Longitudinal or Panel Analysis

David Barron

Trinity Term 2018

Basic concepts

Panel Data

The main characteristic of longitudinal or panel data is that a group of individuals (people, firms, etc.) are surveyed at (usually) regular intervals. Advantages include:

- Can study dynamics
- Sequence of events in time helps show causation. For example, married men generally earn more, but is this a causal effect?
- Can control for unobserved heterogeneity

What analyses can be done?

You can do linear regression, logit, poisson, negative binomial regressions (and a number of others that we won't be covering) in panel or longitudinal format. These versions allow us to deal with some of the issues associated with these kinds of data. In particular, we can't treat each observation as independent. There will almost certainly be more variation from individual to individual than there will be within an individual over time.

Wide data

Using data from chapter 2 of Singer & Willett. *Wide* data has the form:

```
   id tol11 tol12 tol13 tol14 tol15 male exposure
1   9  2.23  1.79  1.90  2.12  2.66    0     1.54
2  45  1.12  1.45  1.45  1.45  1.99    1     1.16
3 268  1.45  1.34  1.99  1.79  1.34    1     0.90
4 314  1.22  1.22  1.55  1.12  1.12    0     0.81
5 442  1.45  1.99  1.45  1.67  1.90    0     1.13
6 514  1.34  1.67  2.23  2.12  2.44    1     0.90
```

Data often come in this form, e.g., BHPS, because there are fewer observations. However, for analysis the data needs to be in *long* format.

Long data

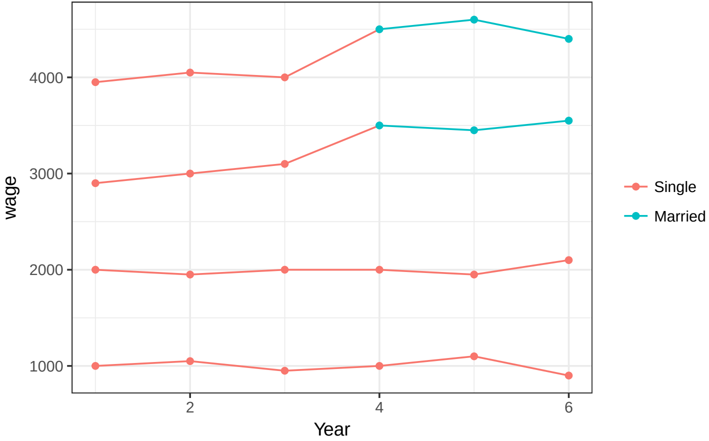
Long data is obtained using the gather command.

```
tol.long <- tolerance %>%
  gather(tol, tolerance, starts_with("tol")) %>%
  mutate(
    age = as.numeric(str_extract(tol, "[1-9]+")),
    time = age - 11
  ) %>%
  arrange(id) %>%
  as_data_frame()
head(tol.long, n = 10)
```

```
# A tibble: 10 x 7
      id  male exposure tol   tolerance   age  time
   <int> <int>    <dbl> <chr>     <dbl> <dbl> <dbl>
 1     9     0     1.54 tol11      2.23    11     0
 2     9     0     1.54 tol12      1.79    12     1
 3     9     0     1.54 tol13      1.9     13     2
 4     9     0     1.54 tol14      2.12    14     3
 5     9     0     1.54 tol15      2.66    15     4
 6    45     1     1.16 tol11      1.12    11     0
 7    45     1     1.16 tol12      1.45    12     1
 8    45     1     1.16 tol13      1.45    13     2
 9    45     1     1.16 tol14      1.45    14     3
10    45     1     1.16 tol15      1.99    15     4
```

Fixed-effects

Plot



The problem

How do we distinguish the *causal* impact of getting married on wages from the possibility that men with higher wages are more likely to get married? Suppose we had only cross-sectional data:

```
lm(formula = wage ~ married, data = bru, subset = year == 4)
            coef.est coef.se t value Pr(>|t|)
(Intercept) 1500.00   500.00    3.00    0.10
married     2500.00   707.11    3.54    0.07
---
n = 4, k = 2
residual sd = 707.11, R-Squared = 0.86
```

Married men earn on average 2500 more than unmarried men. But the mean difference between the same pairs of men the year before was 2075.

Pooled estimates

We could pool the data together and do a standard linear regression:

```
lm(formula = wage ~ married, data = bru)
            coef.est coef.se t value Pr(>|t|)
(Intercept) 2166.67   236.18    9.17    0.00
married     1833.33   472.37    3.88    0.00
---
n = 24, k = 2
residual sd = 1002.04, R-Squared = 0.41
```

This is something of an improvement as we do at least have some highly paid, unmarried men in the sample now, so the effect of marriage appears smaller. But it is still very biased.

Unobserved heterogeneity/Endogeneity

In many ways the fundamental problem with regression is presence of *unobserved heterogeneity*. In this case we are not taking account of factors that explain both why men 3 and 4 are more likely to get married and earn higher wages.

Alternatively, we might think that there is a problem of *endogeneity*: men 3 and 4 are more likely to get married *because* they earn higher wages.

Either way, bias is introduced because there is a correlation between an explanatory variable and the error term.

Differences in differences

Compare mean before and after marriage wages of men 3 and 4 with the change in mean wages of men 1 and 2 over the same time.

ID	Years 1–3	Years 4–6	Difference
1	1000	1000	0
2	2000	2000	0
3	3000	3500	500
4	4000	4500	500

So, the mean increase in wages following marriage is 500. **All the rest of the apparent marriage effect is due to other differences between the men.** NB, if there had been some time-varying effect increasing average wages in the later years, this method would also have controlled for that.

Least Squares Dummy Variables

The easiest way to achieve the same result is to put in a dummy variable for each individual:

```
lm(formula = wage ~ married + id, data = bru)
            coef.est coef.se t value Pr(>|t|)
(Intercept) 1000.00    28.10   35.59    0.00
married      500.00    39.74   12.58    0.00
id2         1000.00    39.74   25.17    0.00
id3         2000.00    44.43   45.02    0.00
id4         3000.00    44.43   67.53    0.00
---
n = 24, k = 5
residual sd = 68.82, R-Squared = 1.00
```

This is the LSDV or **fixed-effects** estimator.

Estimation in R

Using a factor is OK for this toy data, but gets unwieldy quickly. The package plm is a good alternative.

```
Oneway (individual) effect Within Model

Call:
plm(formula = wage ~ married, data = bru, index = c("id", "year"))

Balanced Panel: n = 4, T = 6, N = 24

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
   -100     -50       0      50     100

Coefficients:
        Estimate Std. Error t-value Pr(>|t|)
married    500.0       39.7    12.6  1.2e-10

Total Sum of Squares:    840000
Residual Sum of Squares: 90000
R-Squared:      0.893
Adj. R-Squared: 0.87
F-statistic: 158.333 on 1 and 19 DF, p-value: 1.16e-10
```

Decomposing errors

The basic panel regression model is:

$$y_{it} = \beta_0 + \beta_1 x_{1it} + \beta_2 x_{2it} + \cdots + u_i + \epsilon_{it},$$

where the u_i terms are individual-specific effects and the ϵ_{it} is equivalent to the standard OLS error term (and should fulfill the same assumptions). The mean over time of all components in the equation is:

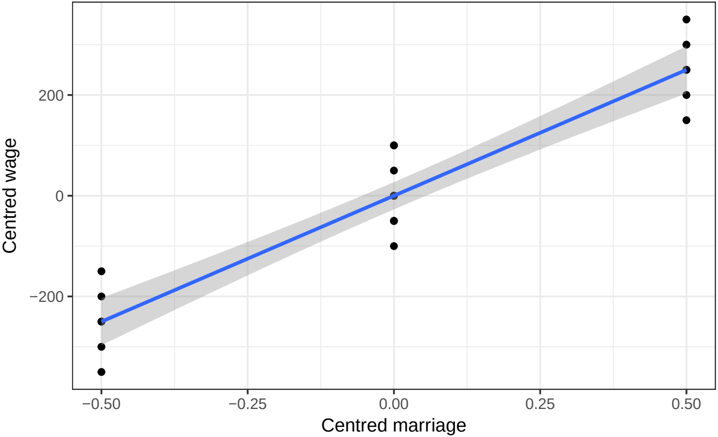
$$\begin{aligned}
\bar{y}_i &= \beta_0 + \beta_1 \bar{x}_{1i} + \beta_2 \bar{x}_{2i} + \cdots + u_i + \bar{\epsilon}_i; \\
y_{it} - \bar{y}_i &= \beta_1 (x_{1it} - \bar{x}_{1i}) + \beta_2 (x_{2it} - \bar{x}_{2i}) + \epsilon_{it} - \bar{\epsilon}_i.
\end{aligned}$$

Removing the means

```
lm(formula = mwage ~ mmarried + 0, data = bru.2)
         coef.est coef.se t value Pr(>|t|)
mmarried 500.00    36.12   13.84    0.00
---
n = 24, k = 1
residual sd = 62.55, R-Squared = 0.89
```

Notice R^2 is same as above.

Plot



Restrictions of FE estimator

- Can't estimate effects of variables that don't vary over time.
- Uses lots of degrees of freedom.
- Multicollinearity of dummy variables inflates standard errors.

Random effects

Random effects model

Looking again at the basic equation, we now specify that the u_i are *random variables*, each iid, and all uncorrelated with the explanatory variables. From this we can obtain:

$$\begin{aligned} y_{it} - \theta \bar{y}_i = & \beta_0(1-\theta) + \beta_1(x_{1it} - \theta \bar{x}_{1i}) \\ & + \beta_2(x_{2it} - \theta \bar{x}_{2i}) + \cdots \\ & + \{(1-\theta)u_i + (\epsilon_{it} - \theta \bar{\epsilon}_i)\}, \end{aligned}$$

where

$$\theta = \sqrt{\frac{\sigma^2_\epsilon}{(T \times \sigma^2_\epsilon) + \sigma^2_u}}$$

Example

```
Oneway (individual) effect Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = wage ~ married.f, data = bru, model = "random",
    index = c("id", "year"))

Balanced Panel: n = 4, T = 6, N = 24

Effects:
                   var  std.dev share
idiosyncratic   4736.8     68.8  0.01
individual    499210.5    706.5  0.99
theta: 0.96

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
 -159.6   -59.6   -14.7    58.8   158.0

Coefficients:
                 Estimate Std. Error t-value Pr(>|t|)
(Intercept)        2499.2      406.0    6.16  3.4e-06
married.fMarried    503.2       45.6   11.03  2.0e-10

Total Sum of Squares:    897000
Residual Sum of Squares: 137000
R-Squared:      0.847
Adj. R-Squared: 0.84
F-statistic: 121.758 on 1 and 22 DF, p-value: 1.96e-10
```

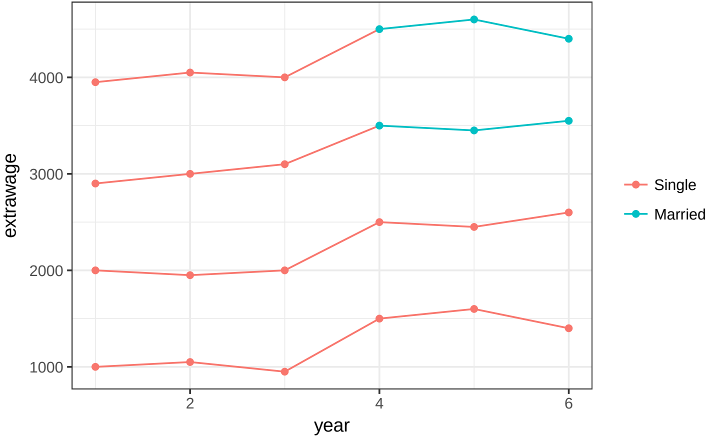
Notice that θ is close to 1. When it is 1, we have the FE estimator again. When it is 0, we have the pooled OLS estimator.

Problems with RE model

Big problem is the assumption that $Cov(x_{it}, u_i) = 0$. Mostly we would doubt this assumption. If it is false, estimates will be biased. FE estimator often thought to be more conservative choice. However, the assumption can be relaxed, and people often want to estimate the effect of variables that don't change over time (sex, ethnicity, etc.), and so use RE.

Other issues

Plot



Time trends

In this modified example, everyone gets an extra 500 added to their wages after year 3. However, the FE estimator still shows a marriage effect:

```
Oneway (individual) effect Within Model

Call:
plm(formula = extrawage ~ married.f, data = bru, index = c("id",
    "year"))

Balanced Panel: n = 4, T = 6, N = 24

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
   -300    -125       0     113     350

Coefficients:
                 Estimate Std. Error t-value Pr(>|t|)
married.fMarried      500        125       4  0.00076

Total Sum of Squares:    1640000
Residual Sum of Squares: 890000
R-Squared:      0.457
Adj. R-Squared: 0.343
F-statistic: 16.0112 on 1 and 19 DF, p-value: 0.000764
```

Period effects

The solution is to include wave dummies:

```
lm(formula = extrawage ~ married.f + id + factor(year), data = bru)
                 coef.est coef.se t value Pr(>|t|)
(Intercept)       958.33    48.67 19.69    0.00
married.fMarried  -16.67    61.56 -0.27    0.79
id2              1000.00    43.53 22.97    0.00
id3              2008.33    53.31 37.67    0.00
id4              3008.33    53.31 56.43    0.00
factor(year)2      50.00    53.31  0.94    0.36
factor(year)3      50.00    53.31  0.94    0.36
factor(year)4     545.83    61.56  8.87    0.00
factor(year)5     570.83    61.56  9.27    0.00
factor(year)6     533.33    61.56  8.66    0.00
---
n = 24, k = 10
residual sd = 75.40, R-Squared = 1.00
```

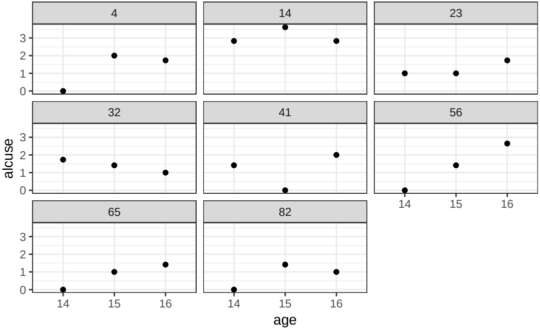
Alternative using plm

```
Twoways effects Within Model

Call:
plm(formula = extrawage ~ married.f, data = bru, effect = "twoways",
    index = c("id", "year"))

Balanced Panel: n = 4, T = 6, N = 24

Residuals:
   Min. 1st Qu.  Median 3rd Qu.    Max.
 -91.67  -58.33   -4.17   41.67  108.33

Coefficients:
                 Estimate Std. Error t-value Pr(>|t|)
married.fMarried    -16.7       61.6   -0.27     0.79

Total Sum of Squares:    80000
Residual Sum of Squares: 79600
R-Squared:      0.00521
```

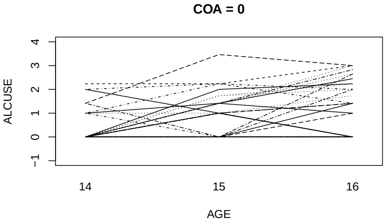
Multilevel model of change

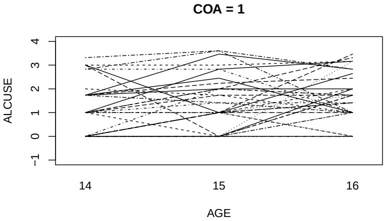
Mulilevel model of change

The ability to model change is a key benefit of panel data. This is really a type of multilevel data, as we have within-person change and between person differences in change. Panel data can distinguish the two. Looking at the example from the textbook (chapter 4), we have three observations on alcohol use among teenagers, at age 14,15 and 16. Here are 9 example cases:



Differences with COA





Multilevel representation

$$Y_{it} = \beta_{0i} + \beta_{1i} T_{it} + \epsilon_{it};$$

$$\beta_{0i} = \gamma_{00} + \gamma_{01} COA_i + u_{0i}$$
$$\beta_{1i} = \gamma_{10} + \gamma_{11} COA_i + u_{1i}$$

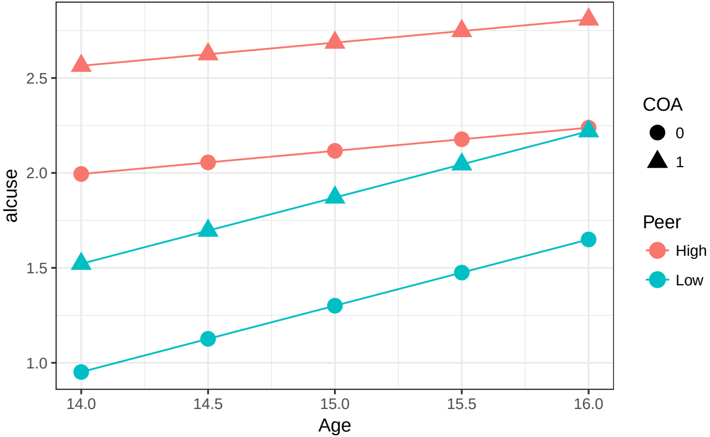
Results

```
lmer(formula = alcuse ~ coa * age_14 + (1 + age_14 | id), data = alc
    REML = FALSE)
            coef.est coef.se t value
(Intercept)  0.32     0.13    2.42
coa1         0.74     0.19    3.82
age_14       0.29     0.08    3.48
coa1:age_14 -0.05     0.13   -0.39

Error terms:
 Groups   Name        Std.Dev. Corr
 id       (Intercept) 0.70
          age_14      0.39     -0.22
 Residual             0.58
---
number of obs: 246, groups: id, 82
AIC = 637.2, DIC = 621
deviance = 621.2
```

Full model

```
lmer(formula = alcuse ~ coa + peer * age_14 + (1 + age_14 | id),
    data = alcohol1, REML = FALSE)
            coef.est coef.se t value
(Intercept) -0.31     0.15   -2.15
coa1         0.57     0.15    3.91
peer         0.70     0.11    6.25
age_14       0.42     0.11    4.02
peer:age_14 -0.15     0.08   -1.79

Error terms:
 Groups   Name        Std.Dev. Corr
 id       (Intercept) 0.49
          age_14      0.37     -0.03
 Residual             0.58
---
number of obs: 246, groups: id, 82
AIC = 606.7, DIC = 589
deviance = 588.7
```

Effects


alcuse
Age
COA
● 0
▲ 1
Peer
High
Low