Week 14: Panel data

Daniel Stegmueller

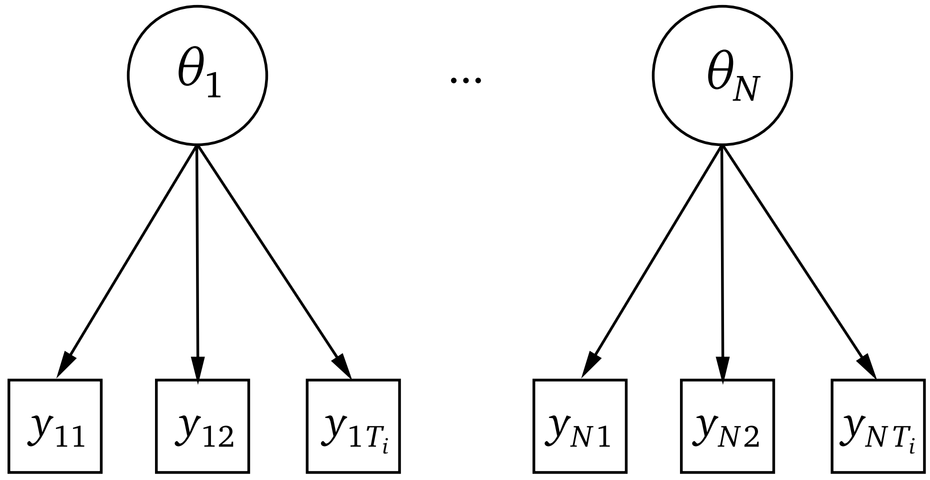
Dept. of Political Science
Duke University

Panel data

- Data observed on cross-section of units and over time
- Panel data of individuals or households
- Clinical trials
- Time series cross section data

Notation

- Units i, $i = 1, \ldots, N$
- Repeated observations t, $t = 1, \ldots, T_i$
- Thus it indexes cross-sections (individuals, countries, firms etc.) and time.



Model setup

- Most flexible model specification:

$$y_{it} = \alpha_{it} + \mathbf{x}_{it}\boldsymbol{\beta}_i + \epsilon_{it}$$

- Response y_{it}
- Intercepts, α_{it} varying over individuals and time
- Covariate vector $\mathbf{x}_{it}$ with time-constant and time varying variables
- Regression coefficients $\boldsymbol{\beta}_t$ varying over individuals
- Stochastic error term, ϵ_{it}, over time and individuals
- Too general to estimate (with $T \times N$ data points)

Pooled model

- Constant coefficients, no heterogeneity

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}$$

- Uses variation over i and t to estimate $\boldsymbol{\beta}$
- Consistently estimated by OLS if $\text{Cov}(\mathbf{x}_{it}, \epsilon_{it}) = 0$ and either T or $N \to \infty$.
- However $\text{Cor}(y_{is}, y_{it}) > 0$ and even after including covariates we are likely to have

$$\text{Cor}(\epsilon_{is}, \epsilon_{it}) > 0$$

- Thus standard errors will be too small and need to be corrected

Between model estimator

- Uses only 'cross-sectional' variation
- Average over all years:
$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i \boldsymbol{\beta} + \bar{\epsilon}_i$$
- Rewrite as between model:
$$\bar{y}_i = \alpha + \bar{\mathbf{x}}_i \boldsymbol{\beta} + (\alpha_i - \alpha + \bar{\epsilon}_i)$$
with $\bar{y}_i = 1/T \sum_t y_{it}$; $\bar{\epsilon}_i = 1/T \sum_t \epsilon_{it}$; and $\bar{\mathbf{x}}_i = 1/T \sum_t \mathbf{x}_{it}$
- Estimated via OLS
- The between estimator is consistent if the composite error term $(\alpha_i - \alpha + \bar{\epsilon}_i)$ is independent of covariates $\bar{\mathbf{x}}_i$

Heterogeneity via individual specific effects

- Alternatively, allow for heterogeneity using individual specific effects

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}$$

- Central assumption: exogenous regressors

$$E(\epsilon_{it}|a_i, \mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}) = 0, \quad \text{for } t = 1, \ldots, T$$

- Model variants:
  - Fixed effects: α_i is unobserved random variable possibly correlated with $\mathbf{x}_{it}$
  - Random effects: 'random intercept' model for individual effects, usually assuming

$$\alpha_i \sim N(0, \sigma^2_\alpha)$$
$$\epsilon_{it} \sim N(0, \sigma^2_\epsilon)$$

- Both models assume that $E(y_{it}|\mathbf{x}_{it}, \alpha_i) = \mathbf{x}_{it}\boldsymbol{\beta}$

Heterogeneity via individual specific effects

- Fixed effects estimation strategies:
    - *LSDV* estimation
    - *Within* estimation
    - *First differences* estimation
- Random effects estimation strategies:
    - *GLS* estimation
    - *ML* estimation

Fixed effects panel regression via within estimator

- Uses panel structure of the data
- Individual-specific deviations from time-averages of covariates and dependent variable
- Starting with individual heterogeneity model

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}$$

- Take averages over time

$$\bar{y}_i = \alpha_i + \bar{\mathbf{x}}_i\boldsymbol{\beta} + \bar{\epsilon}_i$$

- Subtracting yields the within/fixed effects estimator:

$$y_{it} - \bar{y}_i = (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\beta} + (\epsilon - \bar{\epsilon}_i), \quad t = 1, \ldots, T$$

- Note that we got rid of the α_is
- This is a consistent and efficient estimator of the fixed effects model

  (given that α_i are fixed effects and ϵ_{it} are iid)

Fixed effects panel regression via first-differences estimator

- Uses panel structure of the data
- Individual-specific changes of covariates and dependent variable
- Starting with individual heterogeneity model

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}$$

- Lag by one period

$$y_{it-1} = \alpha_i + \mathbf{x}_{it-1}\boldsymbol{\beta} + \epsilon_{it-1}$$

- Subtracting yields first differences estimator

$$y_{it} - y_{it-1} = (\mathbf{x}_{it} - \mathbf{x}_{it-1})\boldsymbol{\beta} + (\epsilon_{it} - \epsilon_{it-1}), \quad t = 2, \ldots, T$$

- Again, the α_is cancel
- This is a consistent estimator of the fixed effects model

  (though it is less efficient than the within variant for $T > 2$ and iid ϵ_{it})

Random effects panel regression via GLS

- Uses panel structure of the data
- Starting with individual heterogeneity model

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}, \quad \epsilon_{it} \sim N(0, \sigma_\epsilon^2),\ \alpha_i \sim N(0, \sigma_\alpha^2)$$

- The FGLS estimator of the RE model is

$$y_{it} - \hat{\lambda}\bar{y}_i = (1-\hat{\lambda})\mu + (\mathbf{x}_{it} - \hat{\lambda}\bar{\mathbf{x}}_i)\boldsymbol{\beta} + \nu_{it}$$

  with an (asymptotically) iid term

$$\nu_{it} = (1-\hat{\lambda})\alpha_i + (\epsilon_{it} - \hat{\lambda}\bar{\epsilon}_i)$$

- $\hat{\lambda}$ is consistent for

$$1 - \frac{\sigma_\epsilon}{\sqrt{\sigma_\epsilon^2 + T\sigma_\alpha^2}}$$

- Pooled OLS model if $\hat{\lambda} \to 0$, within estimation if $\hat{\lambda} \to 1$
- As $T \to \infty$, $\hat{\lambda} \to 1$

Correlation structure of RE model

- Rewrite RE model in error components formulation

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + u_{it}, \;\; u_{it} = \alpha_i + \epsilon_{it}$$

- Thus

$$\text{Cov}((\alpha_i + \epsilon_{it}), (\alpha_i + \epsilon_{is})) = \begin{cases} \sigma_\alpha^2 & \text{if } t \neq s \\ \sigma_\alpha^2 + \sigma_\epsilon^2 & \text{if } t = s \end{cases}$$

- This implies constant error correlations, or equicorrelated errors

$$\text{Cor}(u_{it}, u_{is}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2} \quad \text{for } t \neq s$$

Example: Wages and work hours

- Does labor supply react positively to wages?
- Data from Ziliak 1997.
- Balanced panel of 532 men from 1979–1988
- N=5320, T=10
- Annual hours worked (logged)
- Annual wage (logged)
- Basic model specification:

$$\mathrm{lnhr}_{it} = \beta \mathrm{lnwg}_{it} + \alpha_i + \epsilon_{it}$$

- Cross-sectional correlation: 0.123, p=0.000

Example, data preparation

- Several function make use of `pdata.frame` structure from library `plm`

```
pdata <- pdata.frame(data, index=c("id","year"))
```

- Variation in hours worked

```
summary(pdata$lnhr)

total sum of squares : 433.8
      id     time
0.392212 0.004524
```

- Variation in wages

```
summary(pdata$lnwg)

total sum of squares : 964.8
       id      time
0.8422643 0.0003686
```

Example, pooled model

- Pooled model OLS

```
m1 <- plm(lnhr~lnwg, index=c("id","year"), data=data,
    effect="individual", model="pooling")

Residuals :
    Min.   1st Qu.    Median   3rd Qu.      Max.
-4.83000 -0.08880 -0.00545  0.11400  0.96500

Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
(Intercept)  7.44152    0.02413  308.44   <2e-16
lnwg         0.08274    0.00913    9.07   <2e-16

Total Sum of Squares:    434
Residual Sum of Squares: 427
R-Squared      :  0.0152
Adj. R-Squared :  0.0152
F-statistic: 82.2223 on 1 and 5318 DF, p-value: <2e-16
```

Example, fixed effects model

- Fixed effects / within model

```
m2 <- plm(lnhr~lnwg, index=c("id","year"), data=data,
    effect="individual", model="within")

Residuals :
    Min.   1st Qu.    Median   3rd Qu.      Max.
-4.00000 -0.06170  0.00128  0.07760  1.27000

Coefficients :
     Estimate Std. Error t-value Pr(>|t|)
lnwg   0.1677     0.0189    8.89   <2e-16

Total Sum of Squares:    264
Residual Sum of Squares: 259
R-Squared       :  0.0162
Adj. R-Squared :  0.0146
F-statistic: 78.9578 on 1 and 4787 DF, p-value: <2e-16
```

Example, fixed effects model

- Fixed effects model via individual dummies (LSDV)

```
m2b <- lm(lnhr~ -1 + lnwg + factor(id), data=data)

Residuals:
   Min      1Q Median      3Q     Max
-4.004 -0.062  0.001  0.078  1.272

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
lnwg            0.1677     0.0189    8.89   <2e-16
factor(id)      ....

Residual standard error: 0.233 on 4787 degrees of freedom
Multiple R-squared: 0.999,  Adjusted R-squared: 0.999
F-statistic: 1.08e+04 on 533 and 4787 DF,  p-value: <2e-16
```

Example, random effects model, ML

- Random effects, (RE)ML estimation

```
m3 <- lmer(lnhr~lnwg + (1|id),  data=data)

 AIC BIC logLik deviance REMLdev
 556 583   -274      534     548
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.0264   0.162
 Residual             0.0543   0.233
Number of obs: 5320, groups: id, 532

Fixed effects:
            Estimate Std. Error t value
(Intercept)   7.3453     0.0365   201.4
lnwg          0.1196     0.0137     8.8
```

Example, random effects model, GLS

- Random effects, GLS estimation

```
m3b <- plm(lnhr~lnwg, index=c("id","year"), data=data,
    effect="individual", model="random")

Residuals :
    Min.   1st Qu.    Median   3rd Qu.      Max.
-4.32000 -0.06680  0.00288  0.08720  0.79300

Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
(Intercept)   7.3460     0.0364  201.86   <2e-16
lnwg          0.1193     0.0136    8.75   <2e-16

Total Sum of Squares:    293
Residual Sum of Squares: 289
R-Squared       :  0.0142
Adj. R-Squared  :  0.0142
F-statistic: 76.6383 on 1 and 5318 DF, p-value: <2e-16
```

Example, random effects model, GLS

- Variance decomposition

```
ercomp(m3b)

Effects:
                 var std.dev share
idiosyncratic 0.0542  0.2328  0.68
individual    0.0260  0.1612  0.32
theta:  0.585
```

Example, first differences model

- Fitted to first differences

```
m4 <- plm(lnhr~lnwg, index=c("id","year"), data=data,
    effect="individual", model="fd")

Residuals :
   Min.  1st Qu.   Median 3rd Qu.     Max.
-4.8000 -0.0728 -0.0041  0.0672  4.5500

Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
(intercept) 0.000828   0.004271    0.19     0.85
lnwg        0.108985   0.021335    5.11  3.4e-07

Total Sum of Squares:    420
Residual Sum of Squares: 418
R-Squared      :  0.00542
Adj. R-Squared :  0.00542
F-statistic: 26.0942 on 1 and 4786 DF, p-value: 3.38e-07
```

Estimation comparisons

Estim.	β	s.e.(β)	σ_{ϵ}^2	σ_{ξ}^2	N
Pooled	0.083	0.009	0.080		5320
Between	0.067	0.020			532
Within	0.168	0.019	0.049	0.033	5320
RE (GLS)	0.119	0.014	0.054	0.026	5320
RE (ML)	0.120	0.014	0.054	0.026	5320
First Diff.	0.109	0.021			4788

'Twoway models': adding time effects

- Allow for time effects or 'common shocks' experienced by all units of a cross-section
- Extended individual heterogeneity model:

$$y_{it} = \alpha_i + \gamma_t + \mathbf{x}_{it}\boldsymbol{\beta} + \epsilon_{it}$$

- Estimated by
  - Time dummies
  - Within-estimation, GLS RE
  - ML RE by adding an additional random intercept
- E.g., our within estimator is modified to regressing $y_{it} - \bar{y}_i - \bar{y}_t + \bar{\bar{y}}$ on $\mathbf{x}_{it} - \bar{\mathbf{x}}_i - \bar{\mathbf{x}}_t + \bar{\bar{\mathbf{x}}}_i$ with $\bar{y}_i$ as before, and

$$\bar{y}_t = 1/N \sum_{i=1}^{N} y_{it}, \text{ and } \bar{\bar{y}} = 1/NT \sum_{i=1}^{N} \sum_{i=1}^{T} y_{it}$$

Example contd., twoway FE

- Individual and time fixed effects

```
m6 <- plm(lnhr~lnwg, data=pdata,
    effect="twoway", model="within")

Coefficients :
     Estimate Std. Error t-value Pr(>|t|)
lnwg   0.1665     0.0188    8.84   <2e-16

Total Sum of Squares:    262
Residual Sum of Squares: 258
R-Squared      :  0.0161
Adj. R-Squared :  0.0144
F-statistic: 78.1203 on 1 and 4778 DF, p-value: <2e-16
```

Time fixed effects estimates

- Fixed time effects estimates, levels

```
summary(fixef(m6, effect="time", type="level"))
```

	Estimate	Std. Error	t-value	Pr(>|t|)
1979	7.2367	0.0500	145	<2e-16
1980	7.2267	0.0501	144	<2e-16
1981	7.2329	0.0502	144	<2e-16
1982	7.2107	0.0503	143	<2e-16
1983	7.1783	0.0502	143	<2e-16
1984	7.2035	0.0500	144	<2e-16
1985	7.2329	0.0503	144	<2e-16
1986	7.2259	0.0501	144	<2e-16
1987	7.2386	0.0503	144	<2e-16
1988	7.2426	0.0505	143	<2e-16

Time fixed effects estimates

- Display fixed time effect, deviations from mean

```
summary(fixef(m6, effect="time", type="dmean"))
```

```
     Estimate Std. Error t-value Pr(>|t|)
1979  0.01386    0.04997    0.28     0.78
1980  0.00384    0.05005    0.08     0.94
1981  0.00997    0.05021    0.20     0.84
1982 -0.01221    0.05028   -0.24     0.81
1983 -0.04455    0.05020   -0.89     0.37
1984 -0.01937    0.05001   -0.39     0.70
1985  0.01002    0.05028    0.20     0.84
1986  0.00299    0.05006    0.06     0.95
1987  0.01575    0.05028    0.31     0.75
1988  0.01971    0.05048    0.39     0.70
```

Example contd., twoway RE, GLS

- Individual and time random effects, GLS estimation

```
m7 <- plm(lnhr~lnwg, data=pdata,
    effect="twoway", model="random")
Effects:
                   var  std.dev share
idiosyncratic 0.053894 0.232150  0.67
individual    0.026030 0.161339  0.32
time          0.000324 0.017987  0.00
theta  : 0.586 (id) 0.512 (time) 0.43 (total)

Coefficients :
            Estimate Std. Error t-value Pr(>|t|)
(Intercept)   7.3470     0.0368  199.65   <2e-16
lnwg          0.1190     0.0136    8.74   <2e-16
```

Example contd., twoway RE, GLS

- Individual and time random effects, ML estimation

```
m7b <- lmer(lnhr~lnwg + (1|id) + (1|year),  data=data)

 AIC BIC logLik deviance REMLdev
 544 577   -267      520     534
Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 0.026413 0.1625
 year     (Intercept) 0.000298 0.0173
 Residual             0.053959 0.2323
Number of obs: 5320, groups: id, 532; year, 10

Fixed effects:
            Estimate Std. Error t value
(Intercept)   7.3462     0.0368   199.4
lnwg          0.1193     0.0136     8.7
```

Time random effects estimates

- Time random effects EB estimates and (normal-approx.) confidence intervals

```
m7b.re <- lme4::ranef(m7b)$year
m7b.re.lo <- m7b.re - 1.96*arm::se.ranef(m7b)$year
m7b.re.hi <- m7b.re + 1.96*arm::se.ranef(m7b)$year
m7b.re <- data.frame(m7b.re, m7b.re.lo, m7b.re.hi)
colnames(m7b.re) <- c("re.est", "lo", "hi")
round(m7b.re,3)
```

```
     re.est     lo     hi
1979  0.010 -0.008  0.028
1980  0.003 -0.016  0.021
1981  0.007 -0.011  0.026
1982 -0.009 -0.027  0.009
1983 -0.033 -0.052 -0.015
1984 -0.015 -0.033  0.004
1985  0.008 -0.011  0.026
1986  0.002 -0.016  0.020
1987  0.012 -0.006  0.030
1988  0.015 -0.003  0.034
```

Tests for individual specific effects

- Test for presence of individual specific effects (Breusch Pagan 1980)
- Based on residuals of OLS pooled model
- Null hypothesis: iid errors
- Test distributed χ^2 with 1 df.

```
test <- plm(lnhr~lnwg, data=pdata, model="pooling")
plmtest(test, effect="individual", type="bp")

data:  lnhr ~ lnwg
chisq = 2490, df = 1, p-value < 2.2e-16
alternative hypothesis: significant effects
```

- BP $= 2490 > \chi^2_{0.05}(1) = 3.84$ rejects null hypothesis

F-test of within against pooling model

- F test of pooled model versus within model with individual and/or time specific effects
- Test distributed F with (M_1, M_2) degrees of freedom
- Example: Individual specific effects

```
testw1 <- plm(lnhr~lnwg, data=pdata,
    effect="individual", model="within")
pFtest(testw1, test)

    F test for individual effects

data:  lnhr ~ lnwg
F = 5.833, df1 = 531, df2 = 4787, p-value < 2.2e-16
alternative hypothesis: significant effects
```

- $F = 5.8 > F_{0.05}(531, 4787) = 1.1$ rejects null hypothesis

Fixed versus random effects

- Both models assume that $E(y_{it}|\mathbf{x}_{it}, \alpha_i) = \mathbf{x}_{it}\boldsymbol{\beta}$
- We cannot estimate $E(y_{it}|\mathbf{x}_{it}, \alpha_i)$ in a short (i.e., small T) panel
- Eliminate α_i
$$E(y_{it}|\mathbf{x}_{it}) = E(\alpha_i|\mathbf{x}_{it}) + \mathbf{x}_{it}\boldsymbol{\beta}$$
- RE model assumes $E(\alpha_i|\mathbf{x}_{it}) = \alpha$, thus $E(y_{it}|\mathbf{x}_{it}) = \alpha + \mathbf{x}_{it}\boldsymbol{\beta}$ and one can identify $E(y_{it}|\mathbf{x}_{it})$
- FE model $E(\alpha_i|\mathbf{x}_{it}) = \alpha$ varies with $\mathbf{x}_{it}$, thus we cannot identify $E(y_{it}|\mathbf{x}_{it})$
- It is possible to identify marginal effect
$$\boldsymbol{\beta} = \partial E(y_{it}|\alpha_i, \mathbf{x}_{it})/\partial \mathbf{x}_{it}$$
for time-varying covariates

Hausman test

- If individual effects are fixed (i.e., assuming the FE model is a correct description of the DGP), within estimator is consistent, random effects estimate is inconsistent
- Test for statistically significant difference between within-estimates $\hat{\boldsymbol{\beta}}_W$ of *time-varying* covariates and random-effects estimates $\tilde{\boldsymbol{\beta}}_R$
- Large test value rejects null hypothesis: individual effects α_i are uncorrelated with covariates
- *Two* possible responses
  - Decide to use fixed effects model
  - Respecify random effects model

Hausman test computation

- Start with random effects model and iid error components $\epsilon \sim (0, \sigma^2_\epsilon)$ and $\alpha \sim (0, \sigma^2_\alpha)$
- The Hausman test statistics is

$$H = \left(\tilde{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_W\right)\left[\hat{V}(\tilde{\boldsymbol{\beta}}_W) - \hat{V}(\hat{\boldsymbol{\beta}}_R)\right]^{-1}\left(\tilde{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_W\right)$$

- H is asymptotically distributed χ^2 with dim($\boldsymbol{\beta}$) degrees of freedom
- Alternatively, perform Wald test of $\boldsymbol{\gamma} = 0$ in

$$y_{it} - \hat{\lambda}\bar{y}_i = (1 - \hat{\lambda})\mu + (\mathbf{x}_{it} - \hat{\lambda}\bar{\mathbf{x}}_i)\boldsymbol{\beta} + \nu_{it} + (\mathbf{x}_{it} - \bar{\mathbf{x}}_i)\boldsymbol{\gamma}$$

- $\boldsymbol{\gamma} = 0$ yields RE estimator
- If ν_{it} is correlated with covariates, additional functions of regressors (i.e.,($\mathbf{x}_{it} - \bar{\mathbf{x}}_i$)) will be significantly different from zero

Hausman test, hours & wages example

- Compute FE and RE model

```
fe <- plm(lnhr~lnwg, data=pdata, effect="individual", model="within")
re <- plm(lnhr~lnwg, data=pdata, effect="individual", model="random")
```

- General Hausman test

```
phtest(fe,re)

data:  lnhr ~ lnwg
chisq = 13.73, df = 1, p-value = 0.0002115
alternative hypothesis: one model is inconsistent
```

- $H = 14 > \chi^2_{0.05}(1) = 3.84$ rejects the specified random effects model

Correlated random effects

- Deal with possible correlation between α_i and covariates in RE framework
- Mundlak (1978) allowed α_is to depend on time averages of covariates

$$\alpha_i = \bar{\mathbf{x}}_i \boldsymbol{\delta} + w_i$$

- GLS estimation of $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ yields $\boldsymbol{\beta}$ estimates equal to the FE model
- Easy implementation
  - Create time-constant covariate vector $\bar{\mathbf{x}}_i$
  - Estimate RE model

$$y_{it} = \alpha_i + \mathbf{x}_{it}\boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\delta} + \epsilon_{it}$$

Correlated random effects example

- Correlated RE model

```
m3corr <- plm(lnhr~lnwg+lnwgm, data=pdata,
    effect="individual", model="random")
```

```
              Estimate Std. Error t-value Pr(>|t|)
(Intercept)     7.4830     0.0519  144.23  < 2e-16
lnwg            0.1677     0.0189    8.89  < 2e-16
lnwgm          -0.1008     0.0273   -3.70  0.00022
```

- Previously estimated RE model

```
              Estimate Std. Error t-value Pr(>|t|)
(Intercept)     7.3460     0.0364  201.86   <2e-16
lnwg            0.1193     0.0136    8.75   <2e-16
```

- Previously estimated within/FE model

```
              Estimate Std. Error t-value Pr(>|t|)
lnwg            0.1677     0.0189    8.89   <2e-16
```

Robust standard errors

- Adapt 'robust/Huber-White/sandwich' correction to panel data
- Take into account possible serial correlation over time that is different for different units, i.e.,

$$\mathrm{Cov}(u_{it}, u_{is}) > 0, s \neq t$$

- A general version of our panel models can be written as

$$\tilde{y}_{it} = \tilde{\mathbf{w}}_{it}\boldsymbol{\theta} + \tilde{u}_{it}$$

  where suitable transformations yield each previous estimator

- Stacking observation over time yields

$$\tilde{\mathbf{y}}_i = \tilde{\mathbf{W}}_i\boldsymbol{\theta} + \tilde{\mathbf{u}}_i$$

- The OLS estimator is

$$\hat{\boldsymbol{\theta}} = \left(\sum_{i=1}^{N} \tilde{\mathbf{W}}_i'\tilde{\mathbf{W}}_i\right)^{-1} \sum_{i=1}^{N} \tilde{\mathbf{W}}_i'\tilde{\mathbf{y}}_i$$

Robust standard errors

- Its asymptotic variance is

$$V(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{N} \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i\right)^{-1} \sum_{i=1}^{N} \tilde{\mathbf{W}}_i' \, E\left(\tilde{\mathbf{u}}_i' \tilde{\mathbf{u}}_i | \tilde{\mathbf{W}}_i\right) \tilde{\mathbf{W}}_i \left(\sum_{i=1}^{N} \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i\right)^{-1}$$

assuming $E(\tilde{\mathbf{W}}_i' \tilde{\mathbf{u}}_i) = \mathbf{0}$

- Arrelano (1987) proposed an estimate allowing for both heteroskedasticity and serial correlation:

$$\hat{V}(\hat{\boldsymbol{\theta}}) = \left(\sum_{i=1}^{N} \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i\right)^{-1} \sum_{i=1}^{N} \tilde{\mathbf{W}}_i' \, E\left(\hat{\mathbf{u}}_i' \hat{\mathbf{u}}_i | \tilde{\mathbf{W}}_i\right) \tilde{\mathbf{W}}_i \left(\sum_{i=1}^{N} \tilde{\mathbf{W}}_i' \tilde{\mathbf{W}}_i\right)^{-1}$$

with

$$\hat{\mathbf{u}}_i = \hat{\tilde{\mathbf{u}}}_i = \tilde{y}_{it} - \tilde{\mathbf{w}}_{it} \hat{\boldsymbol{\theta}}$$

Robust standard errors, wages and hours worked

- Use R package `sandwich` to calculate robust SE for fitted `plm` objects
- E.g., results of within model M2

```
coeftest(m2)
     Estimate Std. Error t value Pr(>|t|)
lnwg   0.1677     0.0189    8.89   <2e-16
```

- Panel-robust standard errors

```
coeftest(m2, vcov=vcovHC(m2, method="arellano"))
     Estimate Std. Error t value Pr(>|t|)
lnwg   0.1677     0.0849    1.98    0.048
```

- Discuss! ...

Test: Pooling

- Test for
- Compare models
  - M_1: estimated on full sample
  - M_2: estimated equation for each individual
- F-test with (M_1, M_2) degrees of freedom

```
testpld <- plm(lnhr~lnwg, data=pdata, model="within")
testnp <- pvcm(lnhr~lnwg, data=pdata, model="within")

pooltest(testpld,testnp)

data:  lnhr ~ lnwg
F = 2.835, df1 = 531, df2 = 4256, p-value < 2.2e-16
alternative hypothesis: unstability
```

Random coefficient models

- Model allowing for effect heterogeneity for covariates $\mathbf{w}_{it}$

$$y_{it} = \mathbf{z}_{it}\boldsymbol{\beta} + \mathbf{w}_{it}\boldsymbol{\alpha}_i + \epsilon_{it}$$

  where $\boldsymbol{\alpha}_i$ is a random vector with zero mean.

- The *random parameters* or *random coefficients model* (Swamy 1970) specifies

$$y_{it} = \mathbf{z}_{it}\boldsymbol{\beta}_i + \epsilon_{it}$$

  with

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\alpha}_i$$

- Substituting yields

$$y_{it} = \mathbf{z}_{it}\boldsymbol{\beta} + \mathbf{z}_{it}\boldsymbol{\alpha}_i + \epsilon_{it}$$

- Estimation is possible by (RE)ML or FGLS

Random coefficient model of wages and hours worked, GLS

- Specification $y_{it} = \mathbf{z}_{it}\boldsymbol{\beta}_i + \epsilon_{it}$, $\boldsymbol{\beta}_i = \boldsymbol{\beta} + \boldsymbol{\alpha}_i$
- FGLS estimates

```
vc <- pvcm(lnhr~lnwg, data=pdata,
    effect="individual", model="random")

Estimated mean of the coefficients:
            Estimate Std. Error z-value Pr(>|z|)
(Intercept)   7.7562     0.0666  116.43   <2e-16
lnwg         -0.0301     0.0257   -1.17     0.24

Estimated variance of the coefficients:
            (Intercept)   lnwg
(Intercept)       1.654 -0.629
lnwg             -0.629  0.248

Total Sum of Squares: 69200
Residual Sum of Squares: 442
Multiple R-Squared: 0.994
```

Random coefficient model of wages and hours worked, (RE)ML

- ML (or REML) estimates

```
vcml <- lmer(lnhr~lnwg + (1+lnwg|id), data=data)

   AIC   BIC logLik deviance REMLdev
 -60.5 -21.1   36.3    -86.5   -72.5

Random effects:
 Groups   Name        Variance Std.Dev. Corr
 id       (Intercept) 0.6885   0.830
          lnwg        0.0859   0.293    -0.988
 Residual             0.0463   0.215
Number of obs: 5320, groups: id, 532

Fixed effects:
            Estimate Std. Error t value
(Intercept)   7.5725     0.0554   136.8
lnwg          0.0356     0.0203     1.8
```