Panel Data Models

Contents

1 Introduction

This workshop will take a practical look at panel data methods and models. Political scientists often approach research questions which have important time and space implications. Policy makers might face choices and questions of a similar nature: what effect does the building of a polluting power plant have on house prices? Is education a helpful tool to reduce the gender wage gap?

Panel data are useful to test hypotheses in these contexts, but come with new methodological and substantive challenges. Together we will look at these topics:

- Panel data, what does it look like? How do we describe our samples?
- Adapting Ordinary Least Squares to estimate models using panel data (especially focusing on the challenges of controlling for unobserved heterogeneity, errors, and lagged variables)

- Implementation of the models in R
- Substantive interpretation of model results from a practical policy-making point of view.

Last Updated: May 25, 2017 10:47 AM

2 Acknowledgments

The content of this workshop is based on the material from Introduction to Quantitative Methods course at UCL.

3 Panel Data Models

- Panel data are datasets in which a set of units (for example people) are observed for several time periods.
- If experimental data are not available, then the use of panel data is one important approach to reduce the problem of omitted variable bias.
- There are many empirical research areas where results that are not based on panel data are no longer taken seriously.

3.1 Fixed Effects Model

- The fixed effects model is simply a variation on the linear regression model.
- Its key advantage is that it enables us to control for all variables that vary over the cross-sectional units but are constant over time.

3.1.1 Example: Traffic Fatalities and Beer Tax

Dataset from Stock & Watson (Ch.10), covers state traffic fatality data available for 48 states observed over seven years (from 1982 to 1988), for a total of 336 observations.

state	year	mrall	beertax	mlda	jaild	vmiles	unrate	perinc
AL	1982	0.0002128	1.5393795	19.00	0	7233.887	14.4	10544.15
AL	1983	0.0002348	1.7889907	19.00	0	7836.348	13.7	10732.80
AL	1984	0.0002336	1.7142856	19.00	0	8262.990	11.1	11108.79
AL	1985	0.0002193	1.6525424	19.67	0	8726.917	8.9	11332.63
AL	1986	0.0002669	1.6099070	21.00	0	8952.854	9.8	11661.51
AL	1987	0.0002719	1.5599999	21.00	0	9166.302	7.8	11944.00
AL	1988	0.0002494	1.5014436	21.00	0	9674.323	7.2	12368.62
AZ	1982	0.0002499	0.2147971	19.00	1	6810.157	9.9	12309.07
AZ	1983	0.0002267	0.2064220	19.00	1	6587.495	9.1	12693.81
AZ	1984	0.0002829	0.2967033	19.00	1	6709.970	5.0	13265.93

3.2 Assumptions of the Fixed Effects Model

- The fixed effects model assumes that the true relationship is:

  $$y_{i,t} = \beta_0 + \beta_1 x_{i,t} + \beta_2 z_i + u_{i,t}$$

  where in the S&W example $y_{i,t}$ would be the number of traffic fatalities and $x_{i,t}$ the beer tax in state i in year t.
- Note that the variable z_i does not have a time index and is therefore assumed to be constant over time.
- In this example z_i could be the social attitude towards drunk driving in state i.
- If we define $\alpha_i = \beta_0 + \beta_2 z_i$, then (1) simplifies to

  $$y_{i,t} = \alpha_i + \beta_1 x_{i,t} + u_{i,t}$$
- The graphical interpretation of α_i is that it is the intercept of the relationship between alcohol taxes and traffic fatalities in state i.
- It is straightforward to allow for further variables which are constant over time in (1).
- In this case the intercepts α_i reflect the combined effect of several variables which are constant over time.

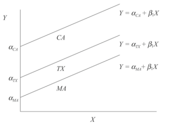
Y
$Y = \alpha_{CA} + \beta_1 X$
CA
α_{CA}
$Y = \alpha_{TX} + \beta_1 X$
$Y = \alpha_{MA} + \beta_1 X$
TX
α_{TX}
MA
α_{MA}
X

Figure 1:

3.3 Advantages and Disadvantages

- The key advantage of the fixed effects model is that it allows us to control for all time invariant omitted variables.
- This is particularly important in the case of variables which are difficult or impossible to observe.
- The key disadvantage is that we have to estimate a number of additional parameters.
- Furthermore, it will be impossible to estimate the effect of variables which do not (or hardly) vary over time.

3.4 Time Fixed Effects

- The basic fixed effects model only prevents omitted variable bias from variables that do not change over time.
- However, panel data allow us to control also for omitted variable bias from one other type of omitted variable.
- In the traffic fatalities example technical progress could be an important determinant of the number of deaths and could also be correlated with alcohol taxes.
- At the same time this variable probably affects all states in the same way (i.e. does not vary across states).

3.5 Time and Unit Fixed Effects

- In most applications we use both unit and time fixed effects at the same time.
- This model is sometimes referred to as the "twoway fixed effects" model.
- In the literature the cross-sectional fixed effects are referred to as "fixed effects", "state (fixed) effects", "firm (fixed) effects" or "person (fixed) effects".
- Similarly, time fixed effects are often referred to as "time effects".

4 Violations of Assumptions

4.1 Assumption 2 in Panel Data

- Panel data is characterized by time dependency for each panel unit.
- As discussed in Week 6, this is a violation of the regression Assumption 2 (X and Y are i.i.d).

4.1.1 Serial Correlation

- Time dependency is often described as autocorrelation or serial correlation.
- The main approach to deal with serial correlation is by adjusting standard errors to take into account autocorrelation.
- If there is substantial autocorrelation (serial correlation) in the error term, even heteroskedasticity-robust standard errors will be inconsistent.
- In panel data as in any other time series data, autocorrelation can be a very serious concern.
- We can test for serial correlation after our fixed effects estimation using the Breusch-Godfrey test.
- The null hypothesis in this test is that the autocorrelation of the error term is 0.

4.1.2 Cross-sectional dependence

- Cross-sectional dependence in panels may arise when e.g. countries respond to common shocks or if spatial diffusion processes are present (think Arab Spring, or shocks from the financial crisis).
- If cross-sectional dependence is present, this results, at least, in the inefficiency of the estimators and invalid inference when using standard estimation techniques.
- This is another instance of the violation of regression Assumption 2.
- If we assume that our earlier two-way fixed effects model specification is consistent, then we can test for residual cross-sectional dependence after the introduction of two-way fixed effects to account for common shocks.

4.2 Panel-corrected Standard Errors

- Panel-corrected standard errors (Beck and Katz 1995)
  - **panel heteroskedasticity**: each country may have its own error variance
  - **contemporaneous correlation of the errors**: the error for one country may be correlated with the errors for other countries in the same year
  - **serially correlated errors**: the errors for a given country are correlated with previous errors for that country

4.3 General Approach to Correlation Between Panels

- Driscoll and Kraay (1998) propose an estimator producing heteroskedasticity- and autocorrelation-consistent standard errors that are robust to general forms of spatial and temporal dependence. Often known as the SCC estimator.
- Panel Corrected Standard Errors (PCSE), while popular in political science, may not work well with shorter panels with large N (ratio of T/N is small).
- SCC estimator performs equally well in large N settings.

4.4 Your Roadmap with Panel Data

- Is it a fixed effects or random effects model?
- Hausman test. But primarily the choice should be driven by theory!
- Use robust standard errors, start with HAC.
- Check whether there is any cross-sectional dependence:
- If not, you can stick to HAC.
- If you have cross-sectional dependence, you need to use PCSE or SCC (use SCC).

5 Hands-on Tutorial

Let's start by load these packages:

```r
library(plm)
library(lmtest)
library(texreg)
```

Clear the environment

```r
rm(list = ls())
```

5.1 More Guns, Less Crimes

Download the guns dataset used by Stock and Watson.

- **Dataset**
- **Codebook**

Gun rights advocate John Lott argues in his book More Guns, Less Crimes that crime rates in the United States decrease when gun ownership restrictions are relaxed. The data used in Lott's research compares violent crimes, robberies, and murders across 50 states to determine whether the so called "shall" laws that remove discretion from license granting authorities actually decrease crime rates. So far 41 states have passed these "shall" laws where a person applying for a licence to carry a concealed weapon doesn't have to provide justification or "good cause" for requiring a concealed weapon permit.

Let's load the dataset used by Lott and see if we can test the arguments made by gun rights advocates.

```
guns <- read.csv("guns.csv")
```

The variables we're interested in are described below.

Variable	Definition
mur	Murder rate (incidents per 100,000)
shall	= 1 if the state has a shall-carry law in effect in that year = 0 otherwise
incarc_rate	Incarceration rate in the state in the previous year (sentenced prisoners per 100,000 residents; value for the previous year)
pm1029	Percent of state population that is male, ages 10 to 29
stateid	ID number of states (Alabama = 1, Alaska = 2, etc.)
year	Year (1977-1999)

We will focus on murder rates in this example but you could try the same with variables measuring violent crimes or robberies as well.

Let's create a factor variable representing whether a state has passed "shall" law or not. The variable already exists as `0` or `1` but we want to convert it to a factor for our analysis.

```
guns$shall <- factor(guns$shall, levels = c(0, 1), labels =c("NO", "YES"))
```

5.2 Fixed Effects

Let's estimate a fixed effect model on panel data using the `plm()` function with `shall`, `incarc_rate`, and `pm1029` as the independent variables.

```
state_effects <- plm(mur ~ shall + incarc_rate + pm1029,
                     data = guns,
                     index = c("stateid", "year"),
                     model = "within",
                     effect = "individual")

summary(state_effects)
```

```
## Oneway (individual) effect Within Model
```

```
##
## Call:
## plm(formula = mur ~ shall + incarc_rate + pm1029, data = guns,
##     effect = "individual", model = "within", index = c("stateid",
##         "year"))
##
## Balanced Panel: n=51, T=23, N=1173
##
## Residuals :
##       Min.    1st Qu.     Median    3rd Qu.       Max.
## -21.102428  -0.958945   0.016047   1.082008  29.031961
##
## Coefficients :
##               Estimate Std. Error t-value  Pr(>|t|)
## shallYES    -1.4513886  0.3154300 -4.6013 4.678e-06 ***
## incarc_rate  0.0174551  0.0011261 15.4998 < 2.2e-16 ***
## pm1029       0.9582993  0.0859610 11.1481 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    12016
## Residual Sum of Squares: 9800
## R-Squared:      0.18444
## Adj. R-Squared: 0.14581
## F-statistic: 84.3526 on 3 and 1119 DF, p-value: < 2.22e-16
```

The `state_effects` model shows that all three of our independent variables are statistically significant, with `shall` decreasing murder rates by 1.45 incidents per 100000 members of the population. The effects of incarceration rate and percentage of male population between 10 and 29 years old are also statistically significant.

Before drawing any conclusions let's make sure whether there are any state effects in our model using `plmtest()`.

```
plmtest(state_effects, effect="individual")
```

```
##
##  Lagrange Multiplier Test - (Honda) for balanced panels
##
## data:  mur ~ shall + incarc_rate + pm1029
## normal = 47.242, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

The p-value suggests the presence of state effects. In addition to state fixed effects, a number of factors could affect the murder rate that are not specific to an individual state. We can model these time fixed effects using the `effect = "time"` argument in `plm()`.

```
time_effects <- plm(mur ~ shall + incarc_rate + pm1029,
                    data = guns,
                    index = c("stateid", "year"),
                    model = "within",
                    effect = "time")

summary(time_effects)
```

```
## Oneway (time) effect Within Model
##
## Call:
## plm(formula = mur ~ shall + incarc_rate + pm1029, data = guns,
##     effect = "time", model = "within", index = c("stateid", "year"))
##
## Balanced Panel: n=51, T=23, N=1173
##
## Residuals :
##      Min.   1st Qu.    Median   3rd Qu.      Max.
## -21.68350  -2.04596  -0.31955   1.76758  35.20084
##
## Coefficients :
##              Estimate Std. Error t-value Pr(>|t|)
## shallYES    -0.2521605  0.3032163 -0.8316   0.4058
## incarc_rate  0.0412157  0.0007704 53.4993   <2e-16 ***
## pm1029       0.2148597  0.1407613  1.5264   0.1272
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    65571
## Residual Sum of Squares: 18141
## R-Squared:      0.72334
## Adj. R-Squared: 0.71731
## F-statistic: 999.623 on 3 and 1147 DF, p-value: < 2.22e-16
```

The `incarc_rate` variable is the only statistically significant variable in the time fixed effects model.

Now let's run `plmtest` on the `time_effects` model to verify if time fixed effects are indeed present in the model.

```
plmtest(time_effects, effect="time")
```

```
##
##  Lagrange Multiplier Test - time effects (Honda) for balanced
##  panels
##
## data:  mur ~ shall + incarc_rate + pm1029
## normal = 16.104, p-value < 2.2e-16
## alternative hypothesis: significant effects
```

The *p-value* tells us that we can reject the null hypothesis so we know that there are time fixed effects present in our model.

We already confirmed the presence of state fixed effects in the first model we estimated. Now, in order to control for both state AND time fixed effects, we need to estimate a model using the `effect = "twoways"` argument.

```
twoway_effects <- plm(mur ~ shall + incarc_rate + pm1029,
                      data = guns,
                      index = c("stateid", "year"),
                      model = "within",
                      effect = "twoways")

summary(twoway_effects)
```

```
## Twoways effects Within Model
##
## Call:
## plm(formula = mur ~ shall + incarc_rate + pm1029, data = guns,
##     effect = "twoways", model = "within", index = c("stateid",
##         "year"))
##
## Balanced Panel: n=51, T=23, N=1173
##
## Residuals :
##        Min.     1st Qu.      Median     3rd Qu.        Max.
## -19.2097691  -0.9748749  -0.0069663   1.0119176  27.1354552
##
## Coefficients :
##               Estimate Std. Error t-value  Pr(>|t|)
## shallYES    -0.5640474  0.3325054 -1.6964 0.0901023 .
## incarc_rate  0.0209756  0.0011252 18.6411 < 2.2e-16 ***
## pm1029       0.7326357  0.2189770  3.3457 0.0008485 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Total Sum of Squares:    11263
## Residual Sum of Squares: 8519.4
## R-Squared:      0.24357
## Adj. R-Squared: 0.19186
## F-statistic: 117.746 on 3 and 1097 DF, p-value: < 2.22e-16
```

In a twoway fixed effects model `shall` is no longer significant and the effect of male population between `10` and `29` years old has decreased from `0.95` to `0.73` incidents per `100,000` population.

The results of all three models are shown below.

```r
screenreg(list(state_effects, time_effects, twoway_effects),
          custom.model.names = c("State Fixed Effects",
                                 "Time Fixed Effects",
                                 "Twoway Fixed Effects"))
```

```
##
## ==========================================================================
##              State Fixed Effects  Time Fixed Effects  Twoway Fixed Effects
## --------------------------------------------------------------------------
## shallYES       -1.45 ***            -0.25                -0.56
##                (0.32)               (0.30)               (0.33)
## incarc_rate     0.02 ***             0.04 ***             0.02 ***
##                (0.00)               (0.00)               (0.00)
## pm1029          0.96 ***             0.21                 0.73 ***
##                (0.09)               (0.14)               (0.22)
## --------------------------------------------------------------------------
## R^2             0.18                 0.72                 0.24
## Adj. R^2        0.15                 0.72                 0.19
## Num. obs.    1173                 1173                 1173
## ==========================================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

5.3 Serial Correlation

For time series data we need to address the potential for serial correlation in the error term. We will test for serial correlation with Breusch-Godfrey test using `pbgtest()` and provide solutions for correcting it if necessary.

```
pbgtest(twoway_effects)
```

```
##
##  Breusch-Godfrey/Wooldridge test for serial correlation in panel
##  models
##
## data:  mur ~ shall + incarc_rate + pm1029
## chisq = 765.16, df = 23, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

The null hypothesis for the Breusch-Godfrey test is that there is no serial correlation. The `p-value` from the test tells us that we can reject the null hypothesis and confirms the presence of serial correlation in our error term.

We can correct for serial correlation using `coeftest()` similar to how we corrected for heteroskedastic errors. We'll use the `vcovHC()` function for obtaining a heteroskedasticity-consistent covariance matrix, but since we're interested in correcting for autocorrelation as well, we will specify `method = "arellano"` which corrects for both heteroskedasticity and autocorrelation.

```
twoway_effects_hac <- coeftest(twoway_effects,
                               vcov = vcovHC(twoway_effects,
                                             method = "arellano",
                                             type = "HC3"))

screenreg(list(twoway_effects, twoway_effects_hac),
          custom.model.names = c("Twoway Fixed Effects",
                                 "Twoway Fixed Effects (HAC)"))
```

```
##
## ==============================================================
##                Twoway Fixed Effects  Twoway Fixed Effects (HAC)
## --------------------------------------------------------------
## shallYES         -0.56                 -0.56
##                  (0.33)                (0.48)
## incarc_rate       0.02 ***              0.02 *
##                  (0.00)                (0.01)
## pm1029            0.73 ***              0.73
##                  (0.22)                (0.54)
## --------------------------------------------------------------
## R^2               0.24
## Adj. R^2          0.19
## Num. obs.      1173
## ==============================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

We can see that with heteroskedasticity and autocorrelation consistent (HAC) standard errors, the percent of male population (10 - 29 yr old) is no longer a significant predictor in our model.

5.4 Cross-sectional Dependence

If a federal law imposed restrictions on gun ownership or licensing requirements then the changes would likely affect all 50 states. This is an example of Cross Sectional Dependence and not accounted for in a fixed effect model. Other scenarios could also trigger cross sectional dependence that we should take into consideration. For example, security policies and law enforcement efforts might change after an extraordinary event (think of mass shootings or terrorist attacks) thus influencing law enforcement practices in all states. We can check for cross sectional dependence using the Pesaran cross sectional dependence test or `pcdtest()`.

```
pcdtest(twoway_effects)
```

```
##
##  Pesaran CD test for cross-sectional dependence in panels
##
## data:  mur ~ shall + incarc_rate + pm1029
## z = 3.9121, p-value = 9.148e-05
## alternative hypothesis: cross-sectional dependence
```

As we've seen with other tests, the null hypothesis is that there is no cross sectional dependence. The p-value, however tells that there is indeed cross-sectional dependence and we need to correct it. There are two general approaches to correcting for cross sectional dependence.

Beck and Katz (1995) method or Panel Corrected Standard Errors (PCSE): We can obtain Panel Corrected Standard Errors (PCSE) by first obtaining a robust variance-covariance matrix for panel models with the Beck and Katz (1995) method using the `vcovBK()` and passing it to the familiar `coeftest()` function.

```
twoway_effects_pcse <- coeftest(twoway_effects,
                                vcov = vcovBK(twoway_effects,
                                              type="HC3",
                                              cluster = "group"))

twoway_effects_pcse
```

```
##
## t test of coefficients:
##
##               Estimate Std. Error t value  Pr(>|t|)
## shallYES    -0.5640474  0.7662556 -0.7361    0.4618
## incarc_rate  0.0209756  0.0028249  7.4253 2.254e-13 ***
## pm1029       0.7326357  0.5118496  1.4313    0.1526
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The results from PCSE are sensitive to the ratio between the number of time periods in the dataset (T) and the total number of observations (N). When we're dealing with large datasets (i.e. the T/N ratio is small), we use the Driscoll and Kraay method instead.

Driscoll and Kraay (1998) (SCC): The cross-sectional and serial correlation (SCC) method by Driscoll and Kraay addresses the limitations of Beck and Katz's PCSE method is therefore preferred for obtaining heteroskedasticity and autocorrelation consistent errors that are also robust to cross-sectional dependence. We can get SCC corrected covariance matrix using the `vcovSCC()` function.

```
twoway_effects_scc <- coeftest(twoway_effects,
                               vcov = vcovSCC(twoway_effects,
                                              type="HC3",
                                              cluster = "group"))

twoway_effects_scc
```

```
##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## shallYES    -0.564047   0.542698 -1.0393  0.29888
## incarc_rate  0.020976   0.010321  2.0324  0.04236 *
## pm1029       0.732636   0.551066  1.3295  0.18396
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
screenreg(list(state_effects,
               time_effects,
               twoway_effects,
               twoway_effects_pcse,
               twoway_effects_scc),
          custom.model.names = c("State Effects",
                                 "Time Effects",
                                 "Twoway Fixed Effects",
                                 "PCSE",
                                 "SCC"))
```

```
##
## ==================================================================================
##              State Effects  Time Effects  Twoway Fixed Effects  PCSE       SCC
## ----------------------------------------------------------------------------------
## shallYES       -1.45 ***      -0.25         -0.56               -0.56      -0.56
##                (0.32)         (0.30)        (0.33)              (0.77)     (0.54)
## incarc_rate     0.02 ***       0.04 ***      0.02 ***            0.02 ***   0.02 *
##                (0.00)         (0.00)        (0.00)              (0.00)     (0.01)
## pm1029          0.96 ***       0.21          0.73 ***            0.73       0.73
##                (0.09)         (0.14)        (0.22)              (0.51)     (0.55)
## ----------------------------------------------------------------------------------
## R^2             0.18           0.72          0.24
## Adj. R^2        0.15           0.72          0.19
## Num. obs.    1173           1173          1173
## ==================================================================================
## *** p < 0.001, ** p < 0.01, * p < 0.05
```

6 Exercises

Download the traffic fatality dataset used in Stock and Watson examples. It covers data from 48 states observed over seven years (from 1982 to 1988), for a total of 336 observations.

- **Dataset**

Variable	Description
mrall	Vehicle Fatality Rate (VFR)
beertax	Tax on Case of Beer
mlda	Minimum Legal Drinking Age
jaild	Mandatory Jail Sentence
vmiles	Average Mile per Driver
unrate	Unemployment Rate
perinc	Per Capita Personal Income

1. Estimate a model for Vehicle Fatality Rate using your choice of variables listed above.
    - Estimate a fixed effect model and test for state and time fixed effects.
    - Run the necessary tests to check whether state and time fixed effects are present.
2. Estimate a twoway model and compare to the previous state and time fixed effect models.
3. Test for serial correlation and cross sectional dependence in the twoway model.
4. If either serial correlation or cross sectional dependence is present, use the methods you've learned to obtain heteroskedastic and autocorrelation consistent standard errors.
5. Compare the HAC and spatially robust standard errors with the twoway model estimated earlier.
6. Display the results in publication-ready tables and discuss the substantively significant findings.