ANALYSIS OF PANEL DATA

An Introduction

datascience@berkeley

Unobserved Effect Models
Pooled OLS
First-Difference Models

Unobserved City Effect in Our Example

With this set up, a simple unobserved effect model for city crime rates for 1982 and 1987 is

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

where $d87$ is an indicator variable for year 1987, and i denotes cities.

In this example, a_i is called an **unobserved city effect** or **city fixed effect**.

Any city-specific features, such as geographical features, that do not change over *the observed time period* and are not observed (by the analysts) are included in a_i. We have to emphasize *constant only observed time period* because every geographical features are not constant forever. Also pay attention to the subscripts being use: a_i only varies across cross-sectional units (and not over time) but constant within each of the cross-sectional unit.

Pooled OLS Applied to the Crime Rate Example

One method is to "*pool*" the two years and use **OLS**. The major drawback of this approach is that the pooled OLS requires that the observed effect a_i and the observed explanatory variable, x_{it}, be *uncorrelated* in order to produce a *consistent* estimator for β_1.

Writing the model using a *composite error* form, we have

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

where $\mu_{it} = a_i + \epsilon_{it}$.

- OLS requires that μ_{it} be uncorrelated with x_{it}. While it may be a reasonable assumption in cross-sectional regression, this assumption is not likely to hold because the same cross-sectional units are observed multiple times.
- Even if ϵ_{it} is uncorrelated, the pooled OLS is likely to be *biased* and *inconsistent* when a_i and x_{it} are correlated. This kind of bias is called *heterogeneity bias*.
- The term *heterogeneity bias* here is referred to the bias is really caused by omitting individual-specific, time-invariant variables.

Pooled OLS Applied to the Crime Rate Example

As you can see in the example above, pooled OLS without even the time indicator variable $d87$ produces unreasonable results in addition to violating the correlation assumption and potentially suffering from omitted variables.

Let's try another pooled OLS model:

```
> pooled.ols.fit <- lm(crmrte ~ d87+unem, data=crime2)
> summary(pooled.ols.fit)

Call:
lm(formula = crmrte ~ d87 + unem, 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 ***
d87           7.9404     7.9753   0.996    0.322
unem          0.4265     1.1883   0.359    0.720
---
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
```

Estimated Model Using Pooled OLS

$$\widehat{crmrte} = 93.42 + 7.94 d87_t + 0.427 unem$$

where $n = 92$ and $R^2 = 0.012$

We drop the subscripts when reporting the estimated model.

- Although the estimated effect of unemployment rate on crime rate directionally makes sense, it is both economically and statistically insignificant.
- The model also does not explain the crime rate well.
- Most importantly, the standard errors and test statistics in this model are incorrect due to the serial correlation caused by the repeated observations. This is a point mentioned before.

The First-Differencing Approach

- Panel data statistical models face this problem directly without making unreasonable assumption regarding the absence of correlation between the individual heterogeneity and the explanatory variables.
- In fact, this kind of models allow for the unobserved effect, a_i, to be correlated with explantory variables.
- In our example, we would like to allow for correlation between the unobserved (to the data scientists) city variables that affect crime rate and the observed explanatory variables, such as unemployment rate.
- For unobserved effect, a_i, that is individual-specific and remain constant over the observed time period, we can use differencing to eliminate the unobservables while estimating the effect of interest.

The First-Differencing Approach

For example,

$$y_{i2} = (\beta_0 + \delta_0) + \beta_1 x_{i2} + a_i + u_{i2} \quad (t = 2)$$
$$y_{i1} = \beta_0 + \beta_1 x_{i1} + a_i + u_{i1} \quad (t = 1).$$

Substracting the second equation from the first, we get

$$(y_{i2} - y_{i1}) = \delta_0 + \beta_1 (x_{i2} - x_{i1}) + (u_{i2} - u_{i1})$$

or

$\triangle y_i = \delta_0 + \beta_1 \triangle x_i + \triangle u_i$ where $\triangle$ denotes the change from $t = 1$ to $t = 2$.

Note that in the "change" equation, the unobserved effect is "*differenced*" away.

The First-Differencing Approach

- The "first-difference" equation is very simple in that it is a cross-sectional equation in which each of the variable is differenced over two consecutive time periods.

- Importantly, we can estimate this model and conduct inference using the estimated models by OLS regression techniques, provided that the underlying assumptions are satisfied.

- Specifically, this model requires that $\triangle u_i$ is uncorrelated with $\triangle x_i$.

- This assumption would hold if the error term, u_{it}, is uncorrelated with the explanatory variables in *both* time periods. This is the version of **strict exogenity** assumption.

The First-Differencing Approach

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

- The differencing to eliminate time-invariant unobservable effects has a substantial inference on the results (relative to the OLS models estimated above). Each 1% change in unemployment rate is associated with an average of a 2.2 increase in crime rate, measured by the number of crimes per $1,000$ residents.

The First-Differencing Approach

- Differencing in this case also makes intuitive sense because instead of estimating a cross-sectional relationship, which possibly suffers from omitted variables, as it models directly how changes in the explanatory variables over time (in this toy example, unemployement rate) affects the change in y (in this case, crime rate) over the same time period.

- One has to remember that this approach will not work for explanatory variables that are "" over the observable time period of interest as they will be differenced out along with the time invariant unobservables.

constant

Berkeley SCHOOL OF INFORMATION