v_{it-1}) &= E(v_{it} v_{it-1}) \\ &= E((u_{it} - u_{it-1})(u_{it-1} - u_{it-2})) \\ &= -E(u_{it-1}^2) \\ &= -\sigma_u^2 \end{aligned}$$

while for any $j > 1$

$$\begin{aligned} \mathrm{cov}(v_{it}, v_{it-j}) &= E(v_{it} v_{it-j}) \\ &= E((u_{it} - u_{it-1})(u_{it-j} - u_{it-j-1})) \\ &= 0 \end{aligned}$$

Thus,

$$\begin{aligned} \Sigma_{N(T-1)*N(T-1)} &= E(\Delta u \Delta u') = \sigma_u^2 \begin{bmatrix} 2 & -1 & & & & \\ -1 & 2 & -1 & & & \\ & -1 & 2 & \ddots & & \\ & & -1 & \ddots & \ddots & \\ & & & \ddots & \ddots & -1 \\ & & & & -1 & 2 \end{bmatrix} \\ &= \sigma_u^2 \Omega \end{aligned}$$

The asymptotic variance of the OLS estimator can now be derived,

$$\begin{aligned}
\text{var}(\widehat{\beta}^{OLS}) &= E\left((\widehat{\beta}^{OLS}-\beta)(\widehat{\beta}^{OLS}-\beta)'\right) \\
&= E(\Delta X'\Delta X)^{-1}E(\Delta X'\Delta u\Delta u'\Delta X)E(\Delta X'\Delta X)^{-1} \\
&= \sigma_u^2 E(\Delta X'\Delta X)^{-1}E(\Delta X'\Omega\Delta X)E(\Delta X'\Delta X)^{-1}
\end{aligned}$$

and the sample analog of this is,

$$\begin{aligned}
\widehat{\text{var}}(\widehat{\beta}^{OLS}) &= \sigma_u^2(\Delta X'\Delta X)^{-1}\Delta X'\Omega\Delta X(\Delta X'\Delta X)^{-1} \\
&= \sigma_u^2\left(\sum_{i=1}^{N}\Delta X_i'\Delta X_i\right)^{-1}\left(\sum_{i=1}^{N}\Delta X_i'\Omega_i\Delta X_i\right)\left(\sum_{i=1}^{N}\Delta X_i'\Delta X_i\right)^{-1}
\end{aligned}$$

(c) *Compare its variance to that of the within groups estimator for β. (Hint: one of the difficulties arises from the fact that $\triangle\ u_{it}$ is an $MA(1)$ process. Hence, there is a special form of serial correlation.)*

Notice that,

$$\Delta X_i = R'X_i$$

where the matrix R is of dimension $T*(T-1)$ and can be defined as,

$$R_{T*(T-1)} = \begin{bmatrix}
1 & 0 & 0 & \dots & 0 & 0 \\
-1 & 1 & 0 & \dots & 0 & 0 \\
0 & -1 & 1 & \dots & 0 & 0 \\
\vdots & \vdots & \vdots & & \vdots & \vdots \\
0 & 0 & 0 & \dots & -1 & 1 \\
0 & 0 & 0 & \dots & -0 & -1
\end{bmatrix}$$

The covariance matrix for the OLS estimator on the first differences can now be written as,

$$\widehat{\text{var}}(\widehat{\beta}^{OLS}) = \sigma_u^2\left(\sum_{i=1}^{N}X_i'RR'X_i\right)^{-1}\left(\sum_{i=1}^{N}X_i'R\Omega_i RX_i\right)\left(\sum_{i=1}^{N}X_i'RR'X_i\right)^{-1}$$

Now notice that,

$$X_i'RR'X_i - X_i'R\Omega R'X_i = X_i'R(I_{T-1}-\Omega)R'X_i$$

and this is negative semi-definite since $I_{T-1} - \Omega$ is *nsd.*

Thus,

$$\begin{aligned}\widehat{\text{var}}\left(\widehat{\beta}^{OLS}\right) &\geqslant \sigma_u^2 \left(\sum_{i=1}^{N} X_i' R R' X_i\right)^{-1} \\ &= \sigma_u^2 \left(\Delta X' \Delta X\right)^{-1}\end{aligned}$$

The Within Groups estimator is an alternative to OLS on the first differences. We have seen that,

$$\widehat{\text{var}}\left(\widehat{\beta}^{WG}\right) = \sigma_u^2 \left(\widetilde{X}'\widetilde{X}\right)^{-1}$$

where $\widetilde{x}_{it} = x_{it} - \overline{x}_i$.

For a one regressor case, these formulas yield,

$$\begin{aligned}\text{var}(\widehat{\beta}^{OLS}) &= \frac{\sigma_u^2}{N(T-1)E\left((\Delta x_{it})^2\right)} \\ \text{var}(\widehat{\beta}^{WG}) &= \frac{\sigma_u^2}{NTE\left((x_{it} - \overline{x}_i)^2\right)}\end{aligned}$$

Thus, in general, $\text{var}(\widehat{\beta}^{OLS}) > \text{var}(\widehat{\beta}^{WG})$.

5. *Suppose you wish to estimate a dynamic model of the form*

$$\begin{aligned}y_{it} &= \beta x_{it} + f_i + u_{it} \\ u_{it} &= \rho u_{it-1} + e_{it}\end{aligned}$$

where f_i is an unobserved fixed effect and the unobservables e_{it} are independent and identically distributed over time. The single regressor x_{it} may be correlated with f, is uncorrelated with e_{it} but is not strictly exogenous.

(a) *Derive a consistent estimator for β. State carefully any assumptions you might have to make and also the minimum number of observations required for estimation.*

OLS, GLS and WG are all inconsistent.

IV in first differences faces serious problems with the selection of an appropriate instrument: the use of the 1st differences method to get rid off f_i does not solve the problem because u is an autoregressive process.

The solution is to apply the Cochrane-Orcut (1949) transformation:

i. take lags:

$$y_{it-1} = \beta x_{it-1} + f_i + u_{it-1}$$

ii. multiply by ρ:

$$\rho y_{it-1} = \rho\beta x_{it-1} + \rho f_i + \rho u_{it-1}$$

iii. subtract from the original model

$$y_{it} - \rho y_{it-1} = \beta x_{it} - \rho\beta x_{it-1} + f_i(1-\rho) + u_{it} - \rho u_{it-1}$$

iv. substitute $u_{it} = \rho u_{it-1} + \epsilon_{it}$:

$$y_{it} - \rho y_{it-1} = \beta x_{it} - \rho\beta x_{it-1} + f_i(1-\rho) + \rho u_{it-1} + \epsilon_{it} - \rho u_{it-1}$$

to obtain

$$y_{it} = \rho y_{it-1} + \beta x_{it} - \rho\beta x_{it-1} + f_i(1-\rho) + \epsilon_{it} \tag{2}$$

v. and now take first differences to get rid of the fixed effect,

$$\Delta y_{it} = \rho \Delta y_{it-1} + \beta \Delta x_{it} - \rho\beta \Delta x_{it-1} + \Delta\epsilon_{it} \tag{3}$$

Now we have the common problem in dynamic panel data: $\Delta\epsilon_{it}$ is correlated with Δy_{it-1} because ϵ_{it-1} is correlated with y_{it-1} but $\Delta\epsilon_{it}$ is not with further back lags of Δy.

Moreover, x_{it} is not strictly exogenous but it is uncorrelated with the present innovation ϵ_{it}. Assume x_{it} is related with past values of ϵ, that is $\epsilon_{it-1}, \epsilon_{it-2}, ...$ Then Δx_{it} is correlated with $\Delta\epsilon_{it}$ in model (3) since x_{it} is correlated with ϵ_{it-1}. Δx_{it-1}, however, is uncorrelated with the residual $\Delta\epsilon_{it}$.

Hence, we need to find instruments for Δy_{it-1} and Δx_{it}. Take, for example y_{it-2} and x_{it-1}. We can form the matrices X and Z,

$$\begin{aligned} X &= \begin{bmatrix} \Delta Y_{(-1)} & \Delta X & \Delta X_{(-1)} \end{bmatrix} \\ Z &= \begin{bmatrix} Y_{(-2)} & X_{(-1)} & \Delta X_{(-1)} \end{bmatrix} \end{aligned}$$

The GMM estimator is,

$$\widehat{\begin{bmatrix} \rho \\ \beta \\ \rho\beta \end{bmatrix}} = \left(X'ZGZ'X\right)^{-1} X'ZGZ'Y$$

where,

$$\begin{aligned} G &= \left(\frac{Z'\Omega Z}{N}\right)^{-1} \\ \Omega &= E\left[\Delta\epsilon\Delta\epsilon'\right] \end{aligned}$$

Method of estimation: use $Z'Z$ in a first stage estimation to replace G, estimate the model and obtain $\widehat{\Delta\epsilon}$. Use predicted errors to obtain $\widehat{G}$ and use it to estimate the parameters by GMM in a second stage regression.

(b) *What is the covariance matrix of your estimator?*

The estimated variance covariance matrix is,

$$\widehat{\text{var}}\left(\widehat{\begin{bmatrix} \rho \\ \beta \\ \rho\beta \end{bmatrix}}\right) = \frac{X'Z}{N}\widehat{G}\frac{Z'X}{N}$$

(c) *Suggest a way of testing the hypothesis that $\rho = 0$ and describe a consistent estimator for β under the hypothesis that $\rho = 0$. State carefully any assumptions you might have to make and also the minimum number of observations required for estimation.*

If $\rho = 0$ the model is,

$$\begin{aligned} y_{it} &= \beta x_{it} + f_i + u_{it} \\ u_{it} &= \epsilon_{it} \quad \text{iid} \end{aligned}$$

- In this case, x_{it} is weakly exogenous.
- Then we can use first differences to get rid of the fixed effect.
- As a result, Δx_{it} will be endogenous and we need to use IV.

- Can apply GMM using as instrument x_{it-1}. To do this estimation need at least 2 periods.

The test of hypothesis is based on the possibility that the error term is correlated over time:

$H_0 :\ \rho = 0$, which means $\Delta u_{it} \Delta u_{it-2} = 0$

$H_1 :\ \rho \neq 0$, which means $\Delta u_{it} \Delta u_{it-2} \neq 0$ (and, in fact, u and Δu will be correlated with other lags as well).

Then we know (see lecture notes),

$$\frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} \sum_{t=3}^{T} \widehat{\Delta u}_{it} \widehat{\Delta u}_{it-2}}{\widehat{V}^{1/2}} \overset{a}{\sim} \mathcal{N}(0,1)$$

where,

$$\widehat{V} = \frac{1}{N} \widehat{\Delta u}' \widehat{\Delta u}_{(-2)} \widehat{\Delta u}'_{(-2)} \widehat{\Delta u} + \frac{1}{N} \widehat{\Delta u}'_{(-2)} \widehat{\Delta X} \text{var}\left(\widehat{\beta}\right) \widehat{\Delta X}' \widehat{\Delta u}_{(-2)}$$

and

$$\text{var}\left(\widehat{\beta}\right) = \frac{\Delta X' Z}{N} \widehat{G} \frac{Z' \Delta X}{N}$$

(d) *Would your estimation strategy change if there was no fixed effect when:*

i. $\rho = 0$?

OLS is consistent.

ii. $\rho \neq 0$?

Need to apply transformation but not differences.

6. *You have a panel data set which contains repeated observations on log real annual earnings (`lny`) for a number of individuals. For each individual you also observe the age (`age`) and an education indicator (`educ`). This takes the values of 1 to 4 with 1 being the lowest education group. Finally `year` is an indicator of time and `newid` is a personal identification code. The data is stored in STATA format and sorted by individual and year. It is named incpanel.dta.*

(a) *Estimate by OLS a dynamic earnings equation using as explanatory variables lagged income, education, age, age squared and time dummies.*

Stata commands to describe data

```
set mem 10m
use incpanel.dta
describe
sort newid year
by newid: gen nobs=_N if _n==1
ta nobs
by newid: gen missobs=1 if year[_n]>year[_n-1]+1
ta missobs
rename realinc lny
```

We want to estimate

$$\begin{aligned}\ln y_{it} &= \alpha_0 + \alpha_1 \ln y_{it-1} + \alpha_2 \text{age}_{it} + \alpha_3 \text{age}_{it}^2 \\ &\quad + \alpha_4 \text{deduc2}_i + \alpha_5 \text{dedud3}_i + \alpha_6 \text{dedud4}_i + \sum_{\tau=2}^{23} \beta_\tau \text{dyear}\tau_{it} + f_i + u_{it}\end{aligned}$$

where f_i is an unobserved individual fixed effect.

As we include age_{it}^2, we are assuming that age affects earnings in a non-linear form.

The necessary conditions for OLS to be consistent are:

i. No contemporaneous correlation between u and X: $E(u_{it}x_{it}) = 0$, where x_{it} includes all the explanatory variables. Since the lagged dependent variable is an explanatory variable, this requires that u is uncorrelated overtime.

ii. $E(f_i x_{it}) = 0$, the unobserved effect is not correlated with x_{it}. Only random effects are consistent with the use of OLS.

iii. The rank condition.

Thus, we expect OLS to be biased: at least condition 2 should not hold if we have fixed effects since y_{it-1} is one of the regressors.

Stata commands:

```
tab(year), gen (dyear)
tab(educ), gen (deduc)
gen age2=age^2
by newid: gen laglny=lny[_n-1] if year[_n-1]==year-1
regress lny laglny age age2 educd2 educd3 educd4 dyear3-dyear23
```

- The elasticity of earnings with respect to lagged earnings is 0.75.
- The positive coefficient of age_{it} and negative coefficient of age_{it}^2 reflect that age affects positively earnings but at a decreasing rate. This is consistent with empirical evidence of the concavity of the earnings profile with respect to age.
- The coefficient of the education dummies increases with level of schooling. In this case the comparison group is individuals with the lowest level of education. The dummies coefficients show the differential in earnings for individuals with higher levels of education with respect to level 1.

(b) *Using the* `egen` *function in STATA construct individual means of the data and using these perform a within groups transformation on income.*

Stata commands:

```
by newid: gen obs1 = 1 if _n==1
by newid: egen mlny=mean(lny) if obs1~=1
list newid year lny mlny
by newid: egen mlaglny=mean(laglny) if obs1~=1
by newid: egen mage=mean(age) if obs1~=1
by newid: egen mage2=mean(age2) if obs1~=1

/*create deviations from the mean*/
gen wglny=lny-mlny
gen wglaglny=laglny-mlaglny
gen wgage=age-mage
gen wgage2=age2-mage2
```

(c) *Estimate the model using within groups. Will you include education? Will you include the time dummies? Will you include age and age squared?*

The transformed model is:

$$\begin{aligned}\ln y_{it} - \overline{\ln y}_i &= \alpha_1 \left(\ln y_{it-1} - \overline{\ln y}_{i(-1)}\right) + \alpha_2 \left(age_{it} - \overline{age}_i\right) \\ &\quad + \alpha_3 \left(age_{it}^2 - \overline{age^2}_i\right) + \sum_{\tau=3}^{23} \beta'_\tau \text{dyear}\tau_{it} + \left(u_{it} - \overline{u}_i\right)\end{aligned}$$

We can rewrite the model as:

$$\begin{aligned}\widetilde{\ln y}_{it} &= \alpha_1 \widetilde{\ln y}_{it-1} + \alpha_2 \widetilde{age}_{it} \\ &\quad + \alpha_3 \widetilde{age^2}_{it} + \sum_{\tau=3}^{23} \beta_\tau \text{dyear}\tau_{it} + \widetilde{u}_{it}\end{aligned}$$

- WG requires strict exogeneity. However, as discussed in question a) this type of dynamic model only satisfies weak exogeneity. Hence, WG is biased.
- The time dummies are not transformed because it would be as rescaling them. Again, the first time dummy is dropped because we have one lagged regressor.

Stata command:

```
regress wglny wglaglny wgage wgage2 dyear3-dyear23, noconstant
```

(d) *Explain what the coefficients on the time dummies mean.*

The coefficients of the time dummies capture the aggregate shocks in a specific year that affect earnings of all individuals in that particular year.

(e) *Create the first differences of income.*

Stata commands

```
sort newid year
quietly by newid: gen dlny=lny-lny[_n-1]
quietly by newid: gen dlaglny=laglny-laglny[_n-1]
quietly by newid: gen dage=age-age[_n-1]
quietly by newid: gen dage2=age2-age2[_n-1]
```

(f) *Estimate the model in first differences using OLS and then IV. Will you include time dummies, age and age squared. What instruments did you use? Compare the results.*

The model is,

$$\begin{aligned}\Delta \ln y_{it} &= \alpha_1 \Delta \ln y_{it-1} + \alpha_2 \\ &\quad + \alpha_3 \Delta age^2_{it} + \sum_{\tau=4}^{23} \beta_\tau \text{dyear}\tau_{it} + \Delta u_{it}\end{aligned}$$

We got rid off the fixed effect and time invariant regressors. OLS requires non-contemporaneous correlation for consistency and unbiasedness. But this is not satisfied in this equation,

$$E(\Delta \ln y_{it-1} \Delta u_{it}) \neq 0$$

since u_{it-1} determines y_{it-1}. Therefore, OLS is biased.

As we have endogeneity problems, $\ln y_{it-1}$ is correlated with u_{it-1}, so we can use IV. The instruments should satisfy:

- Rank condition: $E(z_{it} \Delta \ln y_{it-1}) \neq 0$
- Order condition: $E(z_{it} \Delta u_{it}) = 0$

The minimum number of periods to estimate the equation with IV is 3. We could use as instrument $\ln y_{it-2}$.

Stata commands:

```
/* OLS estimates */
regress dlny dlaglny dage2 dyear4-23
/* IV estimates */
sort newid year
by newid: gen z=lny[_n-2] if year[_n-2]==year-2
ivreg dlny dage2 dyear4-dyear23 (dlaglny=z)
```

(g) *Comment on the validity of the standard errors that the package provides in each case.*

When using OLS, the standard errors of the package may not be correct because the estimation procedure does not take into account heteroskedasticy or correlation. However, we would like to estimate a covariance matrix that is robust to heteroskedasticity and serial correlation. OLS estimation procedure assumes homoskedasticity and no serial correlation. To correct this problem the Stata has the option to include a command: robust.

When using the WG estimator, the estimate of σ_u is inconsistent unless T grows. Need to correct that estimate and re-scale the variance estimates.

7. *Use Cornwell.dta*

   (a) *Estimate a random effects and a fixed effects models relating the logarithm of crime rate to lprbarr, lprbconv, lprbpris, lavgsen, and lpolpc.*

   Stata commands:

   ```
   xtreg lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc yeard2-yeard7, fe
   xtreg lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc yeard2-yeard7, re
   xthausman
   ```

   (b) *Compute the regression-based version of the Hausman test comparing RE and FE.*

   Stata commands:

   ```
   xtreg lcrmrte lprbarr lprbconv lprbpris lavgsen lpolpc lwcon-lwloc yeard2-yeard7,
   fe
   testparm lwcon-lwloc
   ```

   (c) *Add the wage variables (in logarithmic form), and test for joint significance after estimation by fixed effects.*

   Stata commands:

   ```
   sort county year
   by county:  gen clwcon=lwcon-lwcon[_n-1]
   by county:  gen clwtuc=lwtuc-lwtuc[_n-1]
   by county:  gen clwtrd=lwtrd-lwtrd[_n-1]
   by county:  gen clwfir=lwfir-lwfir[_n-1]
   ```

```
by county:  gen clwser=lwser-lwser[_n-1]
by county:  gen clwmfg=lwmfg-lwmfg[_n-1]
by county:  gen clwfed=lwfed-lwfed[_n-1]
by county:  gen clwsta=lwsta-lwsta[_n-1]
by county:  gen clwloc=lwloc-lwloc[_n-1]
reg clcrmrte clprbarr clprbcon clprbpri clavgsen clpolpc clwcon-clwloc yeard3-yeard7
```

(d) *Estimate the equation by first differences, and comment on any notable changes. Do the standard errors change much between fixed effects and first differences?*

Stata commands:

```
predict ehat, resid
sort county year
by county:  gen lagehat=ehat[_n-1]
reg ehat lagehat
```