Panel Data Analysis on Grunfeld Data

Kaustav Khatua

Roll No. 181075

Introduction

In Regression or Time Series Analysis we consider only one aspect, ie. either cross-sectional or time aspect. But, in real life many cases arise where considering only one may not be sufficient. Panel Data is such type of example. Here generally we have **N** individuals and for every individual we have data for **T** (T $\geq$ 2) time points. So, applying concepts of Regression or Time Series analysis alone will not produce good result. Panel Data analysis is the method which we should use in these cases. Here I have applied concepts of Panal Data analysis on **Grunfeld Data**.

About Grunfeld Data

It is a blanced and long panel data. Here cross-sectional units are General Motors, US Steel, General Electric, Chrysler, etc. 11 such companies and for every company data of 20 years (1935 - 1954) are given. Our goal is to predict **Gross invest**(Y) on the basis of **Market value**(X_1) and **Capital**(X_2). All the measurements are in 1947 dollars.

Analysing the Data

We want a model which looks overally like this,

$$Y_{it} = \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \epsilon_{it} \quad \cdots (1)$$
$$i = 1, 2, 3, ..., 11$$
$$t = 1, 2, 3, ..., 20$$

where i denotes ith company and t denotes tth year. Now depending upon the assumption we make on intercept, slope coefficients and error term we get different models. We generally consider three main models; Pooled model, Fixed Effects model and Random Effects model.

Pooled Model

Here we assume that all coefficients are same for all the companies, they are time invariant and error term captures time and cross-sectional effect, ie. we ignore the individual and time dimension of the panel data and do usual **OLS** estimation. Advantage of this model is it is simple and easy to fit but if cross section or time has influence on the data then this model will not perform well.

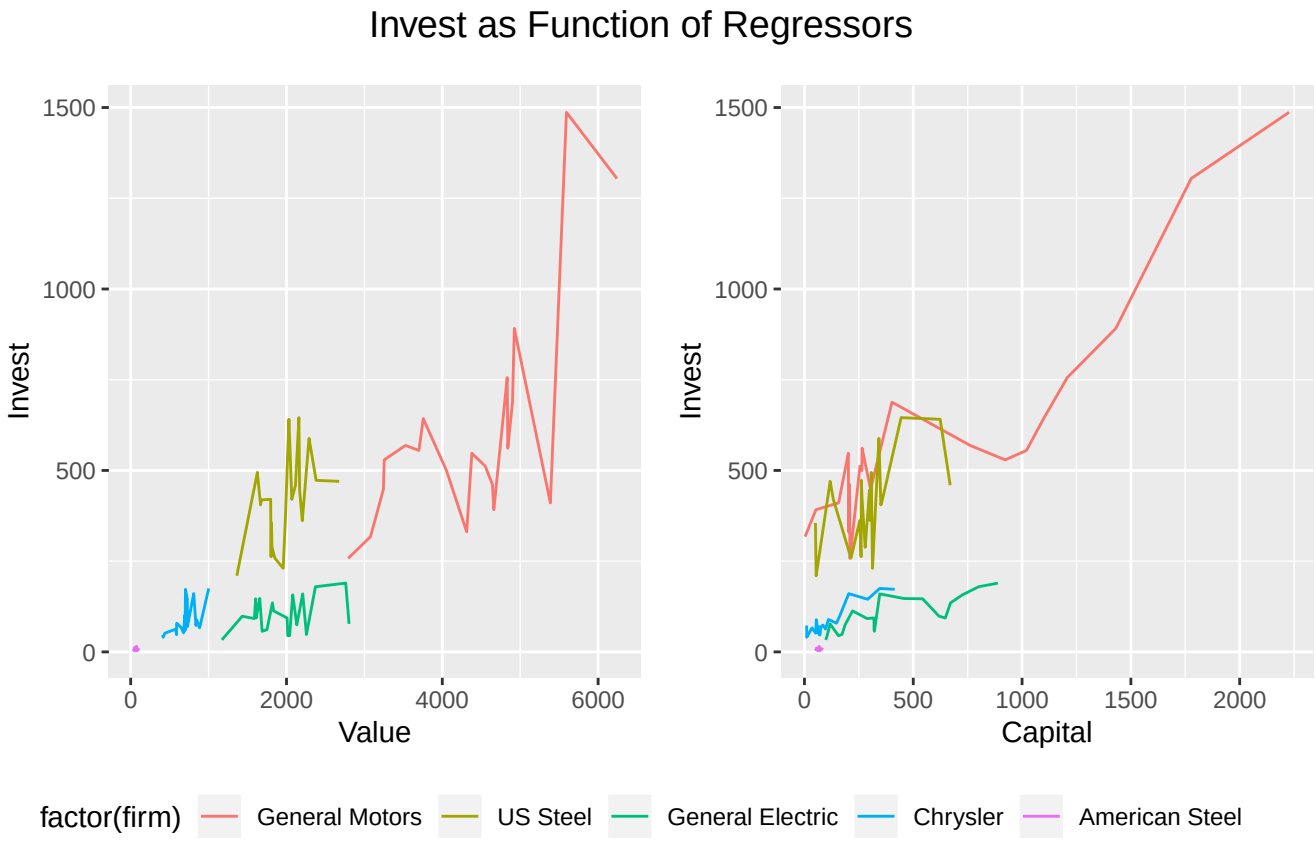
In our case summary of the Pooled model is:

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -38.410054   8.413371  -4.565 8.35e-06 ***
## value         0.114534   0.005519  20.753  < 2e-16 ***
## capital       0.227514   0.024228   9.390  < 2e-16 ***

## Residual standard error: 90.28 on 217 degrees of freedom
## Multiple R-squared:  0.8179, Adjusted R-squared:  0.8162
## F-statistic: 487.3 on 2 and 217 DF,  p-value: < 2.2e-16
```

Every coefficient is significant and p-value of the fit is small indicating that the model is significant overally. It may happen that the data really does not show much time or individual effect, as a result Pooled model is performing well. But when analysing panel data we don't accept the Pooled Model without checking other aspects.

First we check whether distribution of dependent variable is different for different companies. We may get some intuition from the following picture.



To avoid crowdedness plot is shown for only five companies.

In the capital graph when capital is near zero, invest is much higher for General Motors (red line) than Chrysler (blue line), ie. in a linear model General Motors will have bigger intercept than Chrysler. Also, distribution of invest is very different for American Steel than other companies. It is an indication that Pooled Model may not be sufficient and we have to take cross sectional effect into consideration.

Now we may check whether time also has influence on the data or not. One way to check that is checking for autocorrelation, as autocorrelation measures the relationship between a variable and a lagged version of itself over various time intervals. Durbin Watson test checks for autocorrelation by testing whether the errors from a model forms an **AR(1) process**, ie. $\epsilon_{\mathbf{it}} = \rho\epsilon_{\mathbf{i,t-1}} + \mathbf{z_{it}}$, $|\rho| < 1$. Here, $\mathbf{H_0} : \rho = \mathbf{0}$ and $\mathbf{H_1} : \rho \neq \mathbf{0}$ and test statistic is of the form:

$$d = \frac{\sum_{i=1}^{N}\sum_{t=2}^{T}(\hat{\epsilon}_{i,t} - \hat{\epsilon}_{i,t-1})^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\epsilon}_{i,t}^2}$$

where, $\hat{\epsilon}_{i,t}$ is residual from the model for tth observation of ith cross section. If value of the test statistic is near 2 then we can expect that autocorrelation is not present in the data. One thing may be noted that autocorrelation can be a result of model misspecification.

For our case the result of the DW Test is as follows:

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  invest ~ value + capital
## DW = 0.35666, p-value < 2.2e-16
## alternative hypothesis: serial correlation in idiosyncratic errors
```

From the result we can see that, autocorrelation(serial correlation) is present.

It is clear that time and cross sectional effects may not be negligible here and we have to respecify the model(1). There are several possibilities. We explore them one by one.

Fixed Effects Model

Here we assume that all the coefficients are time invariant, slope coefficients are same for all the companies but intercept is different for different companies. The model is of the form,

$$Y_{it} = \beta_{0i} + \beta_1 X_{1it} + \beta_2 X_{2it} + \epsilon_{it}$$

which is equivalent to,

$$Y_{it} = \sum_{j=1}^{11} \alpha_j I_j + \beta_1 X_{1it} + \beta_2 X_{2it} + \epsilon_{it}$$

where, $I_j = 1$ if $j = i$, 0 otherwise. The first equation describes the idea but to fit model we have to use the second equation ie. we have to use **dummy variable technique**. Note that intercept is dropped from the model to avoid multicollinearity problem. We can also use 10 dummy variables with an intercept term.

Summary of the Fixed Effects model in our case:

```
##                          Estimate Std. Error t value Pr(>|t|)
## value                     0.11013    0.01130   9.746  < 2e-16 ***
## capital                   0.31003    0.01654  18.744  < 2e-16 ***
## firmGeneral Motors      -70.29907   47.37535  -1.484   0.1394
## firmUS Steel            101.90474   23.76871   4.287 2.77e-05 ***
## firmGeneral Electric   -235.56939   23.28607 -10.116  < 2e-16 ***
## firmChrysler            -27.80911   13.41858  -2.072   0.0395 *
## firmAtlantic Refining  -114.60252   13.50246  -8.488 4.06e-15 ***
## firmIBM                 -23.16020   12.07589  -1.918   0.0565 .
## firmUnion Oil           -66.54422   12.24204  -5.436 1.53e-07 ***
## firmWestinghouse        -57.54649   13.33791  -4.315 2.48e-05 ***
## firmGoodyear            -87.21454   12.28873  -7.097 1.99e-11 ***
## firmDiamond Match        -6.56803   11.27363  -0.583   0.5608
## firmAmerican Steel      -20.57820   11.29779  -1.821   0.0700 .
```

```
## Residual standard error: 50.3 on 207 degrees of freedom
## Multiple R-squared:  0.9616, Adjusted R-squared:  0.9591
## F-statistic: 398.2 on 13 and 207 DF,  p-value: < 2.2e-16
```

Nine out of eleven dummy variables are significant and the p-value of the fit is small. Adjusted R-Square is high (0.95 may be suspicious, but here we are considering many important regressors, as a result it is quite high).

Results from Durbin-Watson test are as follows:

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  invest ~ value + capital
## DW = 1.0788, p-value = 3.634e-12
## alternative hypothesis: serial correlation in idiosyncratic errors
```

Durbin-Watson test statistic is much higher than Pooled model.

From the above two results we may conclude that different companies have different strategy to invest and by varying the intercepts over companies we are able to explain that to some extent. One next natural choice can be **varying the slope coefficients** also. But it will include 22 more variables in the model. Even in the basic Fixed Effects Model we have to include a coefficient for every cross scetion. If the number of cross section is large then we have to estimate a **large number of coefficients**. To avoid this we use Random Effecs Model.

Random Effects Model

Here also the model is of the form,

$$Y_{it} = \beta_{0i} + \beta_1 X_{1it} + \beta_2 x_{2it} + \epsilon_{it}$$

but the difference is that here β_{0i} is a random variable with mean β_0, ie.

$$\beta_{0i} = \beta_0 + u_i \qquad i = 1, 2, 3, ..., N$$

where, u_i is a random error with mean 0 and variance σ_u^2. So, complete form of the Fixed Effects Model is,

$$\begin{aligned} Y_{it} &= \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + \epsilon_{it} + u_i \\ &= \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + w_{it} \end{aligned}$$

where, $w_{it} = \epsilon_{it} + u_i$. Now we have to estimate σ_u^2 additionally.

The additional assumptions of Random Effects Model are,

$$u_i \sim N(0, \sigma_u^2)$$
$$E(u_i \epsilon_{it}) = E(u_i \epsilon_{jt}) = E(u_i u_j) = 0 \qquad (i \neq j)$$

Due to the assumptions,

$$cov(w_{it}, w_{is}) = cov(\epsilon_{it} + u_i, \epsilon_{is} + u_i) = var(u_i) = \sigma_u^2$$

So, the Random Effects Model is,

$$\begin{aligned} Y_{it} &= \beta_0 + \beta_1 X_{1it} + \beta_2 X_{2it} + w_{it} \\ E(w_{it}) &= 0 \\ var(w_{it}) &= \sigma_\epsilon^2 + \sigma_u^2 \\ cov(w_{it}, w_{is}) &= \sigma_u^2 \end{aligned}$$

So, for a given cross-section errors are correlated. Due to this fact applying OLS for the original model will give inefficient estimators. To get efficient estimators we have to apply GLS, ie. we have to transform the model using covariance matrix of w_{it}, instead we can apply the **FGLS**. FGLS method is described below.

How Unknown Parameters of Random Effects Model are Estimated

Pooled and Fixed Effects Model coefficients are easy to estimate but for Random Effects Model it is not straight forward. In Random Effects Model we have to estimate two variances and the model coefficients. Here instead of applying OLS to estimate the coefficients of the original model we apply OLS on **partial demeaned data**, where partial demeaning is:

$$Y_{it} - \theta \bar{Y_i} = \beta_0(1-\theta) + \beta_1(X_{1it} - \theta \bar{X_{1i}}) + \beta_2(X_{2it} - \theta \bar{X_{2i}}) + (\epsilon_{it} - \theta \bar{\epsilon_i})$$

where, $\theta = 1 - [\sigma_\epsilon^2 / (\sigma_\epsilon^2 + T\sigma_u^2)]^{1/2}$ and $\bar{Y_i} = \sum_{t=1}^{T} Y_{it} \ / \ T$ and $\bar{X_i} = \sum_{t=1}^{T} X_{it} \ / \ T$.
This transformation is also called **time-demeaned** transformation as it removes time component from the data.

We first estimate σ_ϵ^2 and σ_u^2 by,

$$\hat{\sigma_\epsilon^2} = s_{FE}^2 = \frac{RSS\ of\ Fixed\ Effects\ Model}{NT-N-K}$$

and,

$$\hat{\sigma_u^2} = s_{Pooled}^2 - \hat{\sigma_\epsilon^2} = \frac{RSS\ of\ Pooled\ Model}{NT-K-1} - \hat{\sigma_\epsilon^2}$$

Then using these we estimate θ. Note that the denominator in $\hat{\sigma_u^2}$ may be $NT - N - K$. K is number of regressors, in our case it is 2. Then we substitute $\hat{\theta}$ in the partial demeaned model and obtain the coefficient estimates of the model by applying OLSE.

Results from Random Effects Model:

```
[1] Effects:

[1]                    var std.dev share
[3] idiosyncratic 2530.04   50.30  0.29
[5] individual    6201.93   78.75  0.71
[7] theta: 0.8586

[1]
[2] Coefficients:
[3]                Estimate  Std. Error z-value Pr(>|z|)
[4] (Intercept) -53.9436014  25.6969760 -2.0992   0.0358 *
[5] value         0.1093053   0.0099138 11.0256   <2e-16 ***
[6] capital       0.3080360   0.0163873 18.7972   <2e-16 ***

[1] Total Sum of Squares:    2393800
[2] Residual Sum of Squares: 550610
[3] R-Squared:      0.76999
[4] Adj. R-Squared: 0.76787
[5] Chisq: 726.428 on 2 DF, p-value: < 2.22e-16
```

Idiosyncratic Variance and individual variance is referring to σ_ϵ^2 and σ_u^2. The estimate for these are 2530.04 and 6201.93 and the estimate of theta is 0.8586. The coefficient estimates are shown in table.

Results of Durbin-Watson Test is as follows:

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  invest ~ value + capital
## DW = 0.99236, p-value = 1.716e-14
## alternative hypothesis: serial correlation in idiosyncratic errors
```

So, autocorrelation is present and we have to handle this issue. But first we can choose the best among Pooled, Random Effects and Fixed Effects Model.

Note: From the summary table we see that R-Squared has decreased significantly. But it is not the case. To estimate the Random Effects Model **plm** function is used, which calculates R-Squared using following formula.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\epsilon}_{it}^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y_i})^2} \quad , where, \qquad \bar{y_i} = \frac{1}{T}\sum_{t=1}^{T} y_{it}$$

But, to calculate Pooled and Fixed Effects Model, **lm** function is used, which calculates, R-Squared using following formula.

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\sum_{t=1}^{T}\hat{\epsilon}_{it}^2}{\sum_{i=1}^{N}\sum_{t=1}^{T}(y_{it}-\bar{y})^2} \quad , where, \qquad \bar{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$$

If we calculate R-Squared for Random Effects Model using the second formula, then it will be 0.94. So, we don't have to worry about R-Squared.

Choosing Between Pooled, Fixed Effects and Random Effects Model

If cross sectional effect is present in the data then Fixed and Random Effects Model try to consider this effect by varying slope coefficients for cross sections. If this unobserved cross sectional effect is small then Pooled Model will yield efficient estimators. Now if there is cross sectional effect but it is **uncorrelated with regressors** then we can use **Random Effects Model** but if the **correlation is not 0** then we have to use **Fixed Effects Model**.

Choosing Between Pooled and Fixed Effects Model

Pooled Model can be viewed as a restricted form of Fixed Effects Model. When all the dummy variable coefficients, α_j are same then we will get Pooled model. So if we test,

$$H_0 : \alpha_1 = ... = \alpha_{11} \quad (ie.\ \alpha_j - \alpha_l = 0, j \neq l)$$
$$\text{vs} \qquad H_1 : \text{at least one } \alpha_j \text{ is different.}$$

then we will get some intuition which model is appropriate.

Test statistic is,

$$F = \frac{\frac{SS_{Res}(Pooled) - SS_{Res}(FE)}{N-1}}{\frac{SS_{Res}(FE)}{TN-K-N}} \quad \overset{H_0}{\sim} \quad F_{N-1\ ,\ TN-K-N}$$

we reject H_0 if observed value of $F > F_{\alpha\ ,\ N-1,\ TN-K-N}$, where α is level of significance and K is number of regressors. For our case the test statistic value is $49.20708 > F_{0.05,10,207}$. So, we reject the null hypothesis and conclude that, in our case Fixed Effects Model is more appropriate than Pooled Model.

Choosing Between Fixed Effects and Random Effects Model

We have to use **Hausman Test** to determine the better model among Fixed and Random Effects Model. Here we test,

$$H_0 : \text{Correlation between unobserved cross sectional effect and regressors is 0.}$$

against,
$$H_1 : H_0 \text{ is not true.}$$

Here test statistic is,

$$W \ = \ (\ \hat{\beta}_{FE} \ - \ \hat{\beta}_{RE})^T \ \hat{\Psi}^{-1} \ (\ \hat{\beta}_{FE} \ - \ \hat{\beta}_{RE}\ )$$

β is the slope coefficients and

$$\Psi = var(\ \hat{\beta}_{FE} \ - \hat{\beta}_{RE}\ ) = var(\ \hat{\beta}_{FE}\ ) - var(\ \hat{\beta}_{RE}\ ).$$

Under the null hypothesis, W has a limiting chi-squared distribution with k degrees of freedom.

Results of Hausman Test in our case:

```
##
##  Hausman Test
##
## data:  invest ~ value + capital
## chisq = 3.9675, df = 2, p-value = 0.1376
## alternative hypothesis: one model is inconsistent
```

So, between Fixed and Random Effects Model Random Effects Model is more appropriate in our case. So, we choose **Random Effects Model** among Pooled, Fixed Effects and Random Effects Model.

Handling Presence of Autocorrelation

We can observe that Fixed and Random Effects Model have much higher Durbin-Watson Test statistic value than the Pooled Model. So, considering the cross sectional effects by varying the intercept terms is better than ignoring the cross-sectional effect completely. Now varying the slope coefficients along with intercepts for different companies, may give better results, but then we have to estimate a large number of coefficients. So, we drop the idea.

All the three models had low DW Test statistic value. So, we have to fit different model to take care of it. We can start with the following assumptions,

$$\epsilon_{it} = \rho_i \, \epsilon_{i,t-1} \;\; + \;\; z_{it} \; , \qquad | \, \rho \, | < 1$$

First we fit OLS to ith cross section, obtain residuals and estimate, ρ by,

$$\hat{\rho} = r_i = \frac{\sum_{t=2}^{T} \hat{\epsilon}_{it} \hat{\epsilon}_{i,t-1}}{\sum_{t=1}^{T} \hat{\epsilon}_{it}^2}$$

Now, using $\hat{\rho} = r_i$ we do the following transformation (**Prais-Winsten transformation**),

$$y_{i1}^* = \sqrt{1 - r_i^2} \, y_{i1} \qquad and \qquad y_{ij}^* = y_{ij} \; - r_i \, y_{i,j-1} \quad j = 2, ..., T$$

$$x_{i1}^* = \sqrt{1 - r_i^2} \, x_{i1} \qquad and \qquad x_{ij}^* = x_{ij} \; - r_i \, x_{i,j-1} \quad j = 2, ..., T$$

This transformation removes autocorrelation. Now we may apply Random Effects Model on this transformed data.

Results of fitting Random Effects Model on this transformed data is as follows:

```
## [1]
```

```
## [1] Effects:
## [3]                    var std.dev share
## [5] idiosyncratic 1679.33   40.98 0.441
## [7] individual    2127.27   46.12 0.559
```

```
## [1] -122.3391  -13.9861    3.5228   11.0326  194.8439
## [2]
## [3] Coefficients:
## [4]                Estimate  Std. Error z-value Pr(>|z|)
## [5] (Intercept) -27.2420608  15.4028361 -1.7686  0.07695 .
## [6] value         0.0972026   0.0087479 11.1115  < 2e-16 ***
## [7] capital       0.2859625   0.0208663 13.7045  < 2e-16 ***
```

```
## [1]                                  Total Sum of Squares:    963040
```

Result from Durbin-Watson test is as follows:

```
##
##  Durbin-Watson test for serial correlation in panel models
##
## data:  invest ~ value + capital
## DW = 1.2527, p-value = 9.355e-09
## alternative hypothesis: serial correlation in idiosyncratic errors
```

All the coefficients are significant and the Durbin-Watson test statistic value is higher compared to the model applied on the original data. It is our final model.

Way for Further Analysis

Even after applying Prais-Winsten transformation, we could not get rid of autocorrelation completely. So, further analysis in this direction may be conducted. Also, we can conduct Dynamic Panel Data Analysis on this data, where we include one or more lagged dependent variable in the model.