13.3 Two-Period Panel Data Analysis

1. Consider the model for crime introduced earlier. Let us estimate the following functional form:

$$crmrte = \beta_0 + \beta_1 unem + u$$

Lets estimate an OLS for the model without any controls first:

```
> summary(lm(crmrte ~ unem, data = crime2, subset = (year == 87)))

Call:
lm(formula = crmrte ~ unem, data = crime2, subset = (year ==
    87))

Residuals:
   Min     1Q Median     3Q    Max
-57.55 -27.01 -10.56  18.01  79.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  128.378     20.757   6.185  1.8e-07 ***
unem          -4.161      3.416  -1.218     0.23
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 34.6 on 44 degrees of freedom
Multiple R-squared:  0.03262,   Adjusted R-squared:  0.01063
F-statistic: 1.483 on 1 and 44 DF,  p-value: 0.2297
```

We cannot infer the *unem* parameter because it is imprecisely estimated. But even if we did, it tells us that an increase in unemployment reduces crime – which is counter intuitive.

2. Adding controls is one solution: For instance, we could add crmrte from previous years as control with the argument that previous crime intensity could explain current crime intensity.

```
> mod1 = lm(log(crime2$crmrte)  ~  + unem + log(crime2$lawexpc) , data = crime2, subset =
 (year == 87))
> summary(mod1)

Call:
lm(formula = log(crime2$crmrte) ~ +unem + log(crime2$lawexpc),
    data = crime2, subset = (year == 87))

Residuals:
     Min       1Q   Median       3Q      Max
-0.64786 -0.22955 -0.06368  0.22183  0.71164

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)          3.34290    1.25053   2.673   0.0106 *
unem                -0.02900    0.03234  -0.897   0.3748
log(crime2$lawexpc)  0.20337    0.17265   1.178   0.2453
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3231 on 43 degrees of freedom
Multiple R-squared:  0.05712,   Adjusted R-squared:  0.01326
F-statistic: 1.302 on 2 and 43 DF,  p-value: 0.2824

> nobs(mod1)
[1] 46
```

Adding past crime rate

```
> mod2 = lm(subset(log(crime2$crmrte),crime2$year==87)  ~  subset(unem,crime2$year==87)
 + subset(log(crime2$lawexpc),crime2$year==87) + subset(log(crime2$crmrte), crime2$year==
82) , data = crime2)
> summary(mod2)

Call:
lm(formula = subset(log(crime2$crmrte), crime2$year == 87) ~
    subset(unem, crime2$year == 87) + subset(log(crime2$lawexpc),
        crime2$year == 87) + subset(log(crime2$crmrte), crime2$year ==
        82), data = crime2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48081 -0.12202  0.00659  0.14658  0.34428

Coefficients:
                                                Estimate Std. Error t value
(Intercept)                                     0.076450   0.821143   0.093
subset(unem, crime2$year == 87)                 0.008621   0.019517   0.442
subset(log(crime2$lawexpc), crime2$year == 87) -0.139576   0.108641  -1.285
subset(log(crime2$crmrte), crime2$year == 82)   1.193923   0.132099   9.038
                                               Pr(>|t|)
(Intercept)                                       0.926
subset(unem, crime2$year == 87)                   0.661
subset(log(crime2$lawexpc), crime2$year == 87)    0.206
subset(log(crime2$crmrte), crime2$year == 82)   2.1e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1905 on 42 degrees of freedom
Multiple R-squared:  0.6798,    Adjusted R-squared:  0.657
F-statistic: 29.73 on 3 and 42 DF,  p-value: 1.799e-10

> nobs(mod2)
[1] 46
```

3. Unobserved effects: Alternatively, we could think about factors which affect the dependent variable to be categorized into two types: (a) a time varying factor $d2_t$ which is a dummy variable that equals zero when t = 1 and one when t = 2 and (b) a_i which is a time-constant factor that affects y_{it}. The latter is called a **fixed effect or an unobserved effect**:

$$y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it} \qquad t = 1, 2$$

The u_{it} is called the idiosyncratic error or a time varying error because it represents unobserved factors that change over time and affect y_{it}.

4. City fixed effects: consider the model for crime which we used earlier, but now the unobserved effects version of the model:

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + a_i + u_{it}$$

Since the unit of analysis i is a city, the factors affecting city crime rates that do not change over time are captured in a_i. For instance, the area, location, weather, etc might be constant for a city in the dataset.

Are all factors constant in a city? How do we account for these factors?

5. How do we estimate this model? First we could pool these two years and run a pooled regression. But for pooled OLS to be consistent, we need to assume that a_i is uncorrelated with $uunem_{it}$ and $d87_t$ when we run the following model:

$$crmrte_{it} = \beta_0 + \delta_0 d87_t + \beta_1 unem_{it} + v_{it}$$

where $v_{it} = a_i + u_{it}$ is a form of composite error when we ignore the fixed effect. If a_i is correlated with the x_{it}, then the resulting bias is called the **heterogeneity bias**.

```
> library(wooldridge)
> summary(lm(crmrte ~ unem + d87, data = crime2))

Call:
lm(formula = crmrte ~ unem + d87, data = crime2)

Residuals:
    Min      1Q  Median      3Q     Max
-53.474 -21.794  -6.266  18.297  75.113

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  93.4202    12.7395   7.333 9.92e-11 ***
unem          0.4265     1.1883   0.359    0.720
d87           7.9404     7.9753   0.996    0.322
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 29.99 on 89 degrees of freedom
Multiple R-squared:  0.01221,   Adjusted R-squared:  -0.009986
F-statistic: 0.5501 on 2 and 89 DF,  p-value: 0.5788
```

Pooling does not help us obtain trustworthy estimates

6. First differencing methods: consider the two time periods

$$\begin{aligned} crmrte_{i82} =& \beta_0 + \beta_1 unem_{i82} + a_i + u_{i82} \\ crmrte_{i87} =& \beta_0 + \delta_0 d87_t + \beta_1 unem_{i87} + a_i + u_{i87} \end{aligned}$$

If we subtract first from second we get:

$$\begin{aligned} (crmrte_{i87} - crmrte_{i82}) =& \delta_0 + \beta_1 (unem_{i87} - unem_{i82}) \\ & + a_i + (u_{i87} - u_{i82}) \end{aligned}$$

This is written as an estimable first difference equation:

$$\Delta crmrte_{it} = \delta_0 + \beta_1 \Delta unem_{it} + \Delta u_{it}$$

Thus this is a single equation where all variables are differenced over-time. Notice that the unobserved effect which was constant in both time periods gets canceled.

7. Strict Exogeniety and First Difference: Like previously, we need $\Delta unem_{it}$ to be uncorrelated with Δu_{it}. This implies that unemployment in 82 does not affect Δu_{it} in 82 and 87.

   When we observe the OLS estimate of the above equation, we call it the first differenced estimator.

   When would this fail? (a) If law enforcement increases in cities where unemployment rate decreases. This could cause a negative correlation between Δu_{it} and $\Delta unem_{it}$ and a biased OLS.

   (b) Also, $\Delta unem_{it}$ should have variation across i. If there is no variability, then the β would be under-estimated.

   (c) We need to assume homskedasticity.If it does not, we could also correct it for heteroskedasticity.

$$\widehat{\Delta crmrte} = 15.40 + 2.22\ \Delta unem$$
$$(4.70) \quad (.88)$$
$$n = 46, R^2 = .127,$$

Consider the interpretation: Even if $\Delta unem_{it} = 0$, crime rate would increase by 15.40 per 1000 people every year. So there is a trend increase in time series.

8. Limitations of First Differencing Methods: (a) You might not have many data sets conducive to first differencing, especially a panel on Individual behavior might be really scarce. (b) First differencing could throw out a_i that is variation across individuals. We might not be left with a lot of variation across time to obtain small standard errors for our estimates. Even having a large sample might not provide a solution.

9. Explaining difference in wages across individuals using panel data: Consider the following model for wages:

$$log(wage)_{it} = \beta_0 + \delta_0 d2_t + \beta_1 educ_{it} + a_i + u_{it}$$

Because, by definition, innate ability does not change over time, panel data methods seem ideally suited to estimate the return to education. This equation considers intercepts which change overtime represented by δ_0.

To estimate it and remove a_i in the case when a_i is associated with $educ_{it}$, we could potentially difference the equation.

$$\Delta log(wage)_{it} = \delta_0 + \beta_1 \Delta educ_{it} + \Delta u_{it}$$

But here, first differencing does not help as the variation in education in the sample is very less causing us to get imprecise estimates of β.

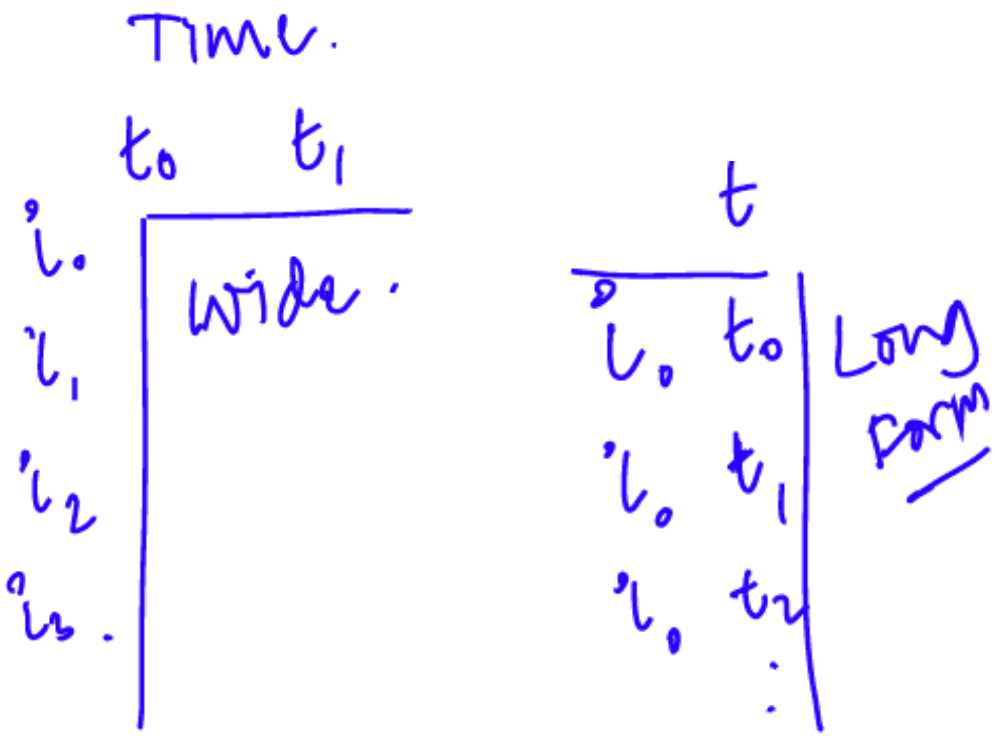
10. Adding several explanatory variables is straight forward. (Example 13.5: Biddle and Hamermesh, 1990)

11. Finite Distributed Lag Model: Consider the model by Eddie (1994) to examine crime. His main idea is that the effect of convictions could take more than a year to impact the crime rates in the city.He calls the rate of conviction = "clear-up percentage". He uses data for 1972 and 1978. He runs the following model:

$$log(crime)_{it} = \beta_0 + \delta_0 d78_t + \beta_1 clrprc_{i,t-1} + \beta_2 clrprc_{i,t-2} + a_t + u_{it}$$

where, he calculates the lag $clrprc_{i,t-2}$ for the year 78 as $= clrprc_{i,78} - clrprc_{i,76}$

13.3.1 Organizing Panel Data



Naturally, the organization of the data set depends on how you want to use it. For reading a panel data set, wide form is better. But for analysis, often long form is more conducive.