Panel Data Analysis for Microeconomic Decision

Mattia Zen – `m.zen@tilburguniversity.edu` – Snr: 2132868

Assignmenet 1

I have checked the code and discussed the assignment together with Natan Ornadel. However, I wrote the final answer on my own.

1 The Effect of Age, Cohort and Time

1.1

The challenge in estimating the relationship between subjective well-being and calendar time, cohort (birth year), and age simultaneously arises due to **perfect multicollinearity**. This occurs because the three variables are linearly dependent: age is the difference between calendar time and birth year (`age = calendar time - birth year`).
In practical terms, this means that the variation in one of these variables is completely explained by the others, making it impossible to isolate their individual effects on subjective well-being. This multicollinearity prevents the regression from distinguishing between the influence of age, calendar time, and cohort, leading to unreliable or indeterminate coefficient estimates.

1.2

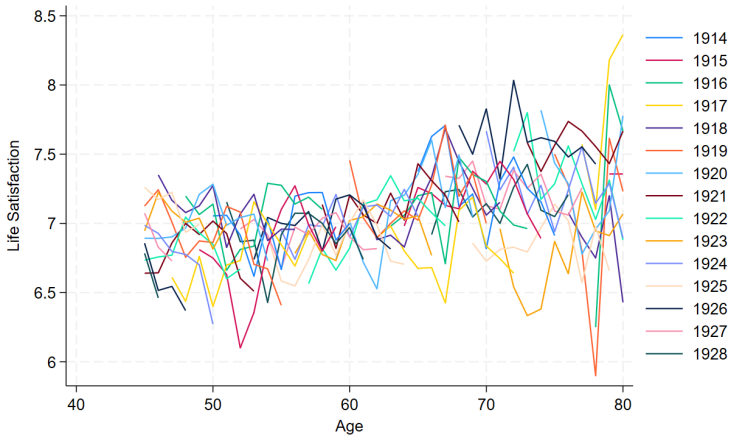

Figure 1: Life Satisfaction against age, different line for each cohort

Figure 1 shows life satisfaction against age. From this graph, we see no clear relations between life satisfaction and age, with some cohorts showing positive trends, while others experiencing decreasing life satisfaction as they age.
Looking at the variation in life satisfaction (Figure 2) we notice that all the variation stays around 0, with only a peak at the age of 80.

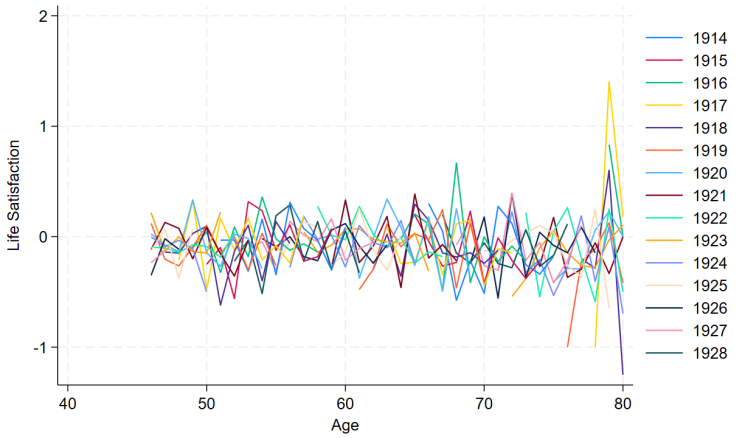

Figure 2: Change in Life Satisfaction against age, different line for each cohort

1.3

In figure 3 we plot the life satisfaction against age, grouping all cohorts together. In this case, we note a positive correlation between the two variables in the range from 55 to 70. This result differs from our previous one. One possible explanation could be the aggregation of all cohorts together and the small positive effects in the previous question sums up in this strongly positive relation.

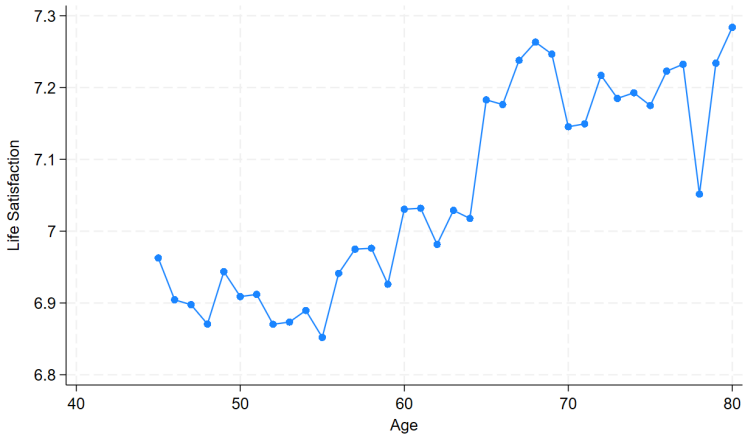

Figure 3: Life Satisfaction distribution by age

2 One draw of simulated data

```
set seed 345398
drawnorm alpha_i, n(200)
expand 5
drawnorm nu_it e_it, n(1000)
g x_it=nu_it+alpha_i
drop nu_it
g y_it=3+alpha_i +2*x_it+e_it
```

2.1

1. setting the seed so that I have the same result every time I run the code
2. Random draw 200 individual terms α_i from a normal distribution. This individual term is constant over time and captures all the time-invariant differences among individuals
3. With this line expand α_i over 5 periods of time, keeping it constant.
4. Random draw the error term and ν_{it} from a normal distribution. This draw is different both among individuals and time.
5. We now have to generate the independent variable as the sum of ν and the individual term. By doing this we have that α_i is correlated with our independent variable x_{it} such that $Cov(x_{it}\alpha_i) \neq 0$
6. We can now drop ν since it is already included in our independent variable. We do it with the `drop` command.
7. The last step of our DGP is to generate the dependent variable y_{it} of our model. The dependent variable is characterised by the presence of an intercept (3), our time-invariant α_i, the slope (2) and the error term e_{it}.

The underlining model from this DGP is:

$$y_{it} = 3 + 2x_{it} + \alpha_i + e_{it}$$

$$x_{it} = \alpha_i + \nu_i t$$

$$n_{it}, \epsilon_{it}, \alpha_i \sim \text{i.i.d. } \mathcal{N}(0,1)$$

2.2

The command `pwcorr, sig` gives me the correlation between all the variables present in our model. As you can see from the table below, there is a strong positive correlation between x_{it} and α_i.
Our result will most likely be biased since I am not properly addressing the issue of the individual-specific error term α_i. Not adding α_i to our estimation means that I am incorporating it in the error term $u_{it} = \alpha_i + e_{it}$. However, since α_i is correlated with x_{it} this violates

Variables	α_i	e_{it}	x_{it}	y_{it}
α_i	1.000			
e_{it}	-0.004	1.000		
	(0.896)			
x_{it}	0.724	-0.003	1.000	
	(0.000)	(0.920)		
y_{it}	0.817	0.255	0.947	1.000
	(0.000)	(0.000)	(0.000)	

our assumption of $Corr(x_{it}u_{it}) = 0$. Recalling from the estimation of $\hat{\beta}_1$ as the fraction between the $Cov(x_{it}y_{it})$ and $Var(x_{it})$ and substituting y_{it} with the model equation we get that:

$$\hat{\beta}_1 = \frac{Cov(x_{it}3)}{Var(x_{it})} + 2\frac{Cov(x_{it}x_{it})}{Var(x_{it})} + \frac{Cov(x_{it}\alpha_i)}{Var(x_{it})} + \frac{Cov(x_{it}e_{it})}{Var(x_{it})}$$

We know that the first and the last term are = 0. We also know that $Cov(x_{it}x_{it}) = Var(x_{it})$ making the second term equal to 2. Thus the bias depends on the $Cov(x_{it}\alpha_i)$ and since there is a positive correlation between α_i and x_{it} (from the table above) we can state that there is an upper bias.

2.3

	y_{it}	y_{it}
x_{it}	2.534***	2.000***
	(0.0272)	(0.0317)
α_i		0.996***
		(0.0428)
_cons	3.117***	3.060***
	(0.0388)	(0.0314)
N	1000	1000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Incorporating α_i in our POLS regression makes our estimation unbiased. In this case, we treat α_i as another variable in our model. In this case, the estimation is not biased and our β_1 converges to its true value of 2. Furthermore, the estimation is efficient. For the OLS model to have an efficient estimator we need to assume that the error term is homoskedastic and serially correlated. By including the individual error term α_i in our equation we are taking out from the error term the part of variation that is serially correlated among time. In our DGP ϵ_{it} is normally distributed among individuals and time, thus it is reasonable to assume that there is homoskedasticity and serial correlation in the error term, after we took α_i out of it.

3 Many draws of simulated data

```
set seed 345398
capture program drop mcprog
program mcprog
        clear
        drawnorm alpha_i, n(200)
        expand 5
        drawnorm nu_it e_it, n(1000)
        g x_it=nu_it+alpha_i
        drop nu_it
        g y_it=3+alpha_i+2*x_it+e_it
        regress y_it x_it
end

simulate _b _se, reps(100): mcprog
sum
```

Variable	Obs	Mean	Std. dev.	Min	Max
_b_x_it	100	2.498296	0.0390307	2.39484	2.586649
_b_cons	100	3.00387	0.0517515	2.887347	3.123169
_se_x_it	100	0.027637	0.0008765	0.0254759	0.0295339
_se_cons	100	0.0388497	0.000763	0.0369637	0.0409154

3.1

Regressing y_{it} on x_{it} with the standard OLS model we are assuming zero serial correlation among periods. This causes the average estimation of the standard error to be underestimated. When using the standard OLS we calculate the variance as $Var(\hat{\beta}) = \sigma^2(X'X)^{-1}$. However, even if the Monte Carlo simulation estimates a higher standard error, the estimation is still biased since we are not considering α_i in our model.

3.2

As mentioned in section 2.2, not addressing α_i in our model causes an upward bias in our estimation. With the Monte Carlo simulation, we empirically see that $\hat{\beta}$ does not converge to the true value of 2 after having estimated the model 100 times. This issue matters especially if we are looking for causality between the two variables. In order to have causal inference we need to isolate the partial effect of x_{it} on the dependent variable. In this case, we incorporate the effect of α_i into the partial effect of x_{it}. Contrarily, if we aim for prediction this bias is less important since our goal is to understand the predictive power of our independent variable on y_{it}.

3.3

Comparing the result obtained in table below we note that the SE of the clustered data and the random effect are similar to each other, while in the POLS it is smaller. We also see that these two standard errors are closer to the standard deviation of the $\hat{\beta}$.
Clustering data allows for heterogeneity and serial correlation within the cluster. This increases the standard error since the amount of independent observation contained in the dataset decreases.
Even if close, the SE_{RE} is smaller than the one of clustered data. This is the case because the Random Effect estimator accounts for heteroskedasticity within the GLS framework. While in the OLS we assume the covariance matrix of the error term to be an identity matrix ($\Omega = I_n$) in the GLS framework we get a BLUE estimation of β under heteroskedasticity thanks to the transformed model. This model Is transformed with the matrix P that follows $\Omega = (P'P)^{-1}$.

	mean	sd	min	max
β_{POLS}	2.494293	.0406014	2.397209	2.583309
SE_{POLS}	.0274331	.0009687	.0250999	.0300382
β_{CL}	2.494293	.0406014	2.397209	2.583309
SE_{CL}	.0350829	.0030678	.0287694	.0420721
β_{RE}	2.372664	.047029	2.279994	2.476421
SE_{RE}	.0347201	.0027344	.0295521	.0423112
N	100			

4 Fixed Effects and First Differences Estimation

	FE	FD	RE
	y_{it}	Δy_{it}	y_{it}
x_{it}	1.979***		2.416***
	(0.0356)		(0.0292)
Δx_{it}		1.977***	
		(0.0356)	
_cons	3.180***		3.130***
	(0.0317)		(0.0458)
N	1000	800	1000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

4.1

To have consistent estimator β_{FE} and β_{FD} we need to assume **strict exogeneity** to eliminate the α_i.
In the **FE effect estimator** we eliminate the α_i by subtracting the variable mean, resulting in a time-demeaning model.

$$y_{it} - \bar{y_i} = \beta(x_{it} - \bar{x_i'}) + (u_{it} - \bar{u_i})$$

For strict exogeneity to hold the $E[u_{it} - \bar{u_i}|x_{it}] = 0$ and we will then estimate $\hat{\beta}_{FE}$ with the same equation as the OLS estimator.

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

In the **FD estimation** we eliminate the α_i subtracting our variable with their one period lagged one.

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

In this case for $E[\Delta x_{it}\Delta u_{it}] = 0$ to hold we need to assume a weaker form of strict exogeneity.

$$E[\Delta x_{it}\Delta u_{it}] = E[(x_{it} - x_{it-1})(u_{it} - u_{it-1})]$$
$$E[x_{it}u_{it}] = 0$$
$$E[x_{it}u_{it-1}] = 0$$
$$E[x_{it-1}u_{it}] = 0$$

In this case, I prefer to use the Fixed Effect estimator since the error terms are serially uncorrelated, as you can see from the table above, this results in $se(\hat{\beta}_{FE}) < se(\hat{\beta}_{FD})$ (in this case the difference is small and cannot be seen from this table, however in the log file attached you see that the SE of the FD model is 0.0356425 for the FE is 0.035616). To have the alternative estimator (First Difference) to be efficient u_{it} should follow a random walk. In this case $u_{it} = u_{it-1} + e_{it}$ resulting in $\Delta u_{it} = e_{it}$.

4.2

As you can see from the table above Random Effect estimation is **not consistent**. Since in the RE estimator, we do not get rid of α_i but we consider it as a random factor we need the assumption of strict exogeneity between the composite error term uit and our independent variable x_{it}.

$$E[u_{it}|x_{it}] = 0$$

$$E[(\alpha_i + e_{it})|(\nu_{it} + \alpha_i)] = 0$$

In our DGP $x_{it} = \nu_{it} + \alpha_i$. Thus, we cannot assume strict exogeneity between the composite error term and the independent variable since part of the variation in both of them is due to α_i

5 Dynamic model

	T=5	T=10	T=20	T=59
	y_{it}	y_{it}	y_{it}	y_{it}
$L.y_{it}$	0.431***	0.471***	0.479***	0.488***
	(25.04)	(49.48)	(75.06)	(124.49)
x_{it}	2.005***	1.954***	1.973***	2.009***
	(45.89)	(77.95)	(118.04)	(192.82)
_cons	3.222***	3.108***	3.178***	3.094***
	(40.00)	(57.54)	(77.67)	(119.68)
N	800	1800	3800	9800

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

To check the unbiasedness of the FE estimator in a lagged dependent variable model we need to rearrange the estimator as:

$$\hat{\gamma}_{FE} = \gamma + \frac{\frac{1}{NT}\sum_i \sum_t (u_{it} - \bar{u}_i)(y_{it-1} - \bar{y}_{i,-1})}{\frac{1}{NT}\sum_i \sum_t (y_{it-1} - \bar{y}_{i,-1})^2}$$

In order for this estimation to be unbiased the $plim N \to \infty$ should be zero. However, this is not the case. Decomposing the fraction we see that the only term that is equal to zero in expectation is the first one. To prove this we need to write all the terms in their expected value substituting $\frac{1}{T} y_{it-1}$ as $\bar{y}_{it-1}$ and do the same for u_{it}. In this case, we obtain $E[\bar{y}_{i,-1}\bar{u}_i]$ which we cannot assume to be equal to zero since we do not have strict exogeneity.

$$E[u_{it} y_{it-1}] = 0$$

$$E[\bar{u}_i \bar{y}_{i,-1}] \neq 0$$

However, as $T \to \infty$ and $N \to \infty$, the value of our bias decreases, as shown in the table below.

	T=5	T=10	T=20	T=50	**True Parameter**
$\hat{\gamma}_{FE}$	0.44	0.46	0.48	0.49	**0.5**
Bias	-0.06	-0.04	-0.02	-0.01	

6 Instrumental variables estimation

In the Arellano-Bond estimator we want to eliminate the α_i and estimate $\hat{\gamma}$ consistently. We achieve this by first taking the first difference of the model.

$$y_{it} - y_{i,t-1} = \gamma(y_{i,t-1} - y_{i,t-2}) + \beta(x_{it} - x_{it-1} + (u_{it} - u_{i,t-1}) \qquad t = 2, \ldots, T$$

Once we have differenced the model we can use the lagged dependent variable as an instrument for the independent variable. As an instrumental variable for the difference $y_{i,t-1} - y_{i,t-2}$ we can use all the $y_{i,t-2-j}$ that satisfy these two conditions:

$$E[(u_{it} - u_{it-1})y_{i,t-2-j}] = 0$$
$$E[(y_{it-1} - y_{i,t-2})y_{i,t-2-j}] \neq 0$$

If these conditions hold I can derive the matrix of instruments (Z_i) as:

$$Z_i = \begin{bmatrix} [y_{i0}] & 0 & 0 & ... & 0 \\ 0 & [y_{i0}, y_{i1}] & 0 & ... & 0 \\ 0 & 0 & [y_{i0}, y_{i1}, y_{i2}] & ... & 0 \\ ... & ... & ... & ... & ... \\ 0 & 0 & 0 & ... & [y_{i0}, y_{i1}, ...y_{iT-2}] \end{bmatrix}$$

The AB estimator is consistent because it isolates the variation correlated with u_{it} from the one that is explained by the instrument matrix Z_i.
In this specific case, we have 4 different instruments $(T-1)$ and 10 different moment conditions $(T(T-1)/2)$.