Problem set 15: week 19

2. Population model:

$$slp_{it} = a + d81_t + \beta_1 W_{it} + \beta_2 Educ_{it} + \beta_3 marr_{it} + \beta_4 kids_{it} + \beta_5 health_{it} + a_i + u_{it}$$

3.

- $b_1 = -0.150$, $se(b_1) = 0.0244$ and $t = -6.15$.
- Working one more minute per week reduces the total weekly sleeping time by 0.15 minutes on average, holding education, general health, marriage status and number of young kids constant.
- Practically speaking, this is a large effect: It implies working 1 extra hour per week reduces sleep by 9 minutes. $(0.15 * 60 = 9)$
- The effect is very statistically significant.
- Our estimator will be unbiased if $E[a_i + u_{it}|W_{it}] = 0$. This is a highly dubious assumption, given a_i is likely to contain factors such as organisation, likely to be correlated with both sleep and work. We return to this point shortly.

4.

- Even if we are willing to maintain the assumption $E[a_i + u_{it}|W_{it}] = 0$.
- We have the problem that $cov(v_{it}, v_{is}) = \sigma_a^2 \neq 0$. Thus, standard error estimation will be incorrect.
- We didn't consider this problem as much with cross-sectional data, since typically, we assumed we had a random sample such that there is no need to suppose the unobserved errors were correlated. That is $cov(\epsilon_i, \epsilon_j|X) = 0$.
- Note, however, this was not guaranteed. A context when this could fail, for example, is the following. Zero correlation between the errors could fail, for example, if our sample contains students from within the same classes, or individuals from the same neighbourhoods. Then unobserved shocks in that neighbourhood/class are likely to cause correlation across the errors of individuals in those neighbourhoods/classes.
- In the cross-section context, we only had one concern correlation in errors across individuals. In panel data, we have two concerns.

1. We are still concerned with whether errors are correlated across individuals.
2. Now also have the additional (more pressing) concern that the errors are correlated across time. This problem is not so easy to assume away (as when did with random sampling in cross-sectional data). As the a_i will be present in all time-periods: causing correlation across time.

- We have two choices to deal with serial correlation: allow for arbitrary correlations by clustering the standard errors, or model the serial correlation with random effect estimation.

5.

- To run POLS with clustered standard errors, use the cluster() option. See do file.
- $b_1 = -0.150$, $se(b_1) = 0.020$ and $t = -5.70$.
- The coefficients are exactly the same. The standard errors have increased marginally. Intuitively, this increase has corrected the standard errors for the presence of serial correlation in the errors within individuals across time (which is typically positive).

6.

- $b_1 = -0.167$, $se(b_1) = 0.025$ and $t = -6.75$.
- The coefficient in the RE model is marginally higher than the POLS (and clustered). The standard errors are marginally different. Under both clustering and RE, nothing much has happened compared to the POLS.

7.

- Our more pressing concern is issues of bias, rather than serial correlation. In particular we are concerned about unobserved heterogeneity (individual unbservables that do not vary across time), since we can potentially account for it. This is bias caused by the a_i in our initial model.
- It is likely the same factors (some biological) that cause individuals to sleep more or less (these are captured by a_i) are also correlated with the amount of time spent working. For instance, some individuals have more or less energy, which allows them to sleep less and work more. Some are just more organised, ... etc.
- Therefore, not accounting for it may, for instance lead to an upward bias: $E[b_1] = \beta_1 + \beta_2 cov(w_{it}, a_i)/V(w_{it}) = \beta_1 + (+)(+) > \beta_1$. Given more organised people may sleep more ($\beta_2 > 0$) and correlated with work ($cov(w_{it}, a_i) > 0$).

8.

- First-dfferencing or fixed effects allows for arbitrary correlation between the individual unobserved heterogeneity and time spent working.
- That is, when using FD of FE, we need to assume:
  - $E[u_{is}|W_{it}] = 0$, for all $s = 0, ..., T$.
  - However we don't need to assume $E[a_i|W_{it}] = 0$.

9. The null is no correlation between the a_i and the Xs. Thus, under the null both random effects and fixed effects are consistent. Given we fail to reject the null, statistically speaking, we can use the random effects. However, note on the totwrk coefficient there is a difference between the FE and RE estimate. So it could be potentially the imprecision the other FE estimates that is causing us to fail to reject the null. (This is a low power problem.)

10. As a general rule, in this type of analysis you should always run all the above models - comparing across models gives you a lot of information about the variation in your data. In this case there is little to choose between POLS, clustering and RE. Thus, the choice becomes a choice between FE or RE. Here even though the Hausman test does not reject the null, I would argue in favour of using fixed effects as the most reliable estimate, as the standard errors are in the same region as the random effects, plus more importantly, the estimator allows for correlation between unobserved heterogeneity and working.