Panel Data

Mauricio Romero

Panel Data

Panel Data

Introduction

Fixed effects

Repeated observations

- Repeated observations
  - Panel data
  - Time-series cross section data
  - Clustered data, etc
- Dynamics effects (dynamic treatment regimes)
- Identification strategies
  - Fixed effects
  - Difference-in-differences

Group Data

- Panel: observe the same units (individuals, firms, countries, schools, etc.) over several time periods
- Time-series cross section: observe different units across time (e.g., different survey rounds of ENOE)
- Clustered data: Natural grouping in the data (e.g., test score data of students across schools)

Panel Data

Introduction

Fixed effects

Introduction

Fixed effects

Notation

- Sample of $i = 1, ..., N$ units from a population
- Time periods $t = 1, ..., T$
- For each i observe (Y_{it}, T_{it})
- The GDP is:

$$y_{it} = \beta_0 + \beta_1 T_{it} + \alpha_i + u_{it}$$

- Do **not** observe α_i
- Assume u_i is white noise (i.i.d. and mean zero)

OLS

- OLS estimator of Y on T yields

$$\begin{aligned}
\mathbb{E}\beta_1 &= \frac{cov(T_{it}, y_{it})}{V(T_{it})} \\
&= \frac{cov(T_{it}, \beta_0 + \beta_1 T_{it} + \alpha_i + u_{it})}{V(T_{it})} \\
&= \frac{cov(T_{it}, \beta_0)}{V(T_{it})} + \frac{cov(T_{it}, \beta_1 T_{it})}{V(T_{it})} + \frac{cov(T_{it}, \alpha_i)}{V(T_{it})} + \frac{cov(T_{it}, u_{it})}{V(T_{it})} \\
&= \underbrace{\frac{cov(T_{it}, \beta_0)}{V(T_{it})}}_{0} + \beta_1 \underbrace{\frac{cov(T_{it}, T_{it})}{V(T_{it})}}_{1} + \frac{cov(T_{it}, \alpha_i)}{V(T_{it})} + \underbrace{\frac{cov(T_{it}, u_{it})}{V(T_{it})}}_{0} \\
&= \beta_1 + \frac{cov(T_{it}, \alpha_i)}{V(T_{it})}
\end{aligned}$$

- Omitted variable bias

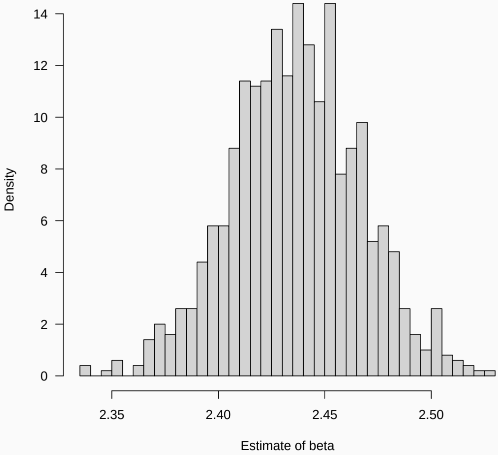
Simulations!

```
#TRUE MODEL
beta0=1 #intercept
beta1=2 #slope
Nobs=1000 #how many observations?
TimePeriods=5 #how many time periods?
unobserved=runif(Nobs,-1,1) #Create individual unobserved factor
Treatment=sample(c(0,1),size=Nobs*TimePeriods,replace=T) #T_{it}
Data=data.frame(IDObs=rep(1:Nobs,each=TimePeriods),
                Period=rep(1:TimePeriods,Nobs),
                alpha_i=rep(unobserved,each=TimePeriods),
                Treatment=Treatment)
#Let's say high unobservable means you always get treated
#Note this negates the random assignment
Data$Treatment[Data$alpha_i>0.5]=1
#Use GDP to generate data
Data$Outcome=beta0+beta1*Data$Treatment+Data$alpha_i+rnorm(Nobs*TimePeriods)
summary(lm(Outcome~Treatment,data=Data)) #seems to be biased
```

Simulations!

```
EstimateBeta=NULL
for(r in 1:1000){
  #Use GDP to generate data
  Data$Outcome=beta0+beta1*Data$Treatment+Data$alpha_i+rnorm(Nobs*TimePeriods)
  EstimateBeta=c(EstimateBeta,lm(Outcome~Treatment,data=Data)$coef[2])
}
hist(EstimateBeta,freq=F,breaks=30,
     main="",las=1,xlab="Estimate of beta")
abline(v=beta1,col='red',lwd=3,lty=1)
```

Our estimate of the coefficient are biased



Fixed effects (intuition I)

- Take one i at a time (like a subset for each i)

$$
\begin{aligned}
y_{i1} &= \beta_0 + \beta_1 T_{i1} + \alpha_i + u_{i1} \\
&\vdots \\
y_{iT} &= \beta_0 + \beta_1 T_{iT} + \alpha_i + u_{iT}
\end{aligned}
$$

- $\beta_0 + \alpha_i$ is the constant now
- We can estimate β_1 for each i ($\widehat{\beta_1}^i$)
- Aggregate for all i to get a more precise estimate of β_1
- Maximal precision weighting by $V(T_{it}|i)$ (why?)

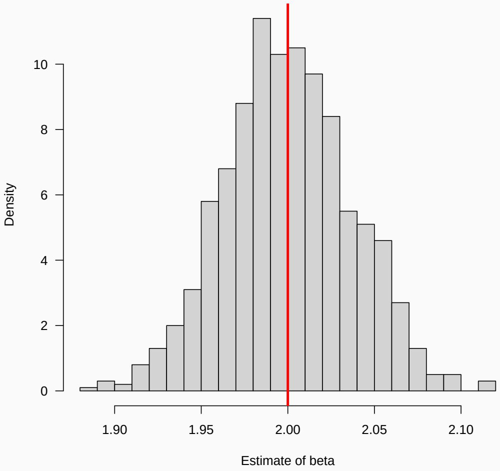
Simulations!

```
#TRUE MODEL
beta0=1 #intercept
beta1=2 #slope
Nobs=1000 #how many observations?
TimePeriods=5 #how many time periods?
unobserved=runif(Nobs,-1,1) #Create individual unobserved factor
Treatment=sample(c(0,1),size=Nobs*TimePeriods,replace=T) #T_{it}
Data=data.frame(IDObs=rep(1:Nobs,each=TimePeriods),
                Period=rep(1:TimePeriods,Nobs),
                alpha_i=rep(unobserved,each=TimePeriods),
                Treatment=Treatment)
#Let's say high unobservable means you always get treated
#Note this negates the random assignment
Data$Treatment[Data$alpha_i>0.5]=1
#Use GDP to generate data
Data$Outcome=beta0+beta1*Data$Treatment+Data$alpha_i+rnorm(Nobs*TimePeriods)
CoefIndividuals=NULL
VarianceTreatment=NULL
for(i in 1:Nobs){
  CoefIndividuals=c(CoefIndividuals,
                    lm(Outcome~Treatment,data=Data,subset=IDObs==i)$coef[2])
  VarianceTreatment=c(VarianceTreatment,
                      var(Data$Treatment[which(Data$IDObs==i)]))
}
mean(CoefIndividuals,na.rm=T)
weighted.mean(CoefIndividuals,w=VarianceTreatment,na.rm=T)
```

Simulations!

```
EstimateBeta=NULL
for(r in 1:1000){
  #Use GDP to generate data
  Data$Outcome=beta0+beta1*Data$Treatment+Data$alpha_i+rnorm(Nobs*TimePeriods)
  CoefIndividuals=NULL
  VarianceTreatment=NULL
    for(i in 1:Nobs){
    CoefIndividuals=c(CoefIndividuals,
                      lm.fit(y=Data$Outcome[which(Data$IDObs==i)],
                             x=as.matrix(cbind(1,Data$Treatment[which(Data$IDObs==i)])))$coefficients[2])
                      #lm(Outcome~Treatment,data=Data,subset=IDObs==i)$coef[2])
    VarianceTreatment=c(VarianceTreatment,
                        var(Data$Treatment[which(Data$IDObs==i)]))
    }
  EstimateBeta=c(EstimateBeta,
                 weighted.mean(CoefIndividuals,w=VarianceTreatment,na.rm=T))
}
hist(EstimateBeta,freq=F,breaks=30,
     main="",las=1,xlab="Estimate of beta")
abline(v=beta1,col='red',lwd=3,lty=1)
```

Our estimate of the coefficient are pretty close to the truth



Fixed effects (intuition II)

- Another idea, is to remove the mean for each observation

$$
\begin{aligned}
y_{it} &= \beta_0 + \beta_1 T_{it} + \alpha_i + u_{it} \\
y_{it} - \overline{Y_i} &= \beta_0 - \beta_0 + \beta_1 (T_{it} - \overline{T_i}) + \alpha_i - \alpha_i + u_{it} - \overline{u_i} \\
y_{it}^* &= \beta_1 T_{it}^* + u_{it}^*
\end{aligned}
$$

- Estimate via OLS

Simulations!

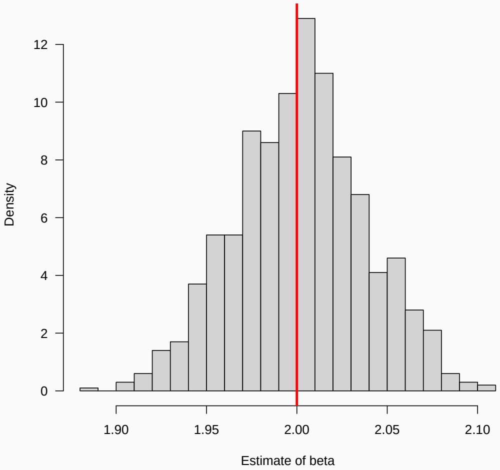
```
#TRUE MODEL
beta0=1 #intercept
beta1=2 #slope
Nobs=1000 #how many observations?
TimePeriods=5 #how many time periods?
unobserved=runif(Nobs,-1,1) #Create individual unobserved factor
Treatment=sample(c(0,1),size=Nobs*TimePeriods,replace=T) #T_{it}
Data=data.frame(IDObs=rep(1:Nobs,each=TimePeriods),
                Period=rep(1:TimePeriods,Nobs),
                alpha_i=rep(unobserved,each=TimePeriods),
                Treatment=Treatment)
#Let's say high unobservable means you always get treated
#Note this negates the random assignment
Data$Treatment[Data$alpha_i>0.5]=1
#Use GDP to generate data
Data$Outcome=beta0+beta1*Data$Treatment+Data$alpha_i+rnorm(Nobs*TimePeriods)
Data$MeanOutcome <- ave(Data$Outcome, Data$IDObs)
Data$MeanTreatment <- ave(Data$Treatment, Data$IDObs)
Data$OutcomeDemean=Data$Outcome-Data$MeanOutcome
Data$TreatmentDemean=Data$Treatment-Data$MeanTreatment
summary(lm(OutcomeDemean~TreatmentDemean,data=Data))
```

Simulations!

```
EstimateBeta=NULL
for(r in 1:1000){
  Data$Outcome=beta0+beta1*Data$Treatment+Data$alpha_i+rnorm(Nobs*TimePeriods)
  Data$MeanOutcome <- ave(Data$Outcome, Data$IDObs)
  Data$MeanTreatment <- ave(Data$Treatment, Data$IDObs)
  Data$OutcomeDemean=Data$Outcome-Data$MeanOutcome
  Data$TreatmentDemean=Data$Treatment-Data$MeanTreatment
  EstimateBeta=c(EstimateBeta,
                lm(OutcomeDemean~TreatmentDemean,data=Data)$coef[2]
               )
}

hist(EstimateBeta,freq=F,breaks=30,
     main="",las=1,xlab="Estimate of beta")
abline(v=beta1,col='red',lwd=3,lty=1)
```

Our estimate of the coefficient are pretty close to the truth



Fixed effects formally

$$y_{it} = \beta_0 + \beta_1 T_{it} + \sum_{j=1}^{N} 1_{i=j} + u_{it}$$

- $1_{i=j}$ is equal to one if $j = i$, zero otherwise — dummy for each group
- By the FWL theorem (decomposition theorem) this is equivalent to:
  - OLS of y_{it} with respect to $\sum_{j=1}^{N} 1_{i=j}$, and take the residuals ($\tilde{y}_{it}$)
  - OLS of T_{it} with respect to $\sum_{j=1}^{N} 1_{i=j}$, and take the residuals ($\tilde{T}_{it}$)
  - OLS of $\tilde{y}_{it}$ with respect to $\tilde{T}_{it}$
- OLS with respect to $\sum_{j=1}^{N} 1_{i=j}$ equivalent to subtracting group mean
- Same as the transformation we discussed in the previous slide

Fixed effects formally

$$y_{it} = \beta_0 + \beta_1 T_{it} + \sum_{j=1}^{N} 1_{i=j} + u_{it}$$

- By the FWL theorem:

$$\widehat{\beta_1} = \frac{\sum_i \widehat{\beta_1^i} V(T_{it}|i)P(i)}{\sum_i V(T_{it}|i)P(i)}$$

- Same as the estimate from doing OLS one i at a time

Simulations!

```r
beta0=1 #intercept
beta1=2 #slope
Nobs=1000 #how many observations?
TimePeriods=5 #how many time periods?
unobserved=runif(Nobs,-1,1) #Create individual unobserved factor
Treatment=sample(c(0,1),size=Nobs*TimePeriods,replace=T) #T_{it}
Data=data.frame(IDObs=rep(1:Nobs,each=TimePeriods),
                Period=rep(1:TimePeriods,Nobs),
                alpha_i=rep(unobserved,each=TimePeriods),
                Treatment=Treatment)
Data$Treatment[Data$alpha_i>0.5]=1
Data$Outcome=beta0+beta1*Data$Treatment+Data$alpha_i+rnorm(Nobs*TimePeriods)
Data$MeanOutcome <- ave(Data$Outcome, Data$IDObs)
Data$MeanTreatment <- ave(Data$Treatment, Data$IDObs)
Data$OutcomeDemean=Data$Outcome-Data$MeanOutcome
Data$TreatmentDemean=Data$Treatment-Data$MeanTreatment
lm(OutcomeDemean~TreatmentDemean,data=Data)$coef[2]
feols(Outcome~Treatment|IDObs,data=Data)$coefficients[1]
CoefIndividuals=NULL
VarianceTreatment=NULL
for(i in 1:Nobs){
  CoefIndividuals=c(CoefIndividuals,
                    lm(Outcome~Treatment,data=Data,subset=IDObs==i)$coef[2])
  VarianceTreatment=c(VarianceTreatment,
                      var(Data$Treatment[which(Data$IDObs==i)]))
}
weighted.mean(CoefIndividuals,w=VarianceTreatment,na.rm=T)
```

Fixed effects formally

- Thus, the following are algebraically equivalent:
    - Dummy variable OLS with $1_{i=j}$
    - Variance weighted average of coefficients from OLS for each i
    - OLS after demeaning (removing i-specific means)
- This is "one-way fixed effects" regression
- Addresses "time-invariant" confounders

Fixed effects visually (from Nick Huntington-Klein)

http://nickchk.com/anim/Animation%20of%20Fixed%20Effects.gif

Fixed effects and causality

- Y_{1it} and Y_{0it} are period-specific potential outcomes
- T_{it} is the treatment assigned to i in period t
- We observe

$$Y_{it} = T_{it} Y_{1it} + (1 - T_{it}) Y_{0it}$$

Fixed effects and causality

- Assumption 1: T_{it} is conditionally mean independent in any given period

$$\mathbb{E}[Y_{0it}|\alpha_i, X_{it}, T_{it}] = \mathbb{E}[Y_{0it}|\alpha_i, X_{it}]$$

- Assumption 2: Linearity

$$\mathbb{E}[Y_{0it}|\alpha_i, X_{it}] = \beta_0 + \alpha_i + X_{it}\gamma$$

- Assumption 3: Constant additive effects:

$$\mathbb{E}[Y_{1it}|\alpha_i, X_{it}] = \mathbb{E}[Y_{0it}|\alpha_i, X_{it}] + \beta_1$$

Under these assumptions, “one-way fixed effects” yields consistent estimator for β_1

Problems that fixed effects cannot solve

- Reverse causality
  - Assumption 1 implies potential outcomes uncorrelated with past, current and future treatment!
- Time-varying unobserved heterogeneity

Regressors that are constant within strata

- If a regressor is constant within an fixed-effect strata, then it is perfectly collinear with that strata dummy
  - A time-invariant regressor in the panel context
- When you fit fixed effects, these strata-invariant regressors must be dropped
- With fixed effects, what matters is whether the demean variables are constant

Clustering standard errors by fixed effect strata

- We cluster to account for dependencies in the treatment
- If treatments are assigned randomly within fixed effect strata (even if treatment probabilities differ across strata), no need to cluster by strata
- If treatment assignment at the strata level (or treatment exhibits positive or negative dependence within a strata), then cluster by strata
- “feols” from the “fixest” package easiest way in R to do fixed effects and cluster

Fixed effects estimators

- Typically we care about β, but unit fixed effects α_i could be of interest
  - $\widehat{\alpha}_i$ from dummy variable regression is unbiased but not consistent for α_i (based on fixed T and $N \to \infty$)

What I will not cover

- Huge literature on panel data
- This is not a review of panel econometrics; for that see Wooldridge and other excellent textbooks
- I won't be covering a lot of it
  - Random effects
  - First difference
  - Arrelano and Bond
  - And much more...
- Goal is to present the modal regression model used in difference-in-differences (next class)