Problem Set 1

Tao Wu

September 15th, 2023

1. **Define the following terms in your own words**

   (a) Panel data have observations on the same n entities over T time periods. Panel data, unlike cross-sectional data, are not just a snapshot of the entities. It provides variation over cross-sections and over time.

   (b) Balanced panel data, as opposed to unbalanced panel data, have no missing data entries. Any missing observation on any entity for one time period will give us unbalanced panel data.

   (c) Fixed effects regression model is used to control for omitted variables in panel data when the omitted variables vary across entities but do not change over time.

   $$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 Z_i + u_{it}$$

   (d) Entity fixed effects, a technique to control for time-invariant characteristics, that are unobserved, of each entity. These characteristics could vary among entities but remain unchanged over time.

   (e) Time fixed effects, a technique to control for time-variant factors, that are unobserved and affect all entities simultaneously. It is used to account for variation in the outcome variable only pertaining to different time periods.

   (f) Entity and time fixed effects regression model where we combine the entity fixed effect with time fixed effect

   $$Y_{it} = \beta_1 X_{it} + \alpha_i + \lambda_t + u_{it}$$

   (g) Heteroskedasticity and autocorrelation are common violations of multiple assumptions of ordinary least squares (OLS) regression. Heteroskedasticity and autocorrelation robust standard errors are what we use to account for these violations and still be able to correctly perform regression analysis while working with panel data.These adjusted standard errors are more reliable for conducting hypothesis tests and constructing confidence intervals for the coefficients.

(h) In random effects models, entity fixed effect α_i and time fixed effect λ_t can be considered independently and identically distributed random variables, independent of u_{it}. In prior fixed effect models, they are thought of as unobserved coorelated with the independent variables.

1. **What are the advantages of using panel data models?**

   (a) Panel data allows us to estimate more efficiently by controlling for hidden factors that cause omitted variable biases.

   (b) Panel data provides more variation to the data source, thus giving us less collinearity among the variables and more accurate estimation.

   (c) Panel data makes it easier to study dynamic models where time-lagged variables are involved.

1. **50 points Exercise**

   (a) The data set is not balanced. When you look at the data table, you will notice missing data entries or observations for many entities across different time periods.

   (b) i. The minimum value of dem_ind in the data set is 0. The maximum is 1. The mean is 0.499 and its standard deviation is 0.371.

   ii. The value of dem_ind for the U.S. in 2000 is 1. The mean of all values of dem_ind for the U.S. is 0.9855556.

   (c) If per capita income in a country increases by 20%, dem_ind is predicted to increase by 0.0472.

```
Estimaiton Results
================================================
                        Dependent variable:
                    ---------------------------
                               dem_ind
------------------------------------------------
log_gdppc                     0.236***
                               (0.007)

Constant                      -1.355***
                               (0.061)

------------------------------------------------
Observations                     958
R2                              0.438
Adjusted R2                     0.438
Residual Std. Error      0.272 (df = 956)
F Statistic          746.477*** (df = 1; 956)
================================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

   (d) Estimate the regression in (c), allowing for country-fixed effects, where α_i is the country-fixed effect and u_it is the idiosyncratic error term. My answer to (c) has changed as follows: If per capita income in a country increases by 20%, dem_ind is predicted to increase by 0.0168. The parameter estimate β_1 has changed in response to the

entity fixed effect where the omitted variable bias in β_1 , caused by unobserved and correlated factors, is reduced by this effect.

$$\texttt{dem\_ind}_{it} = \beta_0 + \beta_1 \times \texttt{log\_gdppc}_{it} + \alpha_i + u_{it}$$

```
Estimaiton Results
=======================================
                  Dependent variable:
              ---------------------------
                         dem_ind
---------------------------------------
log_gdppc              0.084***
                        (0.031)

---------------------------------------
Observations              958
R2                       0.020
Adjusted R2              -0.163
F Statistic    16.202*** (df = 1; 807)
=======================================
Note:         *p<0.1; **p<0.05; ***p<0.01
```

(e) Estimate the regression in (c), allowing for time and country fixed effects. My answer to (c) has changed as follows: If per capita income in a country increases by 20%, dem_ind is predicted to increase by 0.0108. The parameter estimate β_1 has changed in response to the entity-fixed effect and time-fixed effect where the omitted variable bias in β_1 , caused by unobserved and correlated factors that are both entity-invariant and time-invariant, is reduced by these effects. Moreover, Model 3, which includes country-fixed and year-fixed effects, suggests that the upward-biased coefficient in Model 1 is the result of omitted variable bias (the coefficient on $\log_g dppc$ is 0.236). One thing I cannot explain is that the coefficient in Model 3 is estimated to be statistically insignificant.

$$\texttt{dem\_ind}_{it} = \beta_0 + \beta_1 \times \texttt{log\_gdppc}_{it} + \alpha_i + \lambda_t + u_{it}$$

```
Estimaiton Results
=======================================
                 Dependent variable:
               --------------------------
                         dem_ind
---------------------------------------
log_gdppc                 0.054
                         (0.042)

---------------------------------------
Observations               958
R2                        0.005
Adjusted R2              -0.192
F Statistic     3.729* (df = 1; 799)
=======================================
Note:         *p<0.1; **p<0.05; ***p<0.01
```

Estimation Results

	Dependent variable:		
	(1)	(2)	(3)
log_gdppc	0.236***	0.084***	0.054
	(0.007)	(0.031)	(0.042)
Constant	-1.355***		
	(0.061)		
Observations	958	958	958
R2	0.438	0.020	0.005
Adjusted R2	0.438	-0.163	-0.192
Note:	*p<0.1; **p<0.05; ***p<0.01		