# 1 About p-values

Unsignificant values don't permit us to draw the conclusion that there is no real effect. Looking at papers in detail is like looking at the backyard of a slaughterhouse. Only the p-value is almost never sufficient to draw meaningful conclusions on significance of a coefficient. At least we should take the effect size into account.

## 1.1 Panel data

Repeated observations of some individual unit over time. standard case: The same individual over the same unit of time. -> balanced panel

But "atrition" often leads to unbalanced panels. Notation: Notation for subscript

i : index for individual observations

t : index for time periods

$$X_i, t, ...$$

if N individuals for T time periods => sample size NT: balanced panel

everything else: unbalanced people

Further examples:

Rotationg panels (Socio-Economic panel SOEP)

Pseudo panels ( mean cohort values over time): often used for poverty research in developing countries. Advantage: can combine data, once the cohort is identified. Deaton (1985)

Why panel data?

- more observations => more information
- dynamic analysis:
- shocks over time average out
- **unobserved heterogeneity**

# 2 Notation:

Linear Regression

cross section

$$Y_i = \beta_0 + \beta_1 * X_1 i + u_i ...$$

panel structure

$$Y_{it} = \beta_0 + \beta_1 * X_{1it} + uit ...$$

$$Y_{it} = \beta_{0t} + \beta_1 * X_{1it} + uit ...$$

t = 1

$$Y_{i1} = \beta_{01} + \beta_1 * X_{1,1} + u_{i1}...$$

t = T

$$Y_{iT} = \beta_{0T} + \beta_{1T} * X_{1iT} + u_{iT}...$$

$$D_i = 1 if i = j, D_i = 0 if i != j for j = 1, ..., N$$

least-squares dummy variable estimator LSDV
with individual means over time.

$$(Y_{it} - \bar{Y}_{i0}) = \beta_1(X_{it} - x_{1i.}) + (u_{it} - \bar{u}_i)$$

how to get

$$\hat{\beta}_{0i}$$

?

$$\hat{\beta}_{0i} = \bar{Y}_{i0} - \beta$$

■=

library(plm)

setwd("C:/Users/jakob/OneDrive/University/Data$_analysis_Oct19/Panel_data")dataNL <
$-readRDS("dataNL.rds")names(dataNL) < -c("index","year","milk","other","x1","x2","x3","x4","x$

summary(dataNL)

dataNL$lmilk$ < $- - log(-dataNL$milk) dataNL$lx1$ < $-log(dataNL$x1) dataNL$lx2$ < $-log(dataNL$x2) dataNL$lx3$ < $-log(dataNL$x3) dataNL$lx4$ < $-log(dataNL$x4) dataNL$lx5$ < $-log(dataNL$x5)

@ The trend variable remains unlogged.

■= plot(dataNL$lmilk$ $dataNL$lx1) plot(dataNL$lmilk$ $dataNL$lx2) plot(dataNL$lmilk$ $dataNL$lx3) plot(dataNL$lmilk$ $dataNL$lx4) plot(dataNL$lmilk$ $dataNL$lx5)

formula.NL <- lmilk  lx1 + lx2 + lx3 + lx4 + lx5 + trend

lm.NL <- lm(formula.NL , data=dataNL)

summary(lm.NL)

Pool.NL <- plm(formula.NL, data = dataNL, model = "pooling")

summary(Pool.NL)

formula.LSDV <- lmilk  lx1 + lx2 + lx3 + lx4 + lx5 + trend + as.factor(index)
if we run that , index has  140 dummy variables we run into the problem of perfect multicollinearity. So R automatically drops one of the dummies.

lm.LSDV <- lm(formula.LSDV, data = dataNL)

summary((lm.LSDV))

@ In order to extract a coefficient, we use the coef() function ■=

coef(Pool.NL)[2:6] sum(coef(Pool.NL)[2:6])

@ output at 1.06 which is too high. Maybe we get different results with the LSDV estimator. ■=

coef(lm.LSDV)[2:6] sum(coef(lm.LSDV)[2:6])

@ now lower coefficient taking the index dummies into account.

■=

require(car)

linearHypothesis(Pool.NL , ”lx1+lx2+lx3+lx4+lx5=1”)

summary(Pool.NL) summary(lm(formula.NL, data = dataNL)) sum(coef(Pool.NL)[2:6])

WI.NL <- plm(formula.NL, data = dataNL, model = ”within”) cbind(coef(lm.LSDV[2:7], coef(WI.NL)))

@ About manually applying F-Tests : - unrestricted (ignoring H0) - RSS$^U R[residual sum of squares] restricte$ $RSS^R$

$$\star F = \frac{RSS^R - RSS^U R/ + 1}{RSS^U R/(NT - (k-1))}$$

Substract means from every variable.. Using loops (?)

Dummy variables you cannot meaningfully de-mean over time. So we use the LM, but should get out the same results as with the LSDV model. ■= wi2.NL <- plm(formula.NL, data = dataNL, effect = ”twoways”, model = ”within )

plot(density(fixef(WI.NL)))

@ Problem: time-invariant variables and how to deal with them.. ■= dataNL$TimeInvar < -runif(141) formula.TimeInvar < -lmilk + lx1 + lx2 + lx3 + lx4 + lx5 + trend + TimeInvar$

head(dataNL$TimeInvar) WI.NL < -plm(formula.NL, data = dataNL, model =$ ”$within$”)

@ Next steps: random effects model

# 3 scenario

No interest in the unobserved heterogeneity, no need to interpret the individual effects;

$$\alpha_i$$

- parameters are a mere cuisance (guidance?) –> error

$$Y_{i,t} = \alpha_i + \beta_1 * X_{1,i,t} + u_{i,t}$$

alpha is error

$$= \beta_0 + \beta_1 * X_{1,i,t} + \alpha_i + u_{i,t}$$

two error components alpha$_i, u_i t$

Ignore error structure: OLS $\rightarrow unbiased \rightarrow inefficient$

$$\alpha_i\ N(0, \sigma_\alpha^2) with u_{it}\ N(0, \sigma_u^2)$$

Estimating: Feasible Generalised Least Squares FGLS

$$E(Cov[X, u]) = 0$$

$$E(Cov[X, \alpha]) = 0$$

$\leftarrow inmanycontextsthisisacriticalassumptionItisoftenquestionablethatindividualeffectsandregressorsareu$
$effectsmodel.$

$\Rightarrow Waldtest:$

$$(\beta_{FE} - \beta_{RE})(\hat{V}COV_{FE} - \hat{V}COV_{RE})^(-1) * (\beta_{FE} - \beta_{RE})$$

$\Rightarrow$ Hausmann Test: Alternative: Variable addition

FE by within

2) plus all X bar i $\Rightarrow$ should be not having any expl power if $E(cov(x, \alpha)) = 0$

Test by F-test whether all $\bar{X}_{i,s}$ have zero parameters or not.

"Mundlak correction"

■=

RE.NL <- plm(formula.NL, data=dataNL, model = "random") cbind(coef(WI.NL), coef(Pool.NL)[2:7])

cbind(coef(WI.NL), coef(Pool.NL)[2:7], coef(RE.NL)[2:7])

@

We got a lower $R^2$