Assignment 1:
Panel Data Analysis of Microeconomic Decisions

Daniel Redel & André Couder

Fall 2022

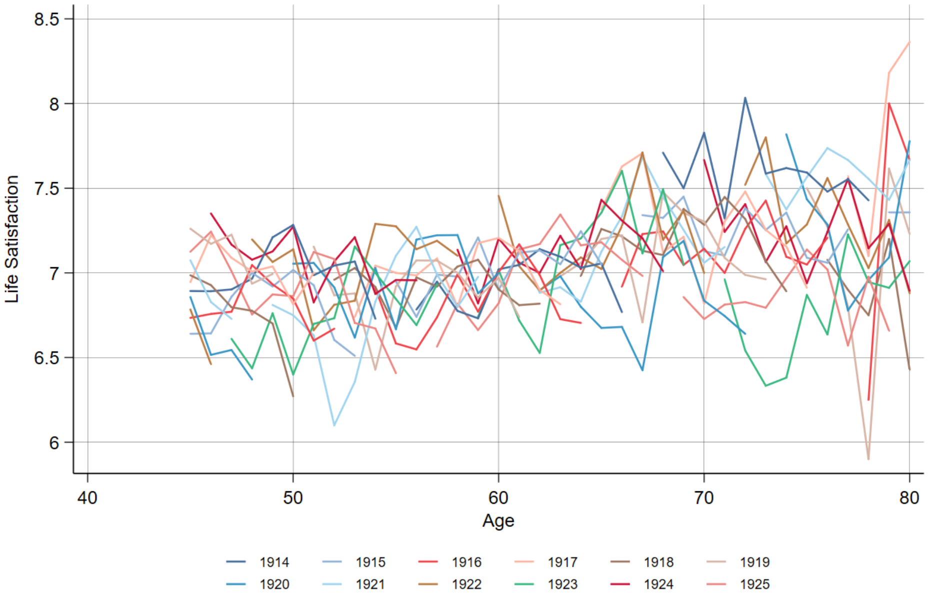
1 The effects of age, cohort and time.

a. It is not possible to run a regression of subjective life satisfaction against `year`, `age`, and `cohort` at the same time because of a **collinearity problem**. In this case, the variable `cohort` is a function of both `year` and `age` ($cohort_i = year_i - age_i$). In such cases, the matrix X has **less than full rank**, as `cohort` is a linear combination of other variables, and therefore $X'X$ cannot be inverted. This leads to non-identifiable parameters when using OLS estimator since it involves inverting the matrix X:

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

b. Figure 1 shows the average life satisfaction by age and cohorts:

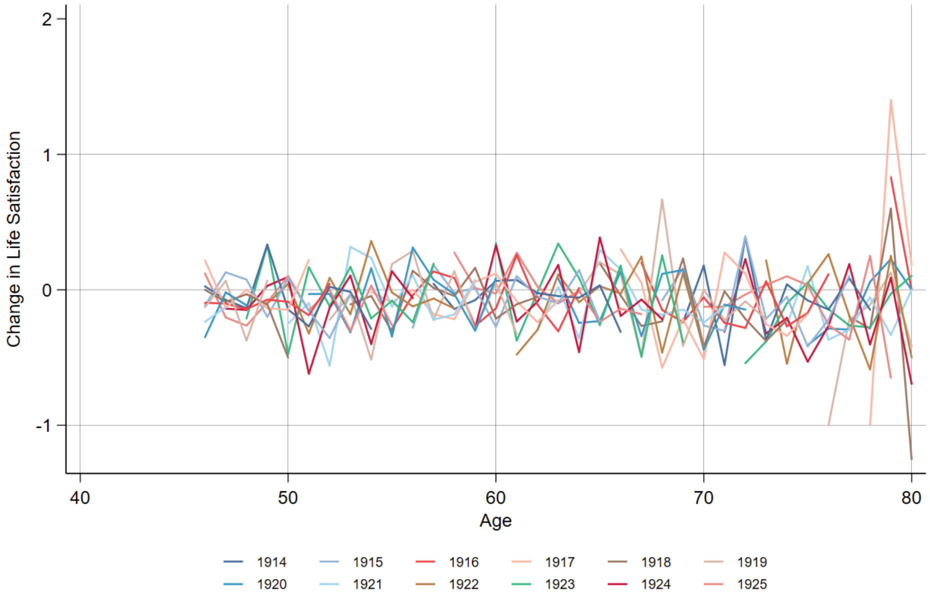
Figure 1: Life Satisfaction by Age and Cohort



We see that the correlation between life satisfaction and age is not unambiguously positive, but varies across different cohorts. For some generations, life satisfaction increases with age, for others life satisfaction is worse as they grow older. Additionally, we see smaller differences in life satisfaction between cohorts in the 55-65 group age than in the younger group 45-44. We see even more dispersion across cohorts at the end stage of life (70-80). One problem with this data is that we do not observe the well-being at every age of each cohort.

Figure 2 shows the average change in life satisfaction:

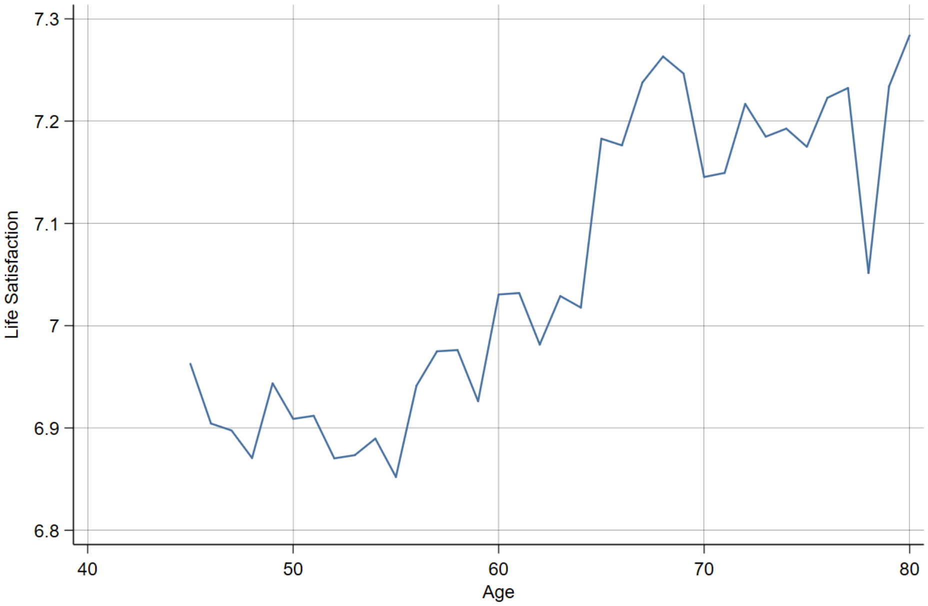
Figure 2: Changes in Life Satisfaction by Age and Cohort



Here we can see that, on average, there are no big changes in life satisfaction at different ages and cohorts, as they all tend to be around 0. Near the age of 80, however, we can see more variation in the changes in life satisfaction, as some cohorts did suffer a negative change in life satisfaction, were others did not experience much change from the past period. Thus, just by looking at the figures presented, there is no clear relationship between subjective life satisfaction and age, if any.

c. Figure 3 shows the average life satisfaction grouping all the cohorts:

Figure 3: Life Satisfaction by Age



In this case, we observe a more evident positive relationship between age and life satisfaction. While between 45 and 55 years old there isn't much correlation, from year 55 to 80 there is a clear positive trend, that is, life satisfaction increases with age. This figure is somehow different from our interpretation of previous Figures 1 and 2, the latter showing a less clear correlation or even zero correlation.

One way to reconcile these two observations is by recognizing the presence of Simpson's Paradox: a trend appears at the group (cohort) level of data but disappears (or reverses) at the aggregate level. This may be the case if cohorts have different sizes or because of larger effects in some cohorts than others.

2 One draw of simulated data.

a. The DGP is generated by the following code:

- `set seed 345398`: This command defines the initial value of the random-number seed used by random-number generators in STATA. With this, we are making sure that the sample we are generating is reproducible, that is, we are always working with the same random sample, because the generator will create the same pseudo-random numbers. In that sense, the numbers generated are a function of the specified seed.

- `drawnorm alpha_i, n(200)`: This line of command draws 200 random numbers from a normal distribution, and then storing them in a variable `alpha_i`. The default is to draw the numbers from a standard normal distribution. In this case, this variable α_i captures all the time-invariant differences across the 200 units (e.g. individuals, households) of the sample. These individual features are given and do not change over time.

- `expand 5`: Here we are duplicating each observation of the random sample 5 times. This command is used for creating our time dimension $T = 5$ and is done after generating the individual unobserved effects in order to make α_i time-invariant.

- `drawnorm nu_it e_it, n(1000)`: As already stated, this line creates two additional random drawings from a standard normal distribution and storing them in variable `nu_it` and `e_it`, respectively. Since this line of code is generating 5×200 numbers across periods, these variables are time-variant.

- `g x_it=nu_it+alpha_i`: Here we are generating our independent variable x_{it} (time-variant) with the `generate` function. In this case, the independent variable is defined by the sum of `nu_it` (which follows a standard normal distribution) and the unobserved effects α_i, so that $Cov(x_{it}, \alpha_i) \neq 0$.

- `drop nu_it`: This command drops the variable `nu_it` from the dataset, as it is already contained in x_{it} and we will not need it anymore.

- `g y_it=3+alpha_i+2*x_it+e_it`: Finally, we generate the dependent variable y_{it} with the `gen` function. The underlying model ends up specified by the following linear relationship:

$$y_{it} = \beta_0 + \beta_1 x_{it} + \alpha_i + u_{it}$$
$$y_{it} = 3 + 2x_{it} + \alpha_i + u_{it}$$

b. Table 1 reports the correlation matrix between the variables in our dataset, which was generated with the command `pwcorr` and enabling the `sig` option to report the significance levels of each pairwise correlation in parenthesis:

Table 1: Correlation Matrix

	α_i	ε_{it}	x_{it}	y_{it}
α_i	1			
ε_{it}	-0.0041	1		
	(0.8963)			
x_{it}	0.7238	-0.0032	1	
	(0.0000)	(0.9198)		
y_{it}	0.8168	0.2549	0.9471	1
	(0.0000)	(0.0000)	(0.0000)	

Note: Significance level in parentheses.

Here we can see a strong positive correlation between our independent variable x_{it} and the unobserved effects α_i (0.7238). This result comes from the fact that the DGP defined x_{it} as the sum of some error term and the individual effects α_i.

If we were to run an OLS estimator from a regression of y_{it} on x_{it}, we should expect the estimate to be upward biased. Recall the OLS estimator:

$$\begin{aligned}\hat{\beta}_1 &= \frac{Cov(x_{it}, y_{it})}{V(x_{it})} = \frac{Cov(x_{it}, 3 + 2x_{it} + \alpha_i + u_i)}{V(x_{it})} \\ &= \frac{Cov(x_{it}, 3)}{V(x_{it})} + 2\frac{Cov(x_{it}, x_{it})}{V(x_{it})} + \frac{Cov(x_{it}, \alpha_i)}{V(x_{it})} + \frac{Cov(x_{it}, u_i)}{V(x_{it})}\end{aligned}$$

The first term goes to 0 because β_0 is a constant[1]. We also know that $Cov(X_i, X_i) = V(X_i)$. Because of strict exogeneity $E(u_{it}|x_{it}) = 0$, we can see that $Cov(x_{it}, u_{it}) = 0$. So, we end up having:

[1] recall property $Cov(X_i, A) = 0$

$$\begin{aligned}
\hat{\beta}_1 &= 2 + \frac{Cov(x_{it}, \alpha_i)}{V(x_{it})} \\
&= 2 + \frac{Cov(\alpha_i + v_{it}, \alpha_i)}{V(\alpha_i + v_{it})} \\
&= 2 + \frac{Cov(\alpha_i, \alpha_i) + Cov(v_{it}, \alpha_i)}{V(\alpha_i) + V(v_{it}) + 2Cov(v_{it}, \alpha_i)} \\
&= 2 + \frac{V(\alpha_i)}{V(\alpha_i) + V(v_{it})} \\
&= 2 + \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_v^2} \approx 2 + \frac{1}{2}
\end{aligned}$$

Since the true parameter of $\beta_1 = 2$ is positive and the correlation between x_{it} and α_i is also positive, we expect OLS to overestimate the parameter.

c. Once we take into account the unobserved effects α_i, we can have unbiased estimates using Pooled OLS:

Table 2: Pooled OLS Results

	(1)	(2)
	POLS 1	POLS 2
x_{it}	2.534***	2.000***
	(0.026)	(0.032)
α_i		0.996***
		(0.043)
Const.	3.117***	3.060***
	(0.039)	(0.031)
R^2	0.897	0.933
N	1000	1000

As we already stated, a simple regression without controlling for α_i will overestimate the true parameter β_1 by about 0.5 units, as it is confirmed in Column (1). Taking α_i into account dissipates this bias, and now we get consistent and unbiased estimates for β_1, as seen in Column (2).

Note, however, that the unobserved effect α_i for each individual i does not change over time, inducing serial correlation across the composite error terms $\epsilon_{it} = \alpha_i + u_{it}$, even with strict exogeneity. Recall the covariance of different composite error terms:

$$Cov(\epsilon_{it}, \epsilon_{is}) = Cov(\alpha_i + u_{it}, \alpha_i + u_{is}) = Cov(u_{it}, u_{is}) + \sigma_\alpha^2$$

Even if we assume $Cov(u_{it}, u_{is}) = 0$, we still have σ_α^2, affecting the efficiency of our estimation. Therefore, the standard errors computed by OLS are not correct, as the covariance matrix

$\Omega \neq \sigma^2 I_N$ and, thus, the variance will not reduce to the usual expression $\sigma^2(X'X)^{-1}$. If we know the structure of the error covariance matrix, we can exploit this to get more efficient estimates by running a GLS estimator.

3 Many draws of simulated data.

a. The following table reports the results of the Monte Carlo Simulation that regresses y_{it} against x_{it}:

Table 3: Summary Results of Monte Carlo Simulation

Variable	Mean	Std. Dev.	Min	Max
β_1	2.498296	.0390307	2.39484	2.586649
β_0	3.00387	.0517515	2.887347	3.123169
$se(\beta_1)$	.027637	.0008765	.0254759	.0295339
$se(\beta_0)$	.0388497	.000763	.0369637	.0409154
N	100			

We observe that the standard deviation of the resulted β_1's from the 100 simulated samples (0.039) is around 41% higher than the average estimate of the standard error $se(\beta_1)$ (0.028). This is because the variance-covariance matrix generated by the standard `reg` command is not properly defined —off-diagonals are not equal to zero—, underestimating the true variance of β_1 since there is autocorrelation:

$$Cov(\varepsilon_{it}, \varepsilon_{is}) = Cov(\alpha_i + u_{it}, \alpha_i + u_{is}) = Cov(u_{it}, u_{is}) + \sigma_\alpha^2$$

Even if we can assume strict exogeneity $Cov(u_{it}, u_{is}) = 0$, we still have σ_α^2: the presence of the time-invariant α_i generates dependence between the error terms across periods, thus, higher dispersion. The variance of β_1 for the Pooled OLS estimator is defined by:

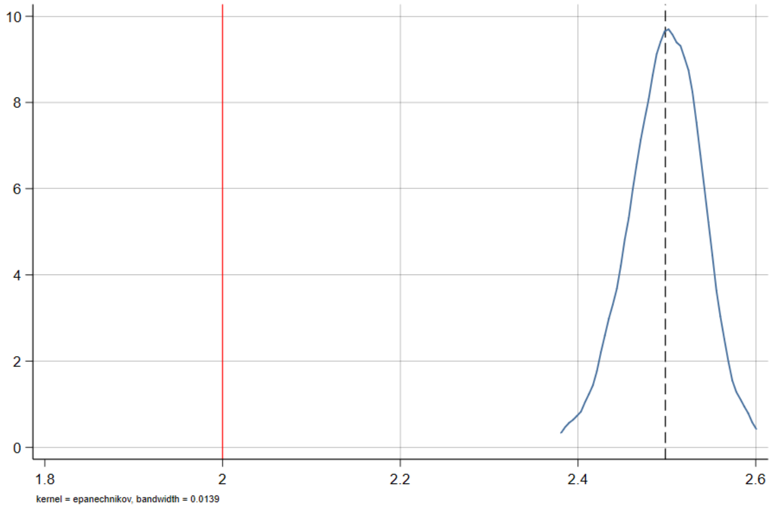
$$V(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

Standard OLS would —mistakenly— assume homoskedasticity and **zero serial correlation** across periods, $\Omega = \sigma^2 I_N$, reducing the variance to the usual expression $V(\hat{\beta}) = \sigma^2(X'X)^{-1}$. The Monte Carlo Simulation, on the other hand, numerically computes a better approximation of the true value of the standard deviation, reporting higher values. The problem, however, is that not controlling for α_i also results in a biased estimator of β_1.

b. As we already stated in Section N°2, a simple regression without controlling for α_i will overestimate the true parameter β_1 by about 0.5 units because of an Omitted Variable Problem.

The following Figure shows the distribution of the resulted β_1's generated from the Monte Carlo simulation:

Figure 4: Distribution of β_1



As we can see, the estimate is upward biased because of the positive correlation between x_{it} and α_i. The partial effect of x_{it} on y_{it} is also incorporating some of the correlation between x_{it} and the unobserved effects α_i.

This is obviously a problem for causal inference, since we are not successfully isolating the true impact of x_{it} on the dependent variable y_{it}. In order to have causal interpretations of our estimates, we need unbiasedness. In prediction, on the other hand, we are more interested in predictive power (accuracy) and are often more willing to accept some bias in order to reduce the variance of our predictions.

c. The following table reports the results of a Monte Carlo Simulation that runs 100 times (i) a Pooled OLS, (ii) a Clustered Pooled OLS, and (iii) a Random Effects Estimator:

Table 4: Summary Results of Monte Carlo Simulation - Model Comparison

	Mean	Std. Dev.	Min	Max
β_1^{POLS}	2.498296	.0390307	2.39484	2.586649
$\beta_1^{Clust.}$	2.498296	.0390307	2.39484	2.586649
β_1^{RE}	2.371546	.0462677	2.26433	2.488911
$se(\beta_1^{POLS})$	.027637	.0008765	.0254759	.0295339
$se(\beta_1^{Clust.})$	.0353637	.0027013	.0294286	.0458886
$se(\beta_1^{RE})$	.0294961	.0007903	.0276135	.0310708
N	100			

In Table 4 we observe that the estimated standard errors $se(\beta_1^{Clust.}) = 0.035$ of the **Clus-**

tered POLS are now closer to the standard deviation of $\beta_1^{Clust.}$ (0.039) because both heteroskedasticity and —especially— autocorrelation within individuals across periods t are taken into account. Compared to the simple POLS regression, clustering the errors by individual reports higher standard errors. The key assumption when clustering is that errors are uncorrelated across clusters (block diagonal), while errors for periods belonging to the same cluster may be correlated. Recall our variance:

$$V(\hat{\beta}) = (X'X)^{-1}X'\Omega X(X'X)^{-1}$$

Given error independence across clusters, we can derive the following equation:

$$X'\Omega X = \sum_{c=1}^{C} X_c'\Omega_c X_c = \sum_{c=1}^{C} X_c' V(u_c|X_c) X_c = B_c$$

The term B_c, and hence $V(\hat{\beta}_1)$, will be bigger when (i) regressors X within cluster are correlated (ii) errors terms within-cluster are correlated and (iii) the within-cluster regressors and error correlations are of the same sign. According to Cameron & Douglas (2015), and estimate of B_c is:

$$\hat{B}_c = \sum_{c=1}^{C} X_c'\hat{u}_c\hat{u}_c' X_c$$

So that:

$$V(\hat{\beta}) = (X'X)^{-1}\sum_{c=1}^{C} X_c'\hat{u}_c\hat{u}_c' X_c(X'X)^{-1}$$

This estimator allows for general forms of heteroskedasticity as well as arbitrary autocorrelation (within a given individual).

The standard errors of the β_1^{RE} in the **Random Effects Estimator** is lower than the Clustered POLS. This is because random effects analysis adds assumptions on the idiosyncratic errors that give the covariance matrix Ω a specific structure that results in more efficient estimates:

$$\Omega_{RE} = \sigma_u^2\mathbf{I}_T + \sigma_\alpha^2\mathbf{j}_T\mathbf{j}_T'$$

This is the result of assuming homoskedasticity $E(u_{it}^2) = \sigma_u^2$ and zero serial correlation of the *idiosyncratic errors* $E(u_{it}u_{is}) = 0$, but **strong persistence** in the unobservables over time due to α_i. FGLS estimator exploits this error covariance structure to gain efficiency:

$$V(\hat{\beta}) = (X'\hat{\Omega}_{RE}^{-1}X)^{-1}$$

Intuitively, the gain in efficiency is due to the use of the *between-variation* in the data $(\bar{x}_i - \bar{x})$. The FGLS estimator, under the current assumptions, is the optimal combination of the within-estimator and the between-estimator and is, therefore, more efficient than either of these two. The problem, however, is that RE estimator relies on $Cov(\alpha_i, x_{it}) = 0$, which is not the case in our simulated data. Additionally, a cluster-robust covariance RE estimator that relaxes the homoskedasticity assumption should get us closer to the true value, which is indeed the case, as $se(\beta_1^{Clust.RE}) = 0.035$.

4 Fixed Effects and First Differences Estimation.

We will compare different panel data models used to analyze the simulated data we generated in Section N°2. Table 5 reports the results of our estimator of interest:

Table 5: Fixed Effects vs. First Differences Estimation

	First Differences	Fixed Effects	Random Effects
	Δy_{it}	y_{it}	y_{it}
Δx_{it}	2.010***		
	(0.037)		
x_{it}		1.989***	2.407***
		(0.035)	(0.029)
Const.		3.179***	3.131***
		(0.031)	(0.047)
N	800	1000	1000

a. Both the FE and FD estimator are useful models when in presence of time-invariant omitted variables that correlate to our independent variables $Cov(x_{it}, \alpha_i) \neq 0$. The central idea in both models is to removed the individual specific effect α_i via running an OLS estimator on transformed data.

The **FE estimator** uses a *within transformation* to eliminate α_i, such that:

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

This is a regression model in deviations from individual means. Since α_i is time-invariant, time-demeaning the data eliminates α_i. Now, to yield consistent estimates, we need **Strict Exogeneity** assumption:

$$E(u_{it}|x_{i1}, ..., x_{iT}) = 0$$

This assumption implies that explanatory variables x_{it} in each time period are uncorrelated with error terms u_{it} in each time period, so that the following orthogonality condition holds:

$$E[(x_{it} - \bar{x}_i)(u_{it} - \bar{u}_i)] = 0$$

Note that this is a stronger assumption than Contemporaneous Exogeneity, $E(x_{it}u_{it}) = 0$, because this assumption does not ensure that x_{is} is uncorrelated with u_{it}, for all $s \neq t$.

The **FD estimator**, on the other hand, eliminates α_i by subtracting the one-period lag from the main regression. This results in:

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

This estimator also relies on the **Strict Exogeneity** assumption for consistency, such that the following FE orthogonality conditions holds:

$$E[(x_{it} - x_{i,t-1})(u_{it} - u_{i,t-1})] = 0$$

However, it is also possible to arrive at this last condition with a weaker assumption. In fact, $E(\Delta x_{it} \Delta u_{it}) = 0$ by itself is weaker than the Strict Exogeneity condition. For example, it would allow correlation between x_{it} and $u_{i,t-2}$ or $u_{i,t+2}$.

Considering the results in Table 5, we confirm that both FE and FD yield consistent estimators of β_1, as both are very close to the true value 2. But FD estimator also report higher standard errors, implying less efficiency. Generally, FE is more efficient when the u_{it} are serially uncorrelated, while FD is more efficient when these are serially correlated, for example when they follow a random walk. In this case, the FE estimator is to be preferred, although with a small difference.

Regarding the standard errors, the FE estimator needs adjusting if we just run a Pooled OLS of the within transformed model as defined above. In that case, the standard errors will be smaller (0.031) because they use the incorrect estimate of σ_u^2:

$$\hat{\sigma}_u^2 = \frac{SSR}{NT - K}$$

Subtracting the mean from the observations takes away one degree of freedom, therefore $(T - 1)$ is the correct amount of degrees of freedom, such that a consistent estimator of the variance will be defined by:

$$\hat{\sigma}_u^2 = \frac{SSR}{N(T - 1) - K}$$

Typically, the standard errors from the FE estimator are corrected by a factor of $[(NT - K)/(N(T-1) - K)]^{1/2}$. In our exercise, the `xtreg` command in STATA already does that adjustment for us, so the reported standard errors in Table 5 are properly estimated.

In contrast, the FD estimator does not need adjustment: all statistics reported from the Pooled regression on the first-differenced data are asymptotically valid. Unlike in FE model, degrees of freedom are correct, as dropping the first period appropriately captures the lost degrees of freedom.

b. In Table 5 we also observe the results from the RE estimator, but it does not report a consistent estimator of β_1, since we know from the GDP process that $Cov(\alpha_i, x_{it}) \neq 0$, as already stated in Section N°3. RE estimator needs **Strict Exogeneity** regarding the *composite error*, $\varepsilon_{it} = \alpha_i + u_{it}$, to derive the asymptotic RE variance:

$$E(\varepsilon_{it} | x_{i1}, ..., x_{iT}) = 0$$

5 Dynamic Model

Consider the following dynamic linear regression with a lagged dependent variable:

$$y_{it} = \gamma y_{i,t-1} + \beta_1 x_{it} + \alpha_i + u_{it}$$

One possible way to estimate γ is by running a Fixed Effects estimator, which can be rearranged as:

$$\hat{\gamma} = \gamma + \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}(u_{it} - \bar{u}_i)(y_{i,t-1} - \bar{y}_{i,t-1})}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{i,t-1} - \bar{y}_{i,t-1})^2}$$

This estimator is biased and inconsistent for $N \to \infty$ and fixed T. This is because the last term on the right-hand side does not have expectation zero neither converge to zero. In particular, it can be shown that the biased comes from the correlation between the within transformed lagged dependent variables and the within transformed error terms:

$$E[(u_{it} - \bar{u}_i)(y_{i,t-1} - \bar{y}_{i,t-1})] = -\frac{\sigma_u^2}{T^2}\frac{(T-1) - T\gamma + \gamma^T}{(1-\gamma)^2} \neq 0$$

The problem is that the within transformed lagged dependent variable is correlated with the within transformed error. Even if we were to assume **Sequential Exogeneity**, we will only rules out correlation with past values, but fixed effects use it for also future values, generating bias. Table 6 reports the results of applying a FE estimator in a Monte Carlo simulation for 5, 10, 20, and 50 time periods:

Table 6: Nickell Bias in Fixed Effects

T	5	10	20	50	True Value
plim $\hat{\gamma}_{FE}$	0.440	0.471	0.485	0.494	**0.5**
Bias	-0.06	-0.029	-0.01	-0.006	0.0

As already proved, the bias is always negative for $\gamma = 0.5$. As T increases, the FE estimator converges to the true parameter. Generally speaking, we get consistent estimates if both $T \to \infty$ and $N \to \infty$.

6 Instrumental Variables Estimation.

The Arellano-Bond estimator (1991) exploits **internal instrumental variables** on the **first-differenced** version the regression with a lagged independent variable:

$$\Delta y_i = \Delta x_i' \beta + \gamma \Delta y_{i,t-1} + \Delta u_{it}$$

In particular, the AB estimator suggests a list of instruments Z_i derived from a series of moment conditions that vary with t. Because this also means more moment conditions than unknown parameters to be estimated (γ), we need to rely on a Generalized Method of Moments (GMM) method. The AB estimator consistently estimates γ when the set of instruments Z_i satisfy the following conditions:

1. **Relevance Condition:** $E[(y_{i,t-1} - y_{i,t-2})y_{i,t-2-j}] \neq 0$
2. **Exogeneity Condition:** $E[(u_{it} - u_{i,t-1})y_{i,t-2-j}] = 0$

With condition (2) also being the moment conditions by which one can derive the instruments for $\Delta y_{i,t-1}$. Intuitively, the AB estimator is consistent because IV decomposes and isolates the non-problematic component —whose variation can be explained by our instruments Z_i— from the problematic component, which is correlated to our error terms Δu_{it}. This only works if the lags used as instruments Z_i have a significant correlation with Δy_{it} —which is usually the case— and if the instruments are also uncorrelated with each Δu_{it}.

In addition to the usual IV assumptions, Arellano-Bond depends on the **Sequential Exogeneity** assumption and **Zero Serial Correlation** assumption:

1. **Sequential Exogeneity:** $E(u_{it}|y_{i,t-1}, y_{i,t-2}, ..., y_{i0}, \alpha_i) = 0$
2. **Zero Serial Correlation:** $E(u_{it}u_{is}) = 0$ for all $t \neq s$

Regarding the count of lagged variables that can be used as instruments, in this case, we can have up to $T - 1 = 4$ different instruments that stem from $T(T-1)/2 = 10$ moment conditions: for $t = 2$, only y_{i0} satisfies $E[(u_{i2} - u_{i1})y_{i0}] = 0$; for $t = 3$ we have y_{i0} and y_{i1}; for $t = 4$, y_{i0}, y_{i1} and y_{i2}, and so on. The way each instrument is valid for a given period can be summarized by referring to the matrix of instruments Z_i:

$$Z_i = \begin{bmatrix} [y_{i0}] & 0 & 0 & 0 \\ 0 & [y_{i0}, y_{i1}] & 0 & 0 \\ 0 & 0 & [y_{i0}, y_{i1}, y_{i2}] & 0 \\ 0 & 0 & 0 & [y_{i0}, y_{i1}, y_{i2}, y_{i3}] \end{bmatrix}_{4\times 10}$$

Going back to our exercise, the results from running an AB estimator in a Monte Carlo Simulation are reported in Table 7:

Table 7: Summary Results of Arellano-Bond Estimator

	Mean	Std. Dev.	Min	Max
γ	.4930889	.0389845	.4127165	.5863011
β_1	1.928794	.4670293	.6058065	2.878655
$se(\gamma)$	.0465118	.0114783	.0300433	.0991765
$se(\beta_1)$	.5737964	.2773561	.2464988	1.919659
N	100			

We see that both our estimates of γ and β_1 are very close to the true parameters.