Homework assignment #6
Panel Data Analysis

MPP-C6: Statistics 2

Prof. Jan C. Minx
minx@hertie-school.org
http://moodle.hertie-school.org/course/view.php?id=1192

29 October 2015

Disclaimer: This document sketches some brief responses to the questions of the homework assignment. It serves to provide students with some guidance. The document may include som flaws - those should be reported to me as students go through the responses.

Project Description

The Environmental Kuznets Curve is at the heart of a long-standing discourse on the relationship between economic development and environmental quality. It hypothesizes an inverted U-shape relationship between indicators of environmental degradation and per capita income. While some have used the EKC to argue that growth policies are also superior for dealing with environmental problems, others have questioned the existence of EKCs for different indicators or stressed very high turning points. We aim to reproduce the results of Stern and Common (2001) which sought to investigate the presence of an environmental Kuznets curve (EKC) for sulfur emissions [1].

Dataset

The dataset stern2.dat contains country data from 1960-1990. The dataset contains the following variables

- *year*: the year in which the country was observed
- *country*: numerical code that uniquely identifies each country (see table 1)
- *gdpppp*: GDP per capita (purchasing power parity) in real 1990 international dollars
- *pop*: population in 1000 residents

- *so*: SO_2 emissions in tonnes
- *sopc*: SO_2 emissions in tonnes per capita
- *oe*: dummy variable describing oecd membership where 1000 represents membership and 2000 represents non-membership

Questions

1. **Read the paper by Stern and Common. Explain the EKC hypothesis in your own words. What is the difference between sulfur and carbon emissions in the empirical discussion and why? What is the authors' perceived contribution to the discussion? What is the role of panel data therein? [conceptual question]**

   The EKC hypothesis estimates that an inverted U-shape describes the relationship between income per capita and measures of environmental degradation. In other words, at low levels of income, increases in income will be associated with increases in environmental degradation, but as income rises, the effect will become smaller until a turning point, beyond which increases in income per capita will lead to decreases in environmental degradation.

   It was thought that carbon had a high turning point and sulfur had a low turning point, because the externalities of carbon emissions are global, whereas the externalities of sulfur emissions are local.

   The authors believe that previous estimates of an EKC for sulfur underestimate the turning point by relying on a sample of higher income countries. Panel data helps to control for time and country effects.

2. **Start by examining your data. Try out some descriptive statistics of the xt command family and report relevant ones. What sort of distribution do our variables of interest display? What transformations could we apply to the data? If necessary, create new variables that are appropriately transformed.**

Interactive Stata Example

```
. /*load the data (delimiters can be either tab or space or a combination,
> collapse tells stata to treat a combination of delimiters as one delimiter
> */
. import delimited ../data/stern2.dat, delimiters("\t ",collapse)
(8 vars, 2294 obs)

.
. * uncomment to remove kuwait
. * drop if country==98
.
. *set up panel
. xtset country year
       panel variable:  country (strongly balanced)
        time variable:  year, 1960 to 1990
                delta:  1 unit

.
. xtdescribe

 country:  1, 14, ..., 147                                    n =          74
```

```
    year:  1960, 1961, ..., 1990                              T =         31
           Delta(year) = 1 unit
           Span(year)  = 31 periods
           (country*year uniquely identifies each observation)

Distribution of T_i:   min      5%      25%       50%       75%     95%     max
                        31      31       31        31        31      31      31

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+-------------------------------
       74    100.00  100.00 |  1111111111111111111111111111111
 ---------------------------+-------------------------------
       74    100.00         |  XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX

. xtsum sopc gdpppp

Variable         |      Mean   Std. Dev.       Min        Max |    Observations
-----------------+--------------------------------------------+----------------
sopc     overall |  .0215023    .0366821   8.90e-07   .4655517 |     N =    2294
         between |              .0337633   .0000997   .2302244 |     n =      74
         within  |              .0148502  -.0996998   .2568297 |     T =      31
                 |                                             |
gdpppp   overall |  5359.908    6244.168        303   80830.76 |     N =    2294
         between |              5443.544   449.3548    38849.2 |     n =      74
         within  |              3121.721  -27160.64   47341.46 |     T =      31

.
. *histogram gdp and sopc (do we need to transform them
. hist gdpppp, normal kdensity
(bin=33, start=303, width=2440.2351)

. graph export hist_gdp.png, replace
(file hist_gdp.png written in PNG format)

. hist sopc, normal kdensity
(bin=33, start=8.900e-07, width=.0141076)

. graph export hist_sopc.png, replace
(file hist_sopc.png written in PNG format)

.
.
. *create new transformed variables and squared term
. cap gen lgdp = log(gdpppp)

. cap gen lsopc = log(sopc)

.
```

Both variables are right-skewed (see figures 1 and 2), so a log transformation is appropriate.

3. **Plot GDP per capita against sulfur emissions per capita (transformed if necessary). Describe the relationship you can see.**

Interactive Stata Example

```
. *plot lgdp and lsopc
. twoway (scatter lsopc lgdp)

. graph export lsopc_lgdp.png, replace
(file lsopc_lgdp.png written in PNG format)

.
```

At low levels of income, a positive effect of GDP per capita on sulphur emissions per capita is clearly visible. There seems to be some sign of a turning point, but it is not unambiguous (see figure 3).

4. **Write the equation for a model that could estimate an EKC for sulfur emissions. Create any extra variables that would be necessary to run this.**

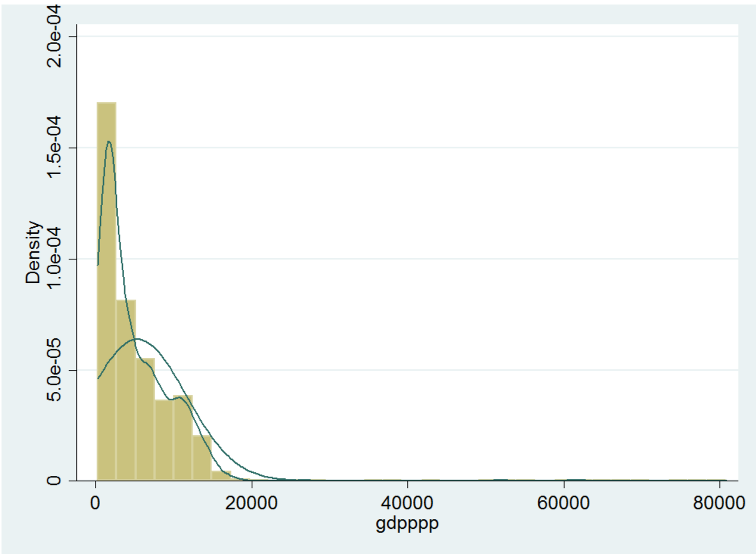

Figure 1: GDP distribution

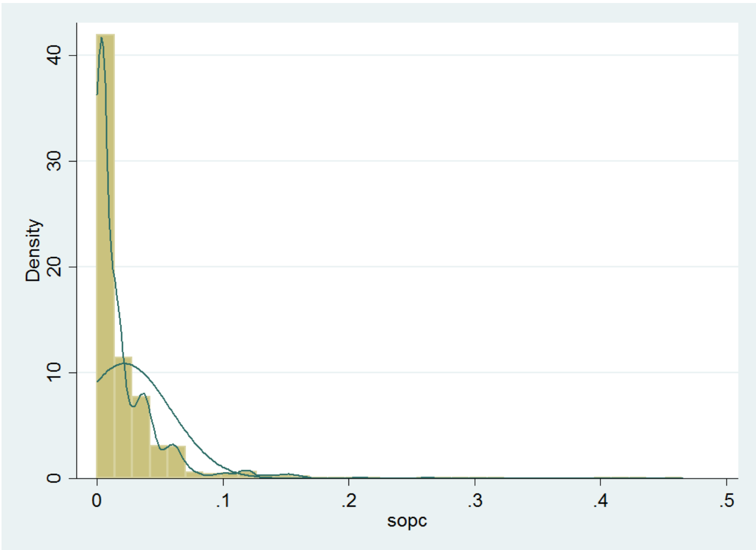

Figure 2: Sopc distribution

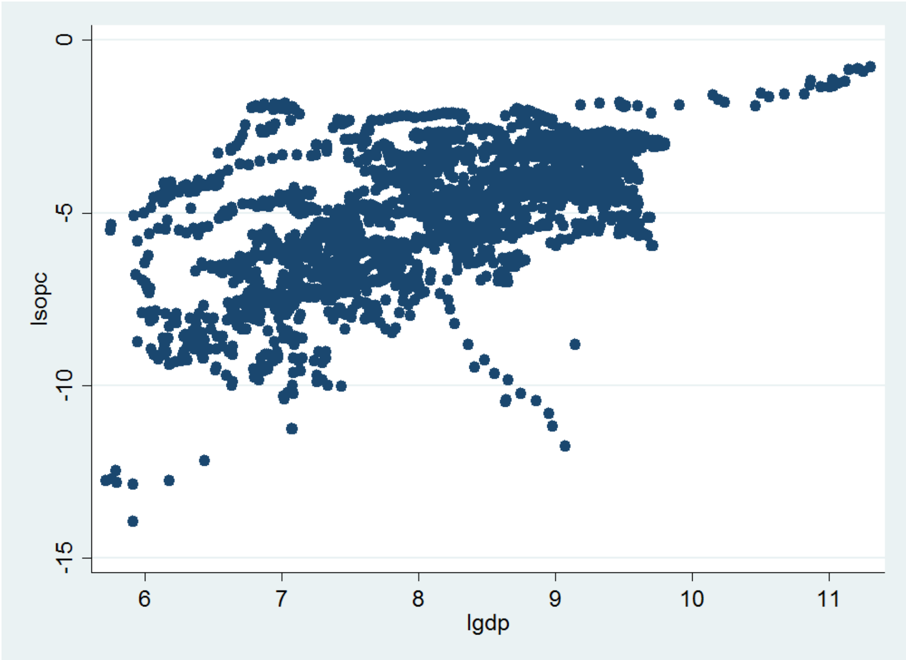

Figure 3: log GDP per capita against log sulfur emissions per capita

$$\ln SO_i = \beta_0 + \beta_1 lgdp + \beta_2 lgdp^2$$

5. **Carry out a pooled regression using the equation described in question 3. Interpret the coefficients. Run fixed-effects and random-effects models and interpret the results.**

Interactive Stata Example

```
. *create squared term
. cap gen lgdpsq = lgdp*lgdp

. *regress using pooled ols
. reg lsopc lgdp lgdpsq

      Source |       SS           df       MS      Number of obs   =     2,294
-------------+----------------------------------   F(2, 2291)      =    679.91
       Model |  3019.46587         2  1509.73294   Prob > F        =    0.0000
    Residual |  5087.10815     2,291  2.22047497   R-squared       =    0.3725
-------------+----------------------------------   Adj R-squared   =    0.3719
       Total |  8106.57402     2,293  3.53535718   Root MSE        =    1.4901

------------------------------------------------------------------------------
       lsopc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |    1.82093   .4380234     4.16   0.000     .9619664    2.679894
      lgdpsq |  -.0417549   .0271007    -1.54   0.124    -.0948993    .0113895
       _cons |  -17.02817   1.752482    -9.72   0.000    -20.46478   -13.59155
------------------------------------------------------------------------------

. est store pooled

.
. *random effects regression
. xtreg lsopc lgdp lgdpsq, re

Random-effects GLS regression                   Number of obs     =      2,294
```

```
Group variable: country                         Number of groups  =         74

R-sq:                                           Obs per group:
     within  = 0.1481                                         min =         31
     between = 0.3912                                         avg =       31.0
     overall = 0.3618                                         max =         31

                                                Wald chi2(2)      =     427.81
corr(u_i, X)   = 0 (assumed)                    Prob > chi2       =     0.0000

------------------------------------------------------------------------------
       lsopc |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |   3.266315   .3497687     9.34   0.000     2.580781    3.951849
      lgdpsq |  -.1520899    .021113    -7.20   0.000    -.1934705   -.1107092
       _cons |  -21.37886   1.453844   -14.71   0.000    -24.22834   -18.52938
-------------+----------------------------------------------------------------
     sigma_u |  1.3905606
     sigma_e |  .58383059
         rho |  .85014041   (fraction of variance due to u_i)
------------------------------------------------------------------------------

. est store ran

.
. *fixed effects regression
. xtreg lsopc lgdp lgdpsq, fe

Fixed-effects (within) regression               Number of obs     =      2,294
Group variable: country                         Number of groups  =         74

R-sq:                                           Obs per group:
     within  = 0.1481                                         min =         31
     between = 0.3903                                         avg =       31.0
     overall = 0.3611                                         max =         31

                                                F(2,2218)         =     192.87
corr(u_i, Xb)  = 0.2231                         Prob > F          =     0.0000

------------------------------------------------------------------------------
       lsopc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |   3.252834   .3533359     9.21   0.000      2.55993    3.945737
      lgdpsq |  -.1525319   .0212964    -7.16   0.000    -.1942948   -.1107689
       _cons |  -21.23976   1.462225   -14.53   0.000    -24.10723   -18.37228
-------------+----------------------------------------------------------------
     sigma_u |  1.4342211
     sigma_e |  .58383059
         rho |  .85784832   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0: F(73, 2218) = 174.06                  Prob > F = 0.0000

. est store fix

.
```

A Pooled regression finds a positive, significant coefficient for the linear GDP term. The squared GDP term is negative but insignificant. We would estimate the turning point of the EKC as

$$\tau = \exp\left(\frac{-\beta_1}{2 \cdot \beta_2}\right) = 2949748538$$

This is far outside our sample and an unreasonable value for GDP per capita.

Both fixed effects models and random effects models find larger effects for both the linear term and the squared term, and the squared term is significant in each. In the case of the fixed effects model the turning point is estimated as 42736.09, the turning point in the random effects model is estimated as 46078.71.

6. **Test which of the three models is preferable. Perform other relevant diagnostics for the model of choice. Is it appropriate to include time-fixed effects in the model?**

Interactive Stata Example

```
. *conduct a Breusch-Pagan test for heteroscedasticity
. quietly reg lsopc lgdp lgdpsq

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of lsopc

         chi2(1)      =   303.50
         Prob > chi2  =   0.0000

.
. *conduct a hausman test
. hausman fix ran

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      fix          ran         Difference          S.E.
-------------+----------------------------------------------------------------
        lgdp |    3.252834     3.266315       -.0134814         .050081
      lgdpsq |   -.1525319    -.1520899        -.000442         .002789
------------------------------------------------------------------------------
                          b = consistent under Ho and Ha; obtained from xtreg
            B = inconsistent under Ha, efficient under Ho; obtained from xtreg

    Test:  Ho:  difference in coefficients not systematic

                  chi2(2) = (b-B)'[(V_b-V_B)^(-1)](b-B)
                          =        6.27
                Prob>chi2 =      0.0435

.
.
```

A Breusch-Pagan test conducted on the pooled OLS model informs us that we should reject the null hypothesis that the model does not suffer from heteroskedasticity. This means that we should consider fixed effects and random effects models.

A Hausman test compares the coefficients of the fixed and random effects models. It finds that the difference between them is significant. A significant difference indicates that the random effects model is esti- mated inconsistently, due to correlation between the explanatory variables and the error components. We should therefore prefer a fixed-effects model.

7. **What is heterogeneity bias and is it relevant according to your results? [conceptual question]**

   Heterogeneity bias comes from ignoring the individual or time-specific effects that exist among cross-sectional or time-series units but are not captured by the included explanatory variables. If the characteristics of individual countries in our dataset that are not captured by our explanatory variables have an effect on sulphur emissions, then a pooled OLS reression will be biased.

8. **A high Hausman statistic implies that there is correlation between country effects and income variables. What could be the most likely cause of this problem? [conceptual question]**

   This could be caused by country attributes that both make countries richer and affect their sulphur emissions.

9. **Compute the relevant turning points of the estimated curves for the world, OECD and non-OECD regressions. Summarize your results in a table.**

Interactive Stata Example

```
. eststo ran_world, title("World"): quietly xtreg lsopc lgdp lgdpsq, fe

. estadd scalar e_tp = exp(-_b[lgdp]/(2*_b[lgdpsq]))

added scalar:
               e(e_tp) =  42736.155

.
. eststo ran_oecd, title("OECD"): quietly xtreg lsopc lgdp lgdpsq if oe==1000, fe

. estadd scalar e_tp = exp(-_b[lgdp]/(2*_b[lgdpsq]))

added scalar:
               e(e_tp) =  9147.0899

.
. eststo ran_non_oecd, title("Non-OECD"): quietly xtreg lsopc lgdp lgdpsq if oe==2000, fe

. estadd scalar e_tp = exp(-_b[lgdp]/(2*_b[lgdpsq]))

added scalar:
               e(e_tp) =  257479.18

.
. esttab ran_world ran_oecd ran_non_oecd, stats(e_tp) mtitle

------------------------------------------------------------
                      (1)             (2)             (3)
                    World            OECD        Non-OECD
------------------------------------------------------------
lgdp                3.253***        12.24***        2.410***
                   (9.21)         (16.76)          (5.47)

lgdpsq             -0.153***       -0.671***      -0.0967***
                  (-7.16)        (-16.28)         (-3.57)

_cons              -21.24***       -59.69***       -18.25***
                 (-14.53)        (-18.46)        (-10.23)
------------------------------------------------------------
e_tp              42736.2          9147.1        257479.2
------------------------------------------------------------
t statistics in parentheses
* p<0.05, ** p<0.01, *** p<0.001

.
```

10. **Discuss why first-differencing may be a more appropriate method for the data.**

    Differencing reduces the serial correlation and removes the country effects that are related to the specification problems in the levels model. If there are omitted integrated variables the first differences estimator is consistent.

11. **Estimate the model for the “world” using first-differences and interpret the results.**

Interactive Stata Example

```
. *First difference
. reg D.lsopc D.lgdp D.lgdpsq, noconstant

      Source |       SS           df       MS      Number of obs   =     2,220
-------------+----------------------------------   F(2, 2218)      =     23.67
       Model |  3.52188216         2  1.76094108   Prob > F        =    0.0000
    Residual |  165.031008     2,218  .074405324   R-squared       =    0.0209
-------------+----------------------------------   Adj R-squared   =    0.0200
```

```
       Total |   168.55289     2,220  .075924725   Root MSE        =    .27277

------------------------------------------------------------------------------
     D.lsopc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        lgdp |
         D1. |   2.081773   .7397481     2.81   0.005      .631102    3.532445
             |
      lgdpsq |
         D1. |  -.0917739   .0461154    -1.99   0.047    -.1822078     -.00134
------------------------------------------------------------------------------

. est store FD

.
```

A first differencing model shows a smaller coefficient for both the linear and squared gdp terms, though both are significant. The turning point is estimated as 84276.43

12. **Comment on any differences between the models you have run.**

    Pooled OLS found no significant turning point for an EKC for sulfur emissions.

    Fixed and random effects models estimated similar turning points at around 43000 and 46000 respectively, though we preferred fixed-effects based on a high Hausman statistic. Further analysis showed that the fixed-effects estimate was highly dependent on the sample. Excluding or including OECD and non-OECD countries changed the turning point estimate significantly.

    A first differencing model, which may be preferable, estimated a much higher turning point close to 84000.

13. **Discuss whether we can observe an EKC for sulfur emissions with reference to your results.**

    We can conclude that a single global EKC model is misspecified. The estimation of the turning point is heavily dependent on the countries in the sample. However, fixed-effects, random-effects and first-difference models all find a significant turning point for sulfur emissions, though for first-difference model the turning point is higher than the maximum GDP value in the sample.

References

[1] David I Stern and Michael S Common. Is there an environmental kuznets curve for sulfur? *Journal of Environmental Economics and Management*, 41(2):162–178, 2001.

Table 1: Country Codes

1	ALGERIA	95	JAPAN
14	EGYPT	97	KOREA,
18	GHANA	98	KUWAIT
22	KENYA	100	MALAYSIA
25	MADAGASCAR	102	MYANMAR
30	MOROCCO	106	PHILIPPINES
31	MOZAMBIQUE	108	SAUDI ARABIA
32	NAMIBIA	109	SINGAPORE
34	NIGERIA	110	SRI LANKA
41	SAFRICA	111	SYRIA
44	TANZANIA	112	TAIWAN
46	TUNISIA	113	THAILAND
48	ZAIRE	116	AUSTRIA
49	ZAMBIA	117	BELGIUM
50	ZIMBABWE	119	CYPRUS
52	BARBADOS	120	CZECHOSLOVAKIA
54	CANADA	121	DENMARK
60	GUATEMALA	122	FINLAND
62	HONDURAS	123	FRANCE
64	MEXICO	125	WGERMANY
65	NICARAGUA	126	GREECE
71	TRINIDAD&TOBAGO	129	IRELAND
72	U.S.A.	130	ITALY
73	ARGENTINA	131	LUXEMBOURG
74	BOLIVIA	133	NETHERLANDS
75	BRAZIL	134	NORWAY
76	CHILE	136	PORTUGAL
77	COLOMBIA	137	ROMANIA
81	PERU	138	SPAIN
83	URUGUAY	139	SWEDEN
84	VENEZUELA	140	SWITZERLAND
88	CHINA	141	TURKEY
89	HONG KONG	142	U.K.
90	INDIA	143	USSR
91	INDONESIA	144	YUGOSLAVIA
92	IRAN	145	AUSTRALIA
94	ISRAEL	147	NZ