Unit 12 Live Session

Analysis of Panel Data: Fixed Effect and Random Effect Models

Figure 1: South Hall

Class Announcements

- Lab-3 is due in 2 weeks

Roadmap

Last week:

- Introduction to Panel Data

This Week:

- Fixed Effects and Random Effects

Next Week:

- Mixed Effects Models that Combine Fixed and Random Effects

Start-up Code

```
if(!"plm"%in%rownames(installed.packages())) {install.packages("plm")}
library(plm)

if(!"plyr"%in%rownames(installed.packages())) {install.packages("plyr")}
library(plyr)

if(!"dplyr"%in%rownames(installed.packages())) {install.packages("dplyr")}
library(dplyr)

if(!"ggplot2"%in%rownames(installed.packages())) {install.packages("ggplot2")}
library(ggplot2)

if(!"ggthemes"%in%rownames(installed.packages())) {install.packages("ggthemes")}
library(ggthemes)

if(!"scales"%in%rownames(installed.packages())) {install.packages("scales")}
library(scales)

if(!"reshape2"%in%rownames(installed.packages())) {install.packages("reshape2")}
library(reshape2)

if(!"gridExtra"%in%rownames(installed.packages())) {install.packages("gridExtra")}
library(gridExtra)

if(!"lubridate"%in%rownames(installed.packages())) {install.packages("lubridate")}
library(lubridate)

if(!"stargazer"%in%rownames(installed.packages())) {install.packages("stargazer")}
library(stargazer)

if(!"mgcv"%in%rownames(installed.packages())) {install.packages("mgcv")}
library(mgcv)
```

Review of Panel Data

Recall that panel data has both cross sectional and time series dimensions. This means that we have multiple units such as individuals, countries, etc. and also repeated observations of those units over time. Typically we refer to the number of units as N and number of time periods as T so our data set has dimensions NxT.

Review of Panel Data Models

We often express a regression model for panel data as:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_p x_{pit} + \gamma_i + \eta_t + \epsilon_{it}$$

The idea is that we do not have NT independent observations and so cannot fit a regular OLS model. The observations have a dependence / correlation structure over time for each individual.

We want our model to account for this more complicated correlation structure to produce valid coefficient estimates, statistical tests, etc.

To account for this structure we often make sure to have fixed effect coefficients included i.e. dummy variables for each unit (γ_i). This means our regression equation has different intercepts for each unit.

These fixed effects adjust for the average effect of each unit, removing unobserved heterogeneity at the individual level from the error term and avoiding omitted variable bias.

We also sometimes include fixed effects for time periods η_t if the data allows it such as different intercepts for each year to control for group invariant but unobserved time trends.

In theory this leaves our error term as random noise, recovering the basic OLS assumptions that we need for valid analysis.

Note this means that our other regressors must be at a different level of variation than the fixed effects. Otherwise they cannot be estimated. For example, if we have observations on units for 10 years and include group fixed effects, we can't also include time invariant individual level regressors like place of birth.

The Pooling Estimator

Recall that the pooled estimator ignores any of the panel data structure and just runs basic OLS:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_p x_{pit} + \epsilon_{it}$$

The Between Estimator

In the between estimator, we average our units across time periods in the data to remove any time invariant confounders:

$$\frac{1}{T}\Sigma_t^T y_{it} = \frac{1}{T}\Sigma_t^T(\beta_0 + \beta_1 x_{1it} + ... + \beta_p x_{pit} + \gamma_i + \epsilon_{it})$$

This simplifies to the following:

$$\bar{y}_i = \beta_0 + \beta_1 \bar{x}_{1i} + ... + \beta_p \bar{x}_{pi} + \gamma_i + \bar{\epsilon}_i)$$

Now the gamma terms are redundant because our data is at the level of each group. If we include them, we will be unable to estimate the gamma coefficients.

The Within Estimator (Fixed Effects)

In the within estimator, we first demean the variables to remove group averages and then run our regression, which eliminates the fixed effect coefficients γ_i:

$$(y_{it} - \bar{y}_i) = \beta_1(x_{1it} - \bar{x}_{1i}) + ... + \beta_p(x_{pit} - \bar{x}_{pi}) + (\epsilon_{it} - \bar{\epsilon}_i)$$

This will produce equivalent results to running a regression with the fixed effects per group included as dummy variables, but it can be faster to run things this way when there are many groups.

The First Difference Model

In the first difference model, we take differences in variables to remove the time invariant group effects γ_i:

$$(y_{it} - y_{it-1}) = \beta_1(x_{1it} - x_{1it-1}) + ... + \beta_p(x_{pit} - x_{pit-1}) + (\epsilon_{it} - \epsilon_{it-1})$$

$$\Delta y_{it} = \beta_1 \Delta x_{1it} + ... + \beta_p \Delta x_{pit} + \Delta \epsilon_{it}$$

Test to Compare the Within Model and First Difference Model

We can run the Wooldridge first difference statistical test to compare the efficiency of the fixed effects within model to the first difference model by analyzing the residuals from each model. Note that both models provide consistent estimates of the coefficients, but one can be more efficient.

If we use the fixed effects within model, we are assuming there is no serial correlation in the error term so that:

$$Cor(\epsilon_{it}, \epsilon_{it-s}) = 0$$

This means that residuals of the first difference model are correlated because:

$$Cor(\Delta\epsilon_{it}, \Delta\epsilon_{it-1}) = \frac{Cov(\epsilon_{it} - \epsilon_{it-1}, \epsilon_{it-1} - \epsilon_{it-2})}{SD(\epsilon_{it} - \epsilon_{it-1})SD(\epsilon_{it-1} - \epsilon_{it-2})} = \frac{Cov(-\epsilon_{it-1}, \epsilon_{it-1})}{\sqrt{2\sigma_\epsilon^2}\sqrt{2\sigma_\epsilon^2}} = -\frac{\sigma_\epsilon^2}{2\sigma_\epsilon^2} = -0.5$$

Given this, we can test the following model for $H0 : \delta = -0.5$ (since residuals have zero mean) to see whether the residuals in the first difference model appear correlated:

$$\Delta\hat{\epsilon}_{it} = \delta\Delta\hat{\epsilon}_{it-1} + \xi_{it}$$

If we fail to reject the null hypothesis, then we can assume the within model is more efficient.

Alternatively, we could have a null hypothesis of $H0 : \delta = 0$ in the regression above, which would mean the first difference model is more valid as the differenced residuals are white noise and unrelated to each other. Failing to reject the null in this case means that the first difference model is better:

Because this test allows both hypotheses, we normally run both. In the case that either or both are rejected, it means the affected models have residuals that suffer from serial correlation, so we need to use autocorrelation robust standard errors.

Random Effects Models

Recall the general fixed effects model:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + ... + \beta_p x_{pit} + \gamma_i + \eta_t + \epsilon_{it}$$

If $Cov(\gamma_i, x_{kit}) = 0$, then the pooling model is consistent but will not be efficient because of variation in residuals across groups and serial correlation.

We can specify the true error term $\nu_{it} = \epsilon_{\gamma_i} + \epsilon_{it}$ as a sum of variation between groups and variation within groups in which case we have serial correlation in our residuals:

$$Cor(\nu_{it}, \nu_{is}) = \frac{Cov(\epsilon_{\gamma_i} + \epsilon_{it}, \epsilon_{\gamma_i} + \epsilon_{is})}{SD(\epsilon_{\gamma_i} + \epsilon_{it})SD(\epsilon_{\gamma_i} + \epsilon_{is})} = \frac{\sigma^2_{\gamma_i}}{\sigma^2_{\gamma_i} + \sigma^2_{\epsilon}}$$

Using a random effects model imposes this error structure on model residuals. This allows us to properly specify the residuals and more efficiently estimate the coefficient(s) of interest. It does require the strong assumption of independence between random effects and other predictors however.

The upside is we can now include group level predictors that are time invariant in the model along with random effects for each group.

Hausman Test for Fixed vs. Random Effects

We can run a statistical test to examine the assumption that residuals are uncorrelated with other predictors in the model in the case of random effects, which means there is no omitted variable bias from omitting fixed effects.

The null hypothesis is that the random effects model is acceptable while the alternative hypothesis is there is correlation between residuals and predictors, meaning we should use the fixed effects model.

If the random effects model is acceptable, then it is also more efficient then the fixed effects model because it specifies the error term and serial correlation correctly. Furthermore, both random effects and fixed effects are consistent estimators of the coefficients, so under the null the coefficients are equal:

$$H_0 : \beta_{RE} = \beta_{FE}$$

We derive a test statistic to examine this and test whether it is significantly different from 0. The test statistic comes from the fact that asymptotically both coefficients follow a joint normal distribution:

$$W = (\beta_{RE} - \beta_{FE})^T \hat{\Sigma}^{-1} (\beta_{RE} - \beta_{FE}) \sim \chi^2_p$$

For this test we need to find the variange of the difference i.e:

$$Var(\beta_{RE} - \beta_{FE}) = Var(\beta_{RE}) + Var(\beta_{FE}) - 2Cov(\beta_{FE}, \beta_{RE})$$

Note that since the covariance of an efficient estimator (random effects under the null) with its difference from an inefficient estimator (fixed effects under the null) is zero:

$$Cov(\beta_{RE}, \beta_{FE}) = Cov(\beta_{RE}, \beta_{FE} - \beta_{RE} + \beta_{RE}) = Cov(\beta_{RE}, \beta_{FE} - \beta_{RE}) + Var(\beta_{RE}) = Var(\beta_{RE})$$

The variance simplifies to the following, allowing us to calculate the test statistic:

$$Var(\beta_{RE} - \beta_{FE}) = Var(\beta_{RE}) + Var(\beta_{FE}) - 2Var(\beta_{RE}) = Var(\beta_{FE}) - Var(\beta_{RE})$$

The intuition behind this test statistic and the fact that the covariance between an efficient estimator and its difference from an inefficient estimator is zero is the following. Under the null hypothesis, random effects is the most efficient estimator while fixed effects is inefficient. If the covariance of a linear combination of random effects and fixed effects was not zero, then we could improve the efficiency of the random effects estimator by taking a linear combination with the fixed effects estimator:

$$Var(\beta_{RE} + \theta\beta_{FE}) = Var(\beta_{RE}) + \theta^2 Var(\beta_{FE}) + 2\theta Cov(\beta_{RE}, \beta_{FE})$$

Now pick a $\theta = -\frac{Cov(\beta_{RE}, \beta_{FE})}{Var(\beta_{FE})}$ then we can have the variance of our new estimator be less than the random effects one, meaning it is more efficient, which violates our assumption above.

$$Var(\beta_{RE} + \theta\beta_{FE}) = Var(\beta_{RE}) - \frac{Cov(\beta_{RE}, \beta_{FE})^2}{Var(\beta_{FE})} < Var(\beta_{RE})$$

Hence under the null hypothesis we have $Cov(\beta_{RE}, \beta_{FE} - \beta_{RE}) = 0$.

If we reject the null hypothesis, then the random effects model is not more efficient, and it is also not consistent. The fixed effects model is better in that case.

Case Study: Presidential Elections and State Unemployment

Let's try to answer the following question: does the unemployment rate affect whether voters vote Democratic or Republican in Presidential elections?

Panel Data Example

Let's view an example data set of how each state (and DC) voted in US presidential elections from 1980 to 2016.

Here we have observations of 51 units over 10 time periods (an election occurs every 4 years), so we have a balanced panel structure because each unit has the same number of observations over time.

```
dat<-read.csv("./data/fund_election_data.csv")
dat<-dat[as.Date(dat$date)<="2016-12-01",]
dat$year<-year(dat$date)

table(dat$state_name)
```

```
##
##              Alabama               Alaska              Arizona
##                   10                   10                   10
##             Arkansas           California             Colorado
##                   10                   10                   10
##          Connecticut             Delaware District of Columbia
##                   10                   10                   10
##              Florida              Georgia               Hawaii
##                   10                   10                   10
##                Idaho             Illinois              Indiana
##                   10                   10                   10
##                 Iowa               Kansas             Kentucky
##                   10                   10                   10
##            Louisiana                Maine             Maryland
##                   10                   10                   10
##        Massachusetts             Michigan            Minnesota
##                   10                   10                   10
##          Mississippi             Missouri              Montana
##                   10                   10                   10
##             Nebraska               Nevada        New Hampshire
##                   10                   10                   10
##           New Jersey           New Mexico             New York
##                   10                   10                   10
```

```
##       North Carolina         North Dakota                 Ohio
##                   10                   10                   10
##             Oklahoma               Oregon         Pennsylvania
##                   10                   10                   10
##         Rhode Island       South Carolina         South Dakota
##                   10                   10                   10
##            Tennessee                Texas                 Utah
##                   10                   10                   10
##              Vermont             Virginia           Washington
##                   10                   10                   10
##        West Virginia            Wisconsin              Wyoming
##                   10                   10                   10
```

```
table(dat$year)
```

```
##
## 1980 1984 1988 1992 1996 2000 2004 2008 2012 2016
##   51   51   51   51   51   51   51   51   51   51
```

Exploratory Analysis

Do states appear to differ in how they vote? Produce a visualization to answer this question.

What does this suggest about fixed effects our model?

```
# Your code here
```

Produce a visualization to answer if states changed their voting patterns over time.

```
# Your code here
```

Which variables seem to matter for whether a state votes Democratic or Republican? Produce a visualization to answer this question.

```
# Your code here
```

How correlated are the X variables with each other?

```
# Your code here
```

Fitting Panel Data Models

Using the plm function, fit a pooled OLS model, a between model, a within model, first difference model, and random effects model of democratic vote share and state unemployment rate.

In each model include the following controls: nat_gdp, age_65_plus, gender_female, race_white, educ_hs_or_less, density, voter_turnout, net_approval, and PVI.

Compare the results of each model. Which model do you think is better? Would you say presidential vote is impacted by unemployment rates?

Pooled OLS Model

```
# pool.model <- # Your code here

# summary(pool.model)
```

Within Model

```
# within.model <- # Your code here

# summary(within.model)
```

Test of Pooling

Conduct a test for pooling using the within and pooled models and `pFtest`. Do you reject the null hypothesis of no fixed effects?

```
# Your code here
```

Between Model

```
# between.model <- # Your code here

# summary(between.model)
```

First Difference Model

```
# fd.model <- # Your code here

# summary(fd.model)
```

Test for First Difference vs. Within Model

Use the `pwfdtest` to test which model is better for first difference vs. the within model. Run both versions of the null hypothesis. Which model is more efficient?

```
# Your code here
```

Random Effects Model

```
# re.model <- # Your code here

# summary(re.model)
```

Test for Random Effects

Conduct a Hausman test for random vs. fixed effects using `phtest`. Which model is more appropriate?

```
# Your code here
```

Model Comparison

```
# stargazer(pool.model, between.model, within.model, fd.model, re.model,
#           style="qje", type="text", keep="state_unemployment", omit.stat=c("adj.rsq","f"),
#           column.labels = c("Pooled OLS","Between","Within","FD","RE"))
```

Heteroskedasticity and Serial Correlation Robust Standard Errors

Review of Robust Standard Errors

If we have homoskedasticity, the variance of OLS β coefficients is:

$$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}$$

Remember from 203 that in the case of heteroskedasticity, the variance of β coefficients is:

$$Var(\hat{\beta}) = \sigma^2(X^TX)^{-1}X^T\Sigma X(X^TX)^{-1}$$

and we can estimate $\hat{\Sigma}$ to get robust standard errors using the following steps:

1. Estimate the OLS model
2. Calculate residuals from the OLS model
3. Square residuals and use these as an estimate of the variance of each observation (since residuals have a mean of zero)

This estimate of the covariance matrix provides more valid estimates of the variance of coefficients for better t tests, etc.

This is known as "white1" standard errors in R (HC parameters have to do with different weighting schemes and HC0 is perfectly fine).

Cluster Robust Standard Errors

In the case of panel data, we often have a grouping structure that we can use to impose additional structure on the covariance matrix when calculating robust standard errors that assumes the variance of observations in each group are the same.

We estimate them by averaging the squared residuals over time for each group in the panel data.

This is known as “white2” standard errors in R. Note these do not adjust for serial correlation.

Serially Correlated Standard Errors

There are two main options in R for handling serial correlation in residuals.

Arellano standard errors allow for both group heteroskedasticity and serial correlation by allowing correlation across time in groups.

The other option is Newey West standard errors that accommodate non zero correlation in the residuals through estimating off diagonal terms in the covariance matrix. These are not robust to heteroskedasticity in groups though, so it is recommended to use Arellano standard errors for full flexibility. These can however suffer with efficiency in small samples.

Testing for Heteroskedasticity in Residuals

Breusch Pagan Test

Suppose we run a fixed effects regression that is valid. The residuals from this regression should not depend on the X variables and in particular the group characteristics because we are assuming homoskedasticity.

Therefore none of the coefficients should be significant in the following regression of the squared, standardized residuals (so they have a variance of one) from our model:

$$(\epsilon_i^2/\hat{\sigma}^2) = X_i\beta + \nu_i$$

We can form a test statistic that compares the residual sum of squares in this regression to the total sum of squares of just using an intercept model (note the intercept is one in this case cause the residuals are standardized).

If the fit in the above regression is sufficienctly better than that from just using an intercept model, we reject the null hypothesis of homoskedasticity.

This statistic follows a chi squared distribution:

$$Test\ Statistic = \frac{(TSS - SSR)}{2} \sim \chi^2_{p-1}$$

Testing for Serial Correlation in Residuals

There are two common tests for serial correlation in regression models: the Durbin Watson test and the Breusch-Godfrey test.

Durbin Watson Test

We can estimate the autocorrelation coefficient in residuals and test if $H0 : \rho = 0$ which would imply no serial autocorrelation for a lag of 1:

$$\epsilon_{it} = \rho\epsilon_{it-1} + \xi_{it}$$

Rejecting the null hypothesis means there is autocorrelation in the residuals, and we should adjust our standard errors.

Breusch-Godfrey Test

We fit the following model on the estimated residuals for a specified number of p autocorrelation lags in the residuals:

$$\hat{\epsilon}_{it} = \beta_0 + \beta_1 X_{1it} + ... + \rho_1\hat{\epsilon}_{it-1} + ... + \rho_p\hat{\epsilon}_{it-p} + \xi_{it}$$

Under $H0 : \rho_i = 0 \; \forall i$ we have $nR^2 \sim \chi^2_p$, so we can use this to test the null hypothesis. Rejecting the null hypothesis means there is autocorrelation up to lag p in the residuals, and we should adjust our standard errors for this.

Case Study: Standard Error Comparison

Perform a Breusch Pagan test for heteroskedasticity in the within model using `pcdtest`. What do you conclude?

```
# Your code here
```

Perform a Durbin Watson test for serial correlation in the within model and first difference models using `pdwtest`.

```
# Your code here
```

Compare the results of the Durbin Watson test above for both the within and first difference models to those from the Breusch-Godfrey test with order=2 using `pbgtest`. Do you notice any differences and what does this mean?

```
# Your code here
```

Using `vcovHC` calculate robust standard errors (white1), cluster robust standard errors (white2), arrelano standard errors (arrellano), and newey west standard errors (using `vcovNW`) for the coefficient on state_unemployment in the within model.

```
# reg.se <- # Your code here

# het.se <- # Your code here
```

```r
# cluster.se <- # Your code here

# nw.se <- # Your code here

# arrellano.se <- # Your code here
```

Compare each type of standard error and discuss the differences.

```r
# data.frame(
#   "Type" = c("Regular OLS", "Robust", "Cluster Robut", "Newey West", "Arrellano"),
#   "SE" = c(reg.se, het.se, cluster.se, nw.se, arrellano.se)
# )
```

Comparing Models With Different Standard Errors

Let's compare the various models we fit against one another but add heteroskedastic and serial correlation robust standard errors for relevant models.

```
# pool.model.se<-sqrt(diag(vcovHC(pool.model, method="arellano", type="HC0")))
# within.model.se<-sqrt(diag(vcovHC(within.model, method="arellano", type="HC0")))
# fd.model.se<-sqrt(diag(vcovHC(fd.model, method="arellano", type="HC0")))
# re.model.se<-sqrt(diag(vcovHC(re.model, method="arellano", type="HC0")))

# stargazer(pool.model, between.model, within.model, fd.model, re.model,
#           se=list(pool.model.se,NULL,within.model.se,fd.model.se,re.model.se),
#           style="qje", type="text", keep="state_unemployment", omit.stat=c("adj.rsq","f"),
#           column.labels = c("Pooled OLS","Between","Within","FD","RE"))
```

Variable Coefficient Models

In the above analysis we have assumed that all coefficients β_i are constant over groups and time.

Alternatively, we could wonder whether the regression coefficients for one or more variables change for by group or over time.

There are many ways to estimate these "varying coefficient" models. We will highlight a couple common ones.

Separate Models

In the plm package, there are two approaches that estimate different regression models at the specified level.

One way is to assume that the coefficient varies randomly around a fixed mean for each group but is time invariant, and we estimate separate regressions for each group across time to get:

$$\beta_{ki} = \beta_k + \eta_i$$

We could also assume instead that the coefficient changes over time but is group invariant, and we estimate separate regressions for each time period across groups to get:

$$\beta_{kt} = \beta_k + \eta_t$$

We can estimate varying coefficients for both fixed effects and random effects models. Generally only the fixed effects varying coefficient model in plm is interesting because random effects only returns the overall coefficient average, which doesn't tell us anything about changes over time/groups (but does examine whether relaxing the varying coefficient assumption keeps results the same).

```
# #we need to use a smaller model because of the insufficient number of observations in each group / time period; we can take the stat sig
# fe.varying.coeff.model1<-pvcm(democratic_vote_share_adj ~ state_unemployment + nat_gdp +
#                                 age_65_plus + educ_hs_or_less + net_approval + PVI,
#                               data=dat, index=c("state_name","date"), model="within",
#                               effect="individual")
# fe.varying.coeff.model1
#
# fe.varying.coeff.model2<-pvcm(democratic_vote_share_adj ~ state_unemployment + nat_gdp +
#                                 age_65_plus + educ_hs_or_less + net_approval + PVI,
#                               data=dat, index=c("state_name","date"), model="within",
#                               effect="time")
# fe.varying.coeff.model2
```

GAMs

A more interesting / flexible way to estimate a varying coefficient model is GAMs. We can still combine information into a single regression model to estimate coefficients but also specify specific terms to vary smoothly over time using cubic regression splines (or some other smoother).

To do so we include a time trend but specify that we want the time trend to differ according to the values of a specific "by" variable, which is the "varying coefficient".

```
# gam.model<-gam(democratic_vote_share_adj ~ s(year, by=state_unemployment, bs="cr") +
#                   nat_gdp + age_65_plus + gender_female + race_white + educ_hs_or_less +
#                   voter_turnout + net_approval+PVI,
#                 data=dat)
# summary(gam.model)
#
# plot(gam.model)
```

Next Week

- Lab 3 is due next week
- Complete week 13 material (lectures + quiz) before next week