PS 406 – Week 8 Section: Panel Methods and Missing Data

D.J. Flynn

May 21, 2014

Today's plan

1. Panel Methods
   - Review
   - Panel models in R
2. Missing Data
   - Review
   - Multiple imputation in R

Recap of panel data

- N individuals, T time periods
- We assume there is correlation within each i over-time, but independence across i
- Types of regressors:
  - **Varying regressors:** Xs that vary across i and T (e.g., income)
  - **Time-invariant regressors:** Xs that vary across i but not T (e.g., race, gender):

$$X_{it} = X_i \, \forall i$$

  - **Individual-invariant regressors:** Xs that vary across T but not i (e.g., unemployment rate, time trends):

$$X_{it} = X_t \, \forall i$$

Panel data models

- **pooled OLS model**: OLS applied to panel data (so you'll end up with $N * T$ observations)

$$y_{it} = \alpha + X_{it}\beta + \epsilon_{it}$$

- **individual-specific effects models:** we assume heterogeneity across i, which we capture with α_i
  - **fixed effects model**: individual-specific effects correlated with regressors:

$$y_{it} = \alpha_i + X_{it}\beta + \epsilon_{it}$$

  - **random effects model**: individual-specific effects uncorrelated with regressors:

$$y_{it} = X_{it}\beta + (\alpha + \epsilon_{it})$$

- Deciding between these requires tests. Let's look at how to estimate panel models and test assumptions...

More on fixed vs. random effects[1]

Such models assist in controlling for unobserved heterogeneity when this heterogeneity is constant over time and correlated with independent variables...There are two common assumptions made about the individual specific effect, the random effects assumption and the fixed effects assumption. The random effects assumption (made in a random effects model) is that the individual specific effects are uncorrelated with the independent variables. The fixed effect assumption is that the individual specific effect is correlated with the independent variables. If the random effects assumption holds, the random effects model is more efficient than the fixed effects model. However, if this assumption does not hold (i.e., if the Durbin-Watson test fails), the random effects model is not consistent.

[1] Thanks Wikipedia.

Panel models in **R**[2]

```
#get lab data:
library(foreign)
data<- as.data.frame(dget(file="http://sekhon.berkeley.edu/gov2000/
R/agl1.dpt"))
names(data)

#Pooled OLS model:
pooled<-lm(y~left*imports, data=data)
summary(pooled)

#PCSEs to correct for contemporaneous correlation across
#means (e.g., exogenous change affects X_i and X_j at same time):
library(pcse)
pcses<-pcse(pooled, groupN=data$country, groupT=data$year)
summary(pcses)
```

[2] For more on PCSEs, see http://cran.r-project.org/web/packages/pcse/pcse.pdf.

The plm package

```
library(plm)
panel<-pdata.frame(data, index=c("country", "year"))

#Pooled OLS model (same as above but with plm):
pooled2<-plm(y~left*imports, data=panel, model="pooling")
summary(pooled2)
```

Fixed effects

```
#Fixed effects for country:
within.model<-plm(y~left*imports,data=panel, model="within")
summary(within.model)

#effects for each country:
summary(fixef(within.model))

#deviation from overall mean:
summary(fixef(within.model,type="dmean"))

#Fixed effects for time:
fe.time<-plm(y~left*imports,data=panel, model="within",
effect="time")
summary(fe.time)

#Fixed effects for time AND country:
fe.time.country<-plm(y~left*imports,data=panel, model="within",
effect="twoways")
```

Fixed effects, cont'd

```
summary(fe.time.country)

#if you include FE for time and country, you can look at the
#coefficients on each:
summary(fixef(fe.time.country,type="dfirst",effect="individual"))
summary(fixef(fe.time.country,type="dfirst",effect="time"))

#Testing whether pooling is OK: estimate a panel variable
#coefficient model (each unit has its own intercept/slope)
#and compare it to pooled model:
summary(pooled2)
pvcm<-pvcm(y~left+imports+ I(left*imports), data=panel,
model="within")
pooltest(pooled2,pvcm)
#low p-value=reject null of poolability (i.e., SHOULDN'T pool)
```

Random effects

- Recall: assumes unit/time effects are random variables drawn from a normal distribution. This is in contrast to fixed effects, which assumes that effects are correlated with regressors (characteristics of each *i*.)

```
random<-plm(y~left+imports+I(left*imports), data=panel,
model="random")
summary(random)

#RE for time:
random2<-plm(y~left+imports+I(left*imports), data=panel,
model="random", effect="time")
summary(random2)

#RE for units/time:
random3<-plm(y~left+imports+I(left*imports), data=panel,
model="random", effect="twoways")
summary(random3)
```

Random effects, cont'd

```
#Comparing fixed/random effects:
phtest(within.model,random)
#low p-value=one model is inconsistent, in which case we
#should go with fixed effects b/c it makes fewer assumptions
#(e.g., characteristics of countries uncorrelated with X)

#Testing for unit-specific omitted variables:
plmtest(pooled2, effect="individual")
#low p-value=omitted variables

#Testing for time-specific omitted variables:
plmtest(pooled2, effect="time")
#low p-value=omitted variables

#Checking for autocorrelation (have we minimized it?):
pdwtest(random)
pbgtest(random)
#low p-value=autocorrelation
```

SEs and CIs for the interaction effects

- You'll need to bootstrap the SEs and CIs. I'll send some sample code out by Friday.

Types of missing data

1. **MCAR**:[3] missing and non-missing cases are representative subsets of larger populations; can use listwise deletion:

$$Pr(R|Y, X) = Pr(R)$$

2. **MAR**:[4] missing and non-missing cases differ on some X, but not on the variable that has missing data. Note that this assumption is testable (regress R on Y_O and X).

$$Pr(R|Y, X) = Pr(R|Y_O, X)$$

3. **Non-Ignorable Missingness**:[5] probability of missingness depends on unobserved value:

$$Pr(R|Y, X) \neq Pr(R|Y_O, X)$$

[3] Example: People flip a coin to decide whether to participate in study.

[4] Example: Republicans less willing to fill out gov't housing survey, but missingness is independent of housing status.

[5] Example: Racial conservatives refuse to answer questions about racial attitudes.

The mi package[6]

```
#We'll use the nes02.csv data (on site):
names(nes02)

#Probit model of vote choice in 2002:
model<-lm(vote.gop~as.factor(pid)+interest+age+white+female,
data=nes02)
summary(model)

#Let's see how many NAs we have:
summary(as.factor(nes02$vote.gop))
summary(as.factor(nes02$pid))
summary(as.factor(nes02$interest))
summary(as.factor(nes02$age))
summary(as.factor(nes02$white))
summary(as.factor(nes02$female))
```

[6] For more on imputation, see http://cran.r-project.org/web/packages/mi/mi.pdf.

```
library(mi)

#tell R which variables to impute:
nes02.temp<-with(nes02, data.frame(vote.gop=vote.gop, pid=pid,
interest=interest, age=age, white=white, female=female))

#info about variables and any NAs:
nes02.temp.info<-mi.info(nes02.temp)
nes02.temp.info

#formulas R will use to impute missings:
nes02.temp.info$imp.formula

#run the imputation (I do just n.imp=3 here to save time)
#(this took me ~4 minutes):
nes02mi<-mi(nes02.temp, nes02.temp.info, n.imp=3, n.iter=5000)

#re-estimate model on imputed dataset (notice syntax):
model.imp<-lm.mi(vote.gop~as.factor(pid)+interest+age+white+female,
nes02mi)
```

```
#compare results across original model (with NAs) and
#imputed model:

summary(model)
display(model.imp)
```

Summing up *everything*....

Method/Estimator	When Would I Use This?
OLS regression	(Quasi-)continuous DV
Logit/probit	Binary DV
Complementary log-log	Binary DV (rare outcomes)
Ordered logit/probit	Ordinal DV
Poisson	Count data
Negative binomial	Count data (overdispersed)
Multinomial logit	Qualitative/categorical DV (IIA)
Multinomial probit	Qualitative/categorical DV (no IIA)
Instrumental variables/2SLS	Endogeneity: $E(X^*e) \neq 0$
Mediation analysis	To test causal mechanisms
Panel models	Anytime you have panel data
Bootstrapping	To find quantities that are difficult to calculate mathematically (e.g., small sample bias of MLE, SEs for interactions in MNP, etc.)
Multiple imputation	Anytime you have missing data that aren't MCAR
Matching	*Many, many* applications (e.g., finding similar cases, botched randomization, etc.)

Note: There are many other instances in which these methods might be appropriate. These are just the applications we focused on in class.