Introduction and Background

Introduction and Background

- Introduction to panel data methods
- Issues with the OLS model in the presence of panel data
- Panel data: Least squares dummy variable (LSDV) method
- Panel data: First difference (FD) estimators
- Panel data: Fixed effects (FE) estimators
- Panel data: Random effects (RE) estimators
- Panel data: FE vs. RE estimators
- Summary and concluding remarks

Introduction to Panel Data Methods

Introduction to Panel Data Methods

Relationship between security returns r_{it} and order imbalance OIB_{it}

- Here $\mathrm{OIB_{it}} = \frac{Buy_{Volume} - Sell_{Volume}}{Buy_{Volume} + Sell_{Volume}}$
- $r_{it} = a_0 + a_1 OIB_{it} + v_t + \alpha_i + \mu_{it}$
- Assume 10 years and 100 securities
- $v_t + \alpha_i + \mu_{it}$ are our error terms; let us discuss them one by one
- v_t ('*t*' from 1…T) is solely time dependent term, e.g., broad market-wide changes
- These time-dependent terms don't vary across the city, and can be accounted for by 'n-1' (10-1 = 9) dummy variables [i.e., least square dummy variable estimation]

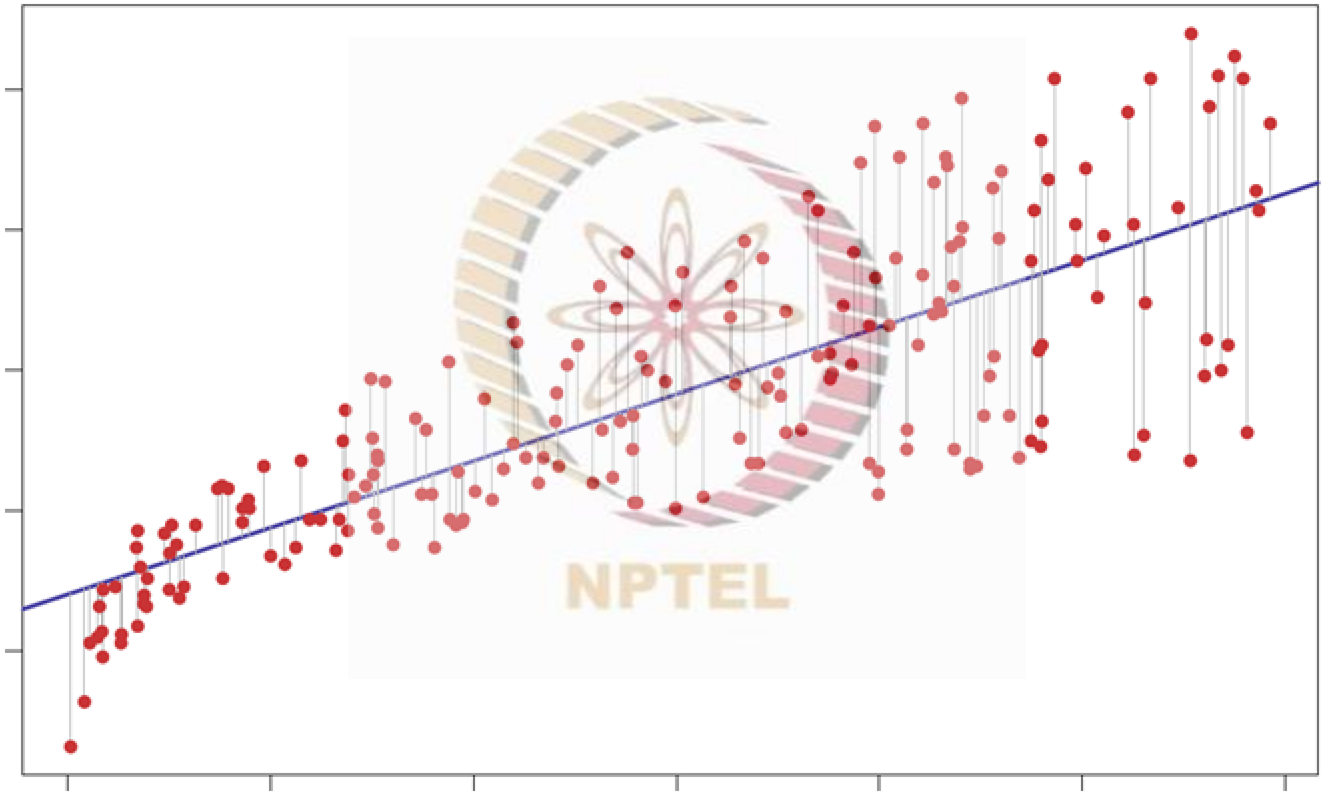
Introduction to Panel Data Methods

Relationship between security returns r_{it} and order imbalance $\mathbf{OIB_{it}}$

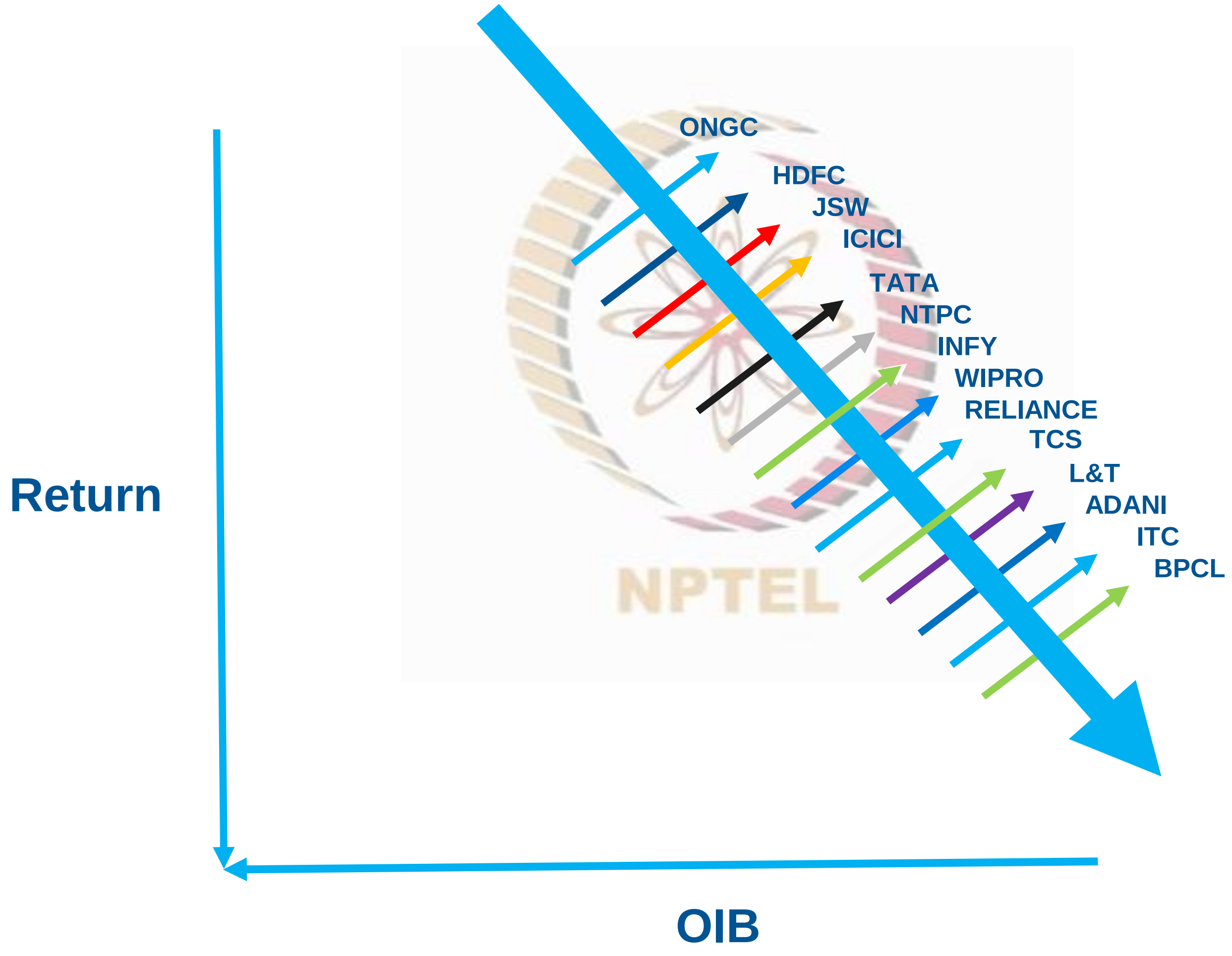
- $r_{it} = a_0 + a_1 OIB_{it} + v_t + \alpha_i + \mu_{it}$
- α_i ('*i*' from 1…n) is the security-specific variable like firm size, firm beta, industry, etc., and not changing frequently overtime
- Usually, T: number of periods is small, and N: number of individual entities is large
- So, accounting for α_i through dummy variable method can make the model extremely inefficient
- So, what if we do not account for this α_i

Issues with Ordinary Least Square (OLS) Model

Fitting OLS Through Scattered Data Points



Fitting OLS with Panel Data



Pooled OLS Estimation with Panel Data

Relationship between security returns r_{it} and order imbalance OIB_{it}

- $r_{it} = a_0 + a_1 OIB_{it} + v_t + \alpha_i + \mu_{it}$
- $n_{it} = \alpha_i + \mu_{it}$ [α_i: Unobserved heterogeneity]
- Cov(n_{it}, OIB_{it}) ≠ 0 [Problem of endogeneity]
- Cov(n_{it},n_{it+1})=Cov(u_{it} +α_i,u_{it+1} +α_i) ≠ 0 [Problem of autocorrelation]
- Pooled OLS estimates will be biased and inconsistent

Lease Squares Dummy Variable (LSDV) Estimators

LSDV Estimators

Relationship between security returns r_{it} and order imbalance OIB_{it}

- Assuming that time-varying effects can be modeled using time-dummies
- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$ (1)
- Include 'N-1' dummy variable for 'N' securities $(S_2, S_3, \ldots, S_N)$ as follows
- $r_{it} = a_0 + a_1 OIB_{it} + \sum_{n=2}^{N} a_n S_n + \mu_{it}$ (2)
- Here, S_2 is a dummy variable that takes a value of 1 for security 2, and 0 otherwise; and so on for securities 3, 4,…, N
- Thus, we are explicitly accounting for the unobserved heterogeneity for each security individually

LSDV Estimators

$$r_{it} = a_0 + a_1 OIB_{it} + \sum_{n=2}^{N} a_n S_n + \mu_{it}$$

- The estimates of a_n are consistent under the following conditions
- Cov(u_{it}, OIB_{it}) = 0 ; no serial correlation in errors; homoscedasticity in error terms
- Theoretically, under these assumptions, the estimates from LSDV are the same as fixed-effects (FE) estimate
- However, dummy variables allow estimating this α_i explicitly, unlike FE estimators
- However, the model is not parsimonious with large N

First Difference (FD) Estimators

First Differences Estimators

Relationship between security returns r_{it} and order imbalance OIB_{it}

- Assuming that time-varying effects can be modeled using time-dummies
- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$ (1)
- $r_{it-1} = a_0 + a_1 OIB_{it-1} + \alpha_i + \mu_{it-1}$ (2)
- Subtract (1) - (2)
- $\Delta r_{it} = a_1 \Delta OIB_{it} + \Delta \mu_{it}$ (3)
- Cov($\Delta \mu_{it}$,ΔOIB_{it}) =0
- This model can be estimated with OLS estimation

First Differences (FD) Estimators

Relationship between security returns r_{it} and order imbalance OIB_{it}

- $\Delta r_{it} = a_1 \Delta OIB_{it} + \Delta\mu_{it}$ (3)
- $Cov(\Delta u_{it}, \Delta u_{it-1}) = Cov(u_{it} - u_{it-1}, u_{it-1} - u_{it-2})$
- However, these is an issue of auto-correlation due to first differencing
- Differencing leads to small variation in variables and therefore considerable increase in standard error of estimates
- Loss of observation
- Time independent factors can not be estimated

First Differences (FD) Estimators

Relationship between security returns r_{it} and order imbalance OIB_{it}

- $\Delta r_{it} = a_1 \Delta OIB_{it} + \Delta \mu_{it}$ (3)
- All those terms with no variance across time will be eliminated; so, we need the dependent and independent variables to have some variation across time and city

Fixed-Effects (FE) Estimator

Fixed-Effects Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \boldsymbol{\alpha_i} + \mu_{\boldsymbol{it}}$ (1)
- Time-demean equation (1) $\frac{1}{T}\sum_{t=1}^{T} r_{it}$ ∀ i's = 1, 2, 3…N
- $\overline{r_i} = a_0 + a_1 \overline{OIB_i} + \boldsymbol{\alpha_i} + \overline{u_i}$ (2)
- Substract (1) - (2)
- r_{it}-$\overline{r_i}$ $= a_1(OIB_{it}$-$\overline{OIB_i})$+$(\mu_{\boldsymbol{it}}$-$\overline{u_i})$
- $\tilde{r}_{it} = a_1 * \widetilde{OIB}_{it} + U_{it};$
- Here, Cov($\widetilde{OIB}_{it}, U_{it}$) = 0, and pooled OLS estimates will be consistent

Fixed-Effects Estimators

- $\tilde{r}_{it} = a_1 * \widetilde{OIB}_{it} + U_{it}$;
- Fixed effects remove any time-constant terms
- Fixed effects are costly (due to transformation of original data)

Fixed-Effects (FE) vs.
First Difference (FD) Estimators

Fixed-Effects vs. First Difference Estimators

- $\tilde{r}_{it} = a_1 * \widetilde{OIB}_{it} + U_{it}$;
- For T = 2, FD = FE
- For T > 2, FD ≠ FE
- With the assumptions that (a) large sample N→ ∞(b) error term (μ_{it}) is uncorrelated with the independent variable (e.g., OIB_{it}) (c) sample is random, and (d) sufficient variance in variables, the following is held
- $E\left[\widehat{a_1}_{FD}\right] = E\left[\widehat{a_1}_{FE}\right] = a_1$(both FE and FD estimates of a_1 are unbiased)
- $\widehat{a_1}_{FD} \xrightarrow{p} \beta$; $\widehat{a_1}_{FE} \xrightarrow{p} a_1$(both the estimators are consistent)

Fixed-Effects vs. First Difference Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$
- (A) $Cov(\mu_{it}, \mu_{it-1}) = 0$: error terms ($\mu_{it}$) are serially uncorrelated
- First differencing introducing serial correlation in error terms
- Due to this, the standard error of estimates for FE estimators are lower (more efficient) than FD estimators: $se(\hat{a}_{1_{FE}}) < se(\hat{a}_{1_{FD}})$
- (B) $\mu_{it} = \mu_{it-1} + e_{it}$: i.e., $\Delta\mu_{it-1} = e_{it}$
- Random walk structure in error term or strong autocorrelation in errors
- $se(\hat{a}_{1_{FE}}) > se(\hat{a}_{1_{FD}})$

Fixed-Effects vs. First Difference Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$
- (C) $\mu_{it} = \rho\mu_{it-1} + e_{it}$: AR(1) structure in error terms
- ρ is close to '1,' then the FD estimator is more efficient
- ρ is close to '0,' then the FE estimator is more efficient
- One solution is to examine the autocorrelation structure in FD errors
- If FD errors have a negative autocorrelation, that indicates original errors have no autocorrelation; hence FE is more appropriate
- If FD errors have a very small correlation, that indicates original errors have random walk; hence FD estimator is more appropriate

Fixed-Effects vs. First Difference Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$

One solution is to examine the autocorrelation structure in FD errors

- For scenarios in between, one can estimate both FD and FE and compare

For non-stationary process, first differences are more useful

For small sample sizes, FD is more appropriate

For data with large time dimension FE estimators are more appropriate

Random Effects Estimator: Part 1

Random Effects (RE) Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$
- Recall that the model would have an issue of endogeneity if the unobserved heterogeneity (α_i) is correlated with one of the independent variables: $Cov(OIB_{it}, \alpha_i) \neq 0$
- Thus, pooled OLS is not effective, and we used FD/FE methods to remove α_i from the model
- However, if $Cov(OIB_{it}, \alpha_i)$ is reasonably close to '0' then, we need not apply FD/FE as they involve a heavy transformation in data
- E.g., FE leads to loss of observations (T-1 periods instead of T)

Random Effects (RE) Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it};$
- $Cov(OIB_{it}, \alpha_i) = 0$; is a reasonable assumption in following cases
  - All the relevant variables are accounted for
  - $\boldsymbol{\alpha_i}$ is very small relative to other variables
- In this scenario, pooled OLS provides consistent estimates
- However, the errors may still be serially correlated: $Cov(\alpha_i + \mu_{it}, \alpha_i + \mu_{is}) \neq 0$
- This serial correlation can be corrected through RE estimation without putting a heavy cost of data (as in FD/FE)
- RE is more efficient than Pooled OLS and FE

Random Effects (RE) Estimators

- If you believe that sufficient variables have been entered in the model and $Cov(OIB_{it}, \alpha_i) \neq 0$ [Problem of Endogeneity] has been resolved
- Then RE is better than FE and OLS
- $r_{it} - \lambda \overline{r_i} = a_0(1 - \lambda) + a_1(OIB_{it} - \lambda \overline{OIB_i}) + (n_{it} - \lambda \overline{n_i})$, where $n_{it} = \alpha_i + \mu_{it}$
- The above random effect is the *pooled* estimation of the above transformation
- $\lambda=0$, then RE ≈ Pooled OLS
- $\lambda=1$, then RE ≈ FE

Random Effects Estimator: Part 2

Random Effects (RE) Estimators

- r_{it}-$\lambda\overline{r_i} = a_0(1-\lambda) + a_1(OIB_{it}$- $\lambda\ \overline{OIB_i}$)+$(n_{it}$-$\lambda\overline{n_i})$, where $n_{it} = \alpha_i + \mu_{it}$
- Typically, 0≤ λ≤1, hence RE is somewhere between pooled OLS and FE
- What is λ?
- $\lambda = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2+T\sigma_a^2}\right)^{0.5}$; here σ_u^2 is the variance of error term, σ_a^2 is the variance of α_i
- If σ_a^2 = 0, then λ = 0; that is α_i is insignificant/not important: RE converges to pool
- $T\sigma_a^2$ >>> σ_u^2, $\lambda = 1$, RE converges to FE
- Thus, unlike FE (fully time-demean) RE is quasi time-demean
- RE also allows to estimate time-constant terms

Random Effects (RE) Estimators

- $r_{it} = a_0 + a_1 OIB_{it} + \alpha_i + \mu_{it}$ (1)
- $r_{it} - \lambda \overline{r_i} = a_0(1-\lambda) + a_1(OIB_{it} - \lambda \overline{OIB_i}) + (n_{it} - \lambda \overline{n_i})$ (2)
- where $n_{it} = \alpha_i + \mu_{it}$
- $\lambda = 1 - \left(\frac{\sigma_u^2}{\sigma_u^2 + T\sigma_a^2}\right)^{0.5}$

  1. First step is to estimate $\hat{\lambda}$: this requires estimation of Eq. (1) through FE or pooled OLS methods
  2. Then estimate the transformed Eq. (2) using $\hat{\lambda}$ using pooled OLS

- The combined system set-up is RE method of estimation

Random Effects Estimator: Part 3

Assumptions of RE

The following assumptions are made for RE estimators to be consistent, i.e., $\widehat{a_1}_{RE} \xrightarrow{p} a_1$ (as $N \to \infty$)

- $Cov(OIB_{it}, \alpha_i) = 0$
- Each cross section is randomly sampled
- $E[u_{it} | X_{it}, \alpha_i] = 0$
- No perfect multicollinearity
- The last three assumptions are applicable to FE/FD also
- Only the first assumption is specific to RE

Estimating Time Constant Variables with RE

Recall the transformed model

- r_{it}-$\lambda\overline{r_i} = a_0(1-\lambda) + a_1(OIB_{it}$- $\lambda\ \overline{OIB_i}$)+ n_{it}-$\lambda\overline{n_i}$)
- In this model let us assume a time constant term $Size_i$, then the resulting model
- $\mathrm{r_{it}}$-$\lambda\overline{\mathrm{r_i}} = a_0(1-\lambda) + a_1(\mathrm{OIB_{it}}$- $\lambda\ \overline{\mathrm{OIB_i}}) + a_2 * \mathrm{Size_i}(1$- $\lambda) + (\mathrm{n_{it}}$-$\lambda\overline{\mathrm{n_i}})$
- As long as λ≠0, we can estimate a_2, the effect of time constant variable $Size_i$
- However, for these estimates to remain consistent, the assumption pertaining to RE need to be held [e.g., $Cov(Size_i, OIB_{it})] = 0$

Fixed Effects (FE) vs. Random Effects (RE)

FE vs. RE

Cov(α_i, X_{it})=0

- Both FE and RE estimates are consistent
- SE (RE estimate) < SE (FE estimate): Efficiency
- RE effect estimation allows for the effect of time-constant variables on dependent variables (For FE, that is not possible)

Cov(α_i, X_{it})$\neq$ 0

- Only the FE estimate is consistent
- SE (RE estimate) < SE (FE estimate)
- Hausman test can be employed to select between the two

Hausman Test

Hausman test statistic tests this hypothesis

- Null H_0 => Cov($\boldsymbol{\alpha_i}$, X_{it}) $= 0$ We should be able to use RE
- Estimate W $= \dfrac{(\widehat{\beta_{FE}} - \widehat{\beta_{RE}})^2}{Var(\widehat{\beta_{FE}}) - Var(\widehat{\beta_{RE}})}$ is distributed as chi-square with one df
- If H_0 is true, the numerator is small (both estimates are consistent), but the denominator is large, the statistic W is close to 0: Fail to reject the null, use RE estimator
- If null is false, the numerator is large, W is away from zero [Cov($\boldsymbol{\alpha_i}$, X_{it})≠0]: reject the null, use fixed effect estimator
- Essentially this estimator compares consistency (in numerator) relative to efficiency (in denominator)

Summary and Concluding Remarks

Summary and Concluding Remarks

- As compared to simple cross-sectional or time-series data, panel data with longitudinal properties is rich in granularity, and the information it offers
- However, it also entails a number of issues related to the estimation
- One of the important issues is the role of unobserved heterogeneity
- That is, the individual-specific time-invariant effects
- Though there are also only time-varying effects, they can be easily modeled by applying ‘T-1’ time dummies for T time periods

Summary and Concluding Remarks

- Usually, there are many individual units as compared to time periods; therefore, accounting for these units explicitly through dummies can make the model extremely non-parsimonious
- One simple approach is to estimate such models using the FD method, which is simply differencing the series by one lag and then applying pooled OLS
- A more evolved FE approach is to estimate time-demeaned series with pooled OLS
- Both FD and FE methods, though useful, put a heavy cost on data due to the extreme nature of data transformation
- A less exacting approach is that of the RE method, which is a compromise between two extremes of FE and pooled OLS approach

Summary and Concluding Remarks

- RE method is more appropriate when the assumption that unobserved heterogeneity is not correlated with the dependent variable [Cov($\boldsymbol{\alpha_i}$, X_{it})=0] can be held
- Cov($\boldsymbol{\alpha_i}$, X_{it})=0: Both RE and FE are consistent by RE is more efficient
- Cov($\boldsymbol{\alpha_i}$, X_{it})≠ 0: Only FE is consistent
- The decision to select FE vs. RE is taken through Hausman test (HT) statistic
- HT statistic essentially is a test of gains in consistency at the cost of efficiency while choosing FE vs. RE method

Thanks!

Lesson 2: Panel Data: Application

Introduction

- Application of Panel data algorithm in prediction of security prices
- Prediction of Broad marketwide returns
- Data Visualization
- Pool vs. Panel models (Fixed and Random effects)
- Interpret the output from different models
- Examine the cross-sectional/serial correlation in errors
- Obtain coefficients using robust standard errors

Case Study: Prediction of Broad marketwide returns

Case Study: Index Return Prediction

- Broad market wide indices are known to reflect the growth of economy and are often correlated with macroeconomic factors such as GDP
- This strategy to invest in market-index based on forecasts related to factors such as GDP has become a very prevalent strategy known as factor investing
- However, this exercise may be vitiated by unobserved heterogeneities such as country specific factors

Case Study: Index Return Prediction

- In this case study, we will employ panel data methods to forecast the market index prices

Sr	Country	Year	GDP	Return
1	A	1990	21.01801	27.8%
2	A	1991	-21.3649	32.1%
3	A	1992	-16.2345	36.3%
4	A	1993	21.69623	24.6%
5	A	1994	21.82465	42.5%
6	A	1995	21.89562	47.7%
7	A	1996	21.73732	50.0%
8	A	1997	21.74277	5.2%
9	A	1998	21.94626	36.6%
10	A	1999	17.49863	39.6%
11	B	1990	-22.5041	-8.2%
12	B	1991	-20.3831	10.6%

Case Study: Index Return Prediction

- The data includes, information about the country, year, GDP, log scaled mean deviated form, and returns
- Using different panel data methods, the relationship between GDP and returns to be modeled
- The subsequent slides provide the problem statement

Case Study: Index Return Prediction

The following tasks need to be performed:

Visualize the data

- Plot year wise market returns for each country
- Plot year wise GDP for each country
- Using the box plot show the heterogeneity in GDP and returns across countries and years
- What do we infer

Case Study: Index Return Prediction

The following tasks need to be performed:

- Model the relationship of GDP with index returns using simple pooled OLS: what are the problems with this approach
- Through visual approach show the fitted line with actual data
- Follow the LSDV approach and model the relationship after adding countrywide dummies

Case Study: Index Return Prediction

The following tasks need to be performed:

- Convert the data into panel format with country and year as the data identifiers (index)
- Model the data using fixed effect, using individual, time, and both the effects
- Perform the tests of poolability to show whether these fixed effects are significant

Case Study: Index Return Prediction

The following tasks need to be performed:

- Using the modeled fixed effect object from the panel data, extract the time and individual fixed effects
- Next model the data with random effects, examine the output and comment on individual and idiosyncratic heterogeneity
- Comment whether the random effects transformation is closer to pool and fixed effect
- Conduct the tests to examine whether random or fixed effects method is more appropriate

Case Study: Index Return Prediction

The following tasks need to be performed:

- Examine the cross-sectional and serial correlation in errors for pool, fixed, and random effects method
- Using robust standard errors, estimate coefficients that are robust to autocorrelation and heteroscedasticity in the model

Summary and Concluding Remarks

Summary and Concluding Remarks

- Broad marketwide returns are modelled with GDP factor
- First, data is visualized to see the individual and time effects
- Fixed and random effects models are estimated
- We performed residual diagnostics for cross-sectional and serial correlation in errors
- Estimate coefficients using robust standard errors

Thanks!