GEORG-AUGUST-UNIVERSITÄT
GÖTTINGEN

Term Paper

Panel Data Analysis in Marketing

Georg-August University, Göttingen

Chair of Marketing, in particular consumer research

Lecturer:

Dr. Ossama Elshiewy

Dept. Chair of Marketing, in particular consumer research

Prepared By:

Hari Raghavendar Rao Bandari

Matriculation No. 11334055

A Report on

Panel Data Analysis in Marketing

Submitted in Partial Fulfillment for the requirements of the course M.WIWI-BWL.0134

In

M.Sc in Applied Computer Science

Submitted By:

Hari Raghavendar Rao Bandari

Matriculation No: 11334055

Under the Guidance of

Dr. Ossama Elshiewy

Dept. Chair of Marketing, in particular consumer research

Department of Economics

Chair of Marketing, in particular Consumer Research

Georg-August-Universität,Göttingen

10 July 2016

The Abstract

"Panel Data Analysis in Marketing". Therefore, "Panel data" is commonly recognized as cross-sectional time-series data analysis, which requires datasets of particular entities . Therefore, it reports the behavior of entities is noticed across time. These entities could be companies, food products, consumer behavior and so on. Therefore, applying data analysis in the marketing domain to acquire knowledge and statistical information about the particular products.

Consequently, course module is focused on consumer behavior and mixed-marketing modeling concepts in panel data analysis, which primarily targeted the consumer behavior. Hence, for analyzing the datasets it requires a software tool so "R language" is adopted and it delivers the large community of Open source. It provides reports of datasets in statistically along with graphical representation to attain insights of the behavior of consumer over certain products.

In conclusion, "panel data analysis in Marketing" plays major criteria to understand or predict the consumer behavior across time with the help of certain modeling concepts.

TABLE OF CONTENTS

1. Introduction to R: 5

2. Refreshment in Regression Analysis: 6

    2.1. Linear Regression(ordinary Least Squares): 6

        2.1.1 Simple Linear Regression: 6

        2.1.2 Multiple Linear Regression: 7

    2.2 Logarithmic Transformation in OLS: 7

    2.3 Elasticity: 7

    2.4 Exercise (With Marketing Topic): 8

3. Fixed Effect Model: 9

    3.1 One-way fixed effect Model: 10

    3.2 Two-way fixed effect Model: 11

    3.3 Model diagnostics for Fixed Effect model: 11

    3.4 HAC Covariance Matrix in FE model: 12

    3.5 Exercise (With Marketing Topic): 12

4. Two Stage Least Squares in Fixed Effects Models: 19

    4.1 Exercise (With Marketing Topic): 20

5. Random Effects Models: 22

    5.1 One-way Random Effect Model: 23

    5.2 Two-way Random Effect Model: 23

    5.3 Model diagnostics for Random Effect model: 24

    5.4 Exercise (With Marketing Topic): 24

6. Dynamic Panel Models: 27

    6.1 Model Diagnostic for Dynamic Panel Models: 28

6.2 Exercise (With Marketing Topic): 29

Bibliology: 34

1. Introduction to R

“R is an Object-Oriented Programming Language used for “statistical computing” and representation of data in graphical mode. The R is widely applied for “data analysis” and it is Open source software with the largest community to resolve the issues and along with beginner tutorial blogs provides clear insight and ease of its software. Therefore, R language is a package oriented style with different kind of analysis requires respective packages need to be installed in the R software and a few are installed by default. Therefore, its Graphical User Interface is simple to interact and gain knowledge of its serviceability.

Consequently, R provides built-in functions for convenient data analysis, e.g. Package “library(car)”. Hence, new user required the knowledge of existing function in R software, e.g. r_functions.pdf for exceptional handling. Therefore, User also required statistical knowledge of the R functions to employ legitimate data analysis and to interpret the error message to resolve and differentiate among software failure and below image is a view of the user interface in R software.

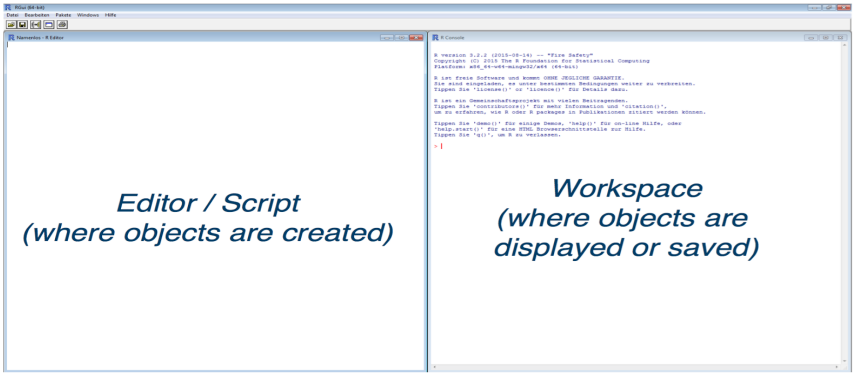


Fig_1: “R User Interface”

Let’s begin with the simple dataset consist of “beer” products of “30 similar supermarkets”. Therefore, generally dataset file format is in beer.xlsx, to avoid incorrect data reading by R software requires a format in CSV (comma separated value) for accurate results convert file from beer.xlsx to beer.csv. To do perform below instruction.Open the “beer.xlsx” file with the help of excel tool and then later export as a file “beer.csv”. Firstly, open the file “beer.csv” to look into the dataset for the product classification with the columns heading names.

Marketid	Sales.Becks	Price.Becks	Display.Becks	Price.Heineken	Region
1	29	4.71	0	4.66	1
2	28	4.42	1	4.58	1
3	29	4.66	1	4.38	1
4	28	4.66	1	4.77	1

Fig_2: beer.csv

From above Fig_2, columns describe the distinct fields with various names to differentiate the data in the “beer.csv” file.

1. Market.ID: It displays distinct Identifier for a supermarket with various integer value as ID.

2. Sales.Becks: It displays becks Unit sales (6-pack of Becks Beer) of the beer for a respective supermarket.
3. Price.Becks: It displays becks Price in € (6-pack of Becks Beer) for each supermarket with respective to its sales.
4. Display.Becks: It display becks beer brand i'e. 1 = Display for Becks, 0 = No display
5. Price.Heineken: It display Heineken beer Price in € (6-pack of Heineken Beer).
6. Region: It displays the region of the two beer products i'e. 1=North, 2=East, 3=South, 4=West.

In conclusion, panel data analysis with R software it requires respective package i'e. library(plm).

2. Refreshment in Regression Analysis:

One of most crucial types of data analysis is "Data Regression". Its approach has mathematically categorized those variables does certainly have a greater impact on the data which has to be analyzed E.x. Panel data analysis in Marketing. The data regression consists of two main variables I'e. Dependent variable mainly focusing on the factors that you're attempting to predict or understand and Independent variable mainly focusing on the factors whose impact is of dependent variables. Therefore, It allows estimating the relationship between variables.

2.1. Linear Regression(ordinary Least Squares):

In statistics, Linear Regression model is a method for modeling the relationship between a scalar dependent variable "Y" and independent variable "X". The case of one independent variable is called "simple linear regression" and in the case of more than one independent variable X then it called "Multiple linear regression". Therefore, Ordinary Least Squares (OLS) is an approach for predicting unidentified parameters in the Linear Regression. In the Linear Regression main focus is estimating b_0, b_1….b_n and evaluating whether they are significantly based on the P-value (< 2.5). Models are estimated using `lm` function to transform the data. However, regression analysis leads to set of techniques for predicting an outcome variable using one or more independent variables.

2.1.1 Simple Linear Regression:

This model describes the relationship between one variable X and Y variable can be formulated these ambiguities by following an equation. However, it considers one independent variable X. The factors b_0, b_1….b_n are called parameters where b_0 is the Y intercept and b_1 is slope of the regression line, which measures the changes in Y with respect to X_1 holding other factors fixed and the variable e, called the error term or disturbance in the relationship. It is also called as two-variable linear Regression.

$$Y = b_0 + b_1 X_1 + e$$

From above equation X is an explanatory variable and Y is a dependent variable and coefficient value of b_0, b_1…...b_n are found by minimizing the error of prediction. This model consists of only one X variable so the regression line is formed and represents an association between two variables on a scatterplot and using that line as linear model for predicting the value of one variable from other variable. Therefore, it is simple to find the error of prediction which will be the difference between the

real value and predicted value. The simple Linear Regression analysis considers all elements affecting Y other than X as being unobserved.

2.1.2 Multiple Linear Regression:

This model describes the relationship between multiple X variables and Y variables can be formulated these ambiguities by following an equation. The factors $b_0, b_1....b_n$ are called parameters where b_0 is the Y intercept and b_1 is slope of the regression line, which measures the changes in Y with respect to X_1 holding other factors fixed, b_2 is slope of the regression line, which measures the changes in Y with respect to X_2 holding other factors fixed and the variable e is called the error term and its depends on only one observed element.

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + e$$

From above equation X is an explanatory variable and Y is a dependent variable and coefficient value of $b_0, b_1.....b_n$ are found by minimizing the error of prediction. This model consists of multiple X variables so the regression plane is formed and it becomes complex to find the error of prediction which will be the difference between the real value and predicted value. It also handles for generalizing functional relationships between variables.

2.2 Logarithmic Transformation in OLS:

These model variables are generally used in the regression model to handle a situation where a non-linear relationship occurs between the independent variable X and dependent variable Y. Therefore, applying the logarithm of one or more variables instead of the un-logged form cause the non-linear relationship efficient.

A logarithmic transformation is also user-friendly of transforming an extremely skewed variable into one that is more approximately ordinary. Now consider a simple linear regression there is four possible combinations of transformation involving logarithm.

1. Linear-Linear $Y = b_0 + b_1 X_1 + e$
2. Linear-Log $Y = b_0 + b_1 log(X_1) + e$
3. Log-Linear $log(Y) = b_0 + b_1 X_1 + e$
4. Log-Log $log(Y) = b_0 + b_1 log(X_1) + e$

From above four possible logarithm choose the best fit and consider the theoretical reasoning for elasticity.

2.3 Elasticity: The percentage change in Y from a 1% change in X_1.

Computation:

Linear- Linear : The Equation as follows : $b_1(X_1 / Y)$

"A one unit change in X_1 is associated with a b_1 unit change in Y on average". The OLS coefficient b_1 is a slope. Note: A full interpretation must specify the context, give variable descriptions, and provide the units of measurement of both x and y.

Linear-Log: The Equation as follows: b_1 / Y

"A one percent change in X_1 is associated with approximately a $b_1/100$ unit change in Y on average".Note: A full interpretation must specify the context, give variable descriptions, and provide the units of measurement of both x and y.

Log-Linear: The Equation as follows: $b_1 X_1$

"A one unit change in X_1 is associated with approximately a 100 ∗ b_1 percent change in Y on average". Note: A full interpretation must specify the context, give variable descriptions, and provide the units of measurement of both x and y

Log-Log: The Equation as follows: b_1

"A one percent change in X_1 is associated with approximately a b_1 percent change in Y on average." The OLS coefficient b_1 is an elasticity. Note: A full interpretation must specify the context, give variable descriptions, and provide the units of measurement of both x and y.

An Individual level elasticities should use X_1 and predicted-Y and an average elasticities should use mean (X_1) and mean (predicted-Y).

2.4 Exercise (With Marketing Topic):

1. Estimate (and evaluate) linear regression models using the beer data (Sales.Becks = b_0 + b_1 *Price.Becks* + b_2 *Display.Becks* + b_3 *Price.Heineken* + *Residuals*)

Solution: From above question, It assumes Multiple Linear Regression as it is in the form of $Y = b_0 + b_1 X_1 + b_2 X_2 + ... + e$. Therefore, Dependent variable Y = Sales.Becks and remaining Independent Variable X_1 = Price.Becks, X_2 = Display.Becks and X_3 = Price.Heineken and residuals = the distance between estimated values and actual values. Now, Let us consider Linear-Linear.

Results:

```
Call:
lm(formula = Sales.Becks ~ Price.Becks + Display.Becks + Price.Heineken,
    data = beer)

Residuals:
    Min      1Q  Median      3Q     Max
-4.1502 -1.9136  0.2482  1.7409  5.0279

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     60.175     11.675   5.154 2.24e-05 ***
Price.Becks    -16.070      1.641  -9.795 3.26e-10 ***
Display.Becks    4.153      1.036   4.009 0.000457 ***
Price.Heineken   8.473      1.835   4.617 9.24e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.602 on 26 degrees of freedom
Multiple R-squared:  0.8865,    Adjusted R-squared:  0.8734
F-statistic: 67.66 on 3 and 26 DF,  p-value: 2.062e-12
```

From above results, here we can interpret residuals provides information about the wrong prediction about sales.becks based on the price.becks, display.becks and price.Heineken (under -4.1502 and over 5.0279) and intercept b_0 = 60.175 and b_1 = -16.070, b_2 = 4.153, b_3 = 8.473 and all our coefficients estimated are significant. As p-value is much less than 0.05, we can reject the null hypothesis that variables *Price.Becks, Display.Becks* and *Price.Heineken* collectively has no effect on Sales.Becks. Hence there is a significant relationship between the variables in the linear regression model of the data set beer. Therefore, Multiple R-squared gives approximately 89% of the variation in sales.Becks can be explained by our model (Price.becks, Display.Becks and Price.Heineken) and F-statistics (68%) and the p-value is for an overall test for significance of our model this test the null hypothesis that all the model coefficient are 0 basically its means slope for Price.becks = Display.Becks = Price.Heineken = 0. The Residual Standard error gives an idea of how far observed Sales.Becks are

from predicted Sales.Becks. Therefore, we associate an increase of 1 unit in *Price.Becks* with a decrease of -16.070 in *Sales.Becks* adjusting for *Display.Becks* and *Price.Heineken* and also hypothesis test that the slope of *Price.Becks* is zero. we associate an increase of 1 unit in *Display.Becks* with an increase of 4.153 in *Sales.Becks* adjusting for *Price.Becks* and *Price.Heineken* also hypothesis test that the slope of *Display.Becks* is zero. we associate an increase of 1 unit in *Price.Heineken* with an increase of 8.473 in *Sales.Becks* adjusting for *Price.Becks* and *Display.Becks* also hypothesis test that the slope of *Price.Heineken* is zero.

2. Compute (own-)price elasticity for Becks sales

Results: we consider Linear-Linear

```
   Min.   1st Qu.  Median   Mean    3rd Qu.   Max.
-4.259   -3.024   -2.596  -2.534   -2.152   -1.324
```

From the above results becks, sales have a mean value of -2.534 and the beer dataset price.becks variables vary from a beer with a price of -4.259 to -1.324, respectively. It is not symmetrically distributed. The price of price.becks should be in between above min and max values. From above we can see the prices of becks is decreased.

3. Compute display elasticity for Becks sales

Results: we consider Linear-Linear

```
   Min.    1st Qu.   Median    Mean    3rd Qu.    Max.
0.00000  0.00000  0.04514   0.06299   0.12490   0.15970
```

From the above results becks, display have a mean value of 0.06299 and the beer dataset display.becks variables vary from a beer with a price of 0.00000 to 0.15970, respectively. The sales of display.becks should be in between above min and max values.

4. Compute (own-)price elasticity of Heineken prices for Becks sales

Results: we consider Linear-Linear

```
   Min.    1st Qu.   Median    Mean     3rd Qu.    Max.
0.8878   1.1840   1.3760     1.3600    1.5450    1.9870
```

From the above results price of Heineken has a mean value of 1.3600 and The beer dataset price.heineken variables varies from a beer with a price of 0.8878 to 1.9870, respectively.

5. Discuss how you can use these results to decide about optimal price and display levels

Results: we consider Linear - Linear elasticity transformation as Price.Becks is in Negative sign i.e. with an increase of 1 unit in Price.Becks there will be a decrease in beck's sales and display.becks increase of 1 unit the sales of becks will be increased. Conclusion, Price.Becks price should be optimal in the range -4.259 to -1.324, and display.Becks should in the range 0.00000 to 0.15970 to increase sales of the becks.

3. Fixed Effect Model:

It has different labels in the literature namely “Least squares Dummy Variable Model”, “Within Estimator”, “Error Component Model” and most commonly “Fixed Effect Model” is implemented or considered only observed in analyzing the impact of variables that change over time. This model helps to focus on the heterogeneity across years basically it focus on the unobserved variables that do not change over time. To explain in the statistical way it considers observation from different individuals

(index *x*) over several time periods (index *t*). Its assumes that the individual specific effect will be correlated with the independent variables.

The Mathematical Equation : $Y_{it} = b_0 + b_1 X_{it} + e_{it}$

The disadvantage of OLS model it does not consider the unobserved heterogeneity of individuals *i* and time periods *t*. Therefore Fixed Effect model handle the unobserved heterogeneity of individuals and time period *t*. This model account for the time-invariant heterogeneity of individuals or time-periods (one-way FE model) or both individuals and time-periods (two-way FE model).

An Alternative way of estimating fixed effect models using Linear Squares Dummy Variable Model (LSDV) if you have less observation *i* and time period *t* and within effects estimation methods. It also includes a dummy variable for each observation and/or time in the panel data matrix. Therefore, these dummy variables as additional explanatory variables and considered as the part of the intercept and respective parameters estimates of dummy variables are the unobserved effects of individuals and time-periods. Ordinary least squares(OLS) regressions with dummies in fact are fixed effect method.

FE model for panel data equation for observation *i* and time period *t*.

The Mathematical Equation : $Y_{it} = b_0 + b_1 X_{it} + e_{it}$ --------- *(1)*

After using OLS for Eq. (1) leads to biased estimates if unobserved heterogeneity for *i* and *t* exists because below equation explains it.

The Mathematical Equation : $Y_{it} = b_0 + b_1 X_{it} + ind_i + time_t + u_{it}$ --------- *(2)*

The ambiguity with the LSDV model is cumbersome with many *i* and *t*. Therefore, as a solution for this cumbersome, we apply FE model by demeaning as One-way FE model and two-way FE model.

3.1 One-way fixed effect Model:

This model account for time-invariant heterogeneity of individuals or time-periods. The specification for the one-way fixed model is individual fixed effect.

The Mathematical Equation :$(Y_{it} - Y_{mi}) = b_0(1-1) + b_1 (X_{it} - X_{mi}) + (e_{it} - e_{mi}) + time_t$ --------- *(3)*

From above equation Y_{mi} and X_{mi} are the average value within observations of individual *i*. The individual fixed effects remove the effect ind_i and b_0.

The specification for the one-way fixed model is Time fixed effect.

The Mathematical Equation :$(Y_{it} - Y_{mt}) = b_0(1-1) + b_1 (X_{it} - X_{mt}) + (e_{it} - e_{mt}) + ind_i$ --------- *(4)*

From above equation Y_{mt} and X_{mt} are the average value within observations of Time-period *t*. The individual fixed effects remove the effect $time_t$ and b_0. Therefore, demeaned residuals *e* are just a

result of the estimation with demeaned Y and X and fixed effect model is equal to the OLS applied to demeaned data matrix.

3.2 Two-way fixed effect Model:

This model account for time-invariant heterogeneity of individuals and time-periods. The specification for the two-way fixed model is individual and time fixed effect.

The Mathematical Eq :$(Y_{it} - Y_{mi} - Y_{mt} + Y_m) = b_1 (X_{it} - X_{mi} - X_{mt} + X_m) + (e_{it} - e_{mi} - e_{mt} + e_m)$------(5)

From above equation Y_{mi} and X_{mi} are the average value within observations of individual *i,* Y_{mt} and X_{mt} are the average value within observations of Time-period *t* and Y_m and X_m are the average overall value. The individual and time fixed effects remove the effect of ind_i and $time_t$ and b_0. Therefore, demeaned residuals *e* are just a result of the estimation with demeaned Y and X and fixed effect model is equal to the OLS applied to demeaned data matrix.

The specification for the two-way fixed model is to compute individual and time fixed effect.

The Mathematical Eq : $ind_i = (Y_{mi} - Y_m) - b_1 (X_{mi} - X_m)$------(6)

$time_t = (Y_{mt} - Y_m) - b_1 (X_{mt} - X_m)$------(7)

From above equation, If ind_i and $time_t$ have no significant effect, then Fixed effect model is equal to the Ordinary Least Squares (OLS).

3.3 Model diagnostics for Fixed Effect model:

It requires corrections necessary for F-value and R^2 and consider the specification tests.

s.no	Specification test name	Specification test Method	Specification test Method
1	Cross-section dependence	(Pesaran test) H_0: No Cross-section dependence (H_0 good!)	If yes (bad!): Estimate Fixed Effect-SUR model
2	Stationary	(Im, Pesaran and Shin Test) H_0: Non Stationary/ Trend (H_0 bad!)	If no (bad!): Use first difference or linear time trend as X
3	Auto-correction	(Breusch-Godfrey Test) H_0: No Auto-correction (H_0 good!)	If yes (bad!): Estimate HAC covariance matrix
4	Heteroskedasticity	(Breusch-Pagan Test) H_0: No Heteroskedasticity (H_0 good!)	If yes (bad!): Estimate HAC covariance matrix
5	Multicollinearity	(Variance inflation Factors) VIFs (< 5 good!)	If yes (bad!): Exclude correlated X or use

			Principal Comp.

3.4 HAC Covariance Matrix in FE model:

HAC stands for Heteroskedasticity and Autocorrelation Consistent. This model considers Unbiased parameter estimates but biased standard errors affect significance testing. Although, standard error are called cluster-robust standard error and it follows the cluster observations of individual *i* for covariance matrix estimation. Therefore, It also captures within-cluster correlation of the residuals.

3.5 Exercise (With Marketing Topic):

Firstly, open the file "slice.csv" to look into the dataset for the product classification with the columns heading names.

store	week	sales	price	display
ALBANY,NY - PRICE CHOPPER	1	778	3,052699	0
ALBANY,NY - PRICE CHOPPER	2	695	3,053237	0
ALBANY,NY - PRICE CHOPPER	3	661	3,049924	0
ALBANY,NY - PRICE CHOPPER	4	789	3,051965	0

From the above figure "slice.csv" columns describe the distinct fields with various names to differentiate the data in the "slice.csv" file.

1. Estimate an OLS Model with (Sales = $b_0 + b_1\ Price + b_2\ Display + Residuals$).

Solution:

```
Call:
lm(formula = sales ~ price + display, data = slice)

Residuals:
   Min     1Q Median     3Q    Max
 -9647  -2816   -987   1038 138655

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  10885.1      633.9  17.173   <2e-16 ***
price        -2216.7      216.6 -10.233   <2e-16 ***
display       7201.7      730.1   9.864   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6468 on 3596 degrees of freedom
Multiple R-squared:  0.06067,   Adjusted R-squared:  0.06015
F-statistic: 116.1 on 2 and 3596 DF,  p-value: < 2.2e-16
```

From above results, here we can interpret residuals provides information about the wrong prediction about sales based on the price and display (under -9647 and over 138655) and intercept b_0 = 10885.1 and b_1 = -2216.7, b_2 = 7201.7 and all our coefficients estimated are significant. As p-value is much less than 0.05, we can reject the null hypothesis that variables *Price, Display* collectively have no effect on Sales. Hence there is a significant relationship between the variables in the linear regression

model of the data set slice. Therefore, Multiple R-squared gives approximately 0.060 of variation in sales can be explained by our model (Price, Display) and F-statistics and p-value are for overall test for significance of our model this test the null hypothesis that all the model coefficient are zero(0) basically its means slope for Price = Display = 0. The Residual Standard error gives an idea of how far observed Sales are from predicted Sales. If we see price coefficient value is a Negative sign (dropped) and display value in the Positive sign. In conclusion, price value is in Negative value so if the prices increases sales will decrease and vice versa and display value is in positive value so it increases then the sales value will be increased.

2. Estimate a LSDV Model with (Sales = $b_0 + b_1\ Price + b_2\ Display + ind_i + time_t + Residuals$).

Results: **One-way FE for "store"**

```
Call:
lm(formula = sales ~ price + display + store - 1, data = slice)

Residuals:
   Min     1Q Median     3Q    Max
-16309  -1156    -72    967 110586

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
price                               -8072.8      231.9  -34.82   <2e-16 ***
display                              7099.5      642.7   11.05   <2e-16 ***
storeALBANY,NY - PRICE CHOPPER      22990.7      854.4   26.91   <2e-16 ***
storeATLANTA - KROGER CO            27651.9      859.4   32.17   <2e-16 ***
.                                   .            .       .       .
.                                   .            .       .       .
storeTAMPA/ST. PETE - PUBLIX        27482.5      854.5   32.16   <2e-16 ***
storeTAMPA/ST. PETE - WINN DIXIE    26335.9      824.1   31.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4290 on 3538 degrees of freedom
Multiple R-squared:  0.752,     Adjusted R-squared:  0.7477
F-statistic: 175.9 on 61 and 3538 DF,  p-value: < 2.2e-16
```

From above results, few results are exempted. We are considering “Store” as Individuals as fixed effect model consider heterogeneity across group or time. Therefore, it considers stores as a group gives the estimated Intercept value for each respectively and P-value is less than 0.05 so our model is significant. Therefore, above figure convey the price value is in a negative value, then price increases then there will decrease in the sales. Hence there is a significant relationship between the variables in the linear regression model of the data set beer. Therefore, Multiple R-squared gives approximately 75% of the variation in sales.Becks can be explained by our model (Price, Display) and F-statistics (175.9) and the p-value is for an overall test for significance of our model this test the null hypothesis that all the model coefficient are 0.

One-way FE for "week"

```
Call:
lm(formula = sales ~ price + display + factor(week) - 1, data = slice)

Residuals:
   Min     1Q Median     3Q    Max
 -9516  -2867  -1008   1135 137019

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
price           -2199.4      219.6 -10.015   <2e-16 ***
display          7021.8      749.8   9.365   <2e-16 ***
factor(week)1   11205.2     1050.0  10.672   <2e-16 ***

.                .           .      .       .
.                .           .      .       .
factor(week)60  12070.8     1038.4  11.624   <2e-16 ***
factor(week)61  10234.8     1048.3   9.763   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6491 on 3536 degrees of freedom
Multiple R-squared:  0.4326,    Adjusted R-squared:  0.4225
F-statistic:  42.8 on 63 and 3536 DF,  p-value: < 2.2e-16
```

From above results, few results are exempted. We are considering "week" as time-period as fixed effect model consider heterogeneity across group or time. Therefore, it considers week as a group gives the estimated Intercept value for each respectively and P-value is less than 0.05 so our model is significant. Therefore, above figure convey the price value is in a negative value, then price increases then there will decrease in the sales. Hence there is a significant relationship between the variables in the linear regression model of the data set beer. Therefore, Multiple R-squared gives approximately 43% of the variation in sales.Becks can be explained by our model (Price, Display) and F-statistics (42%) and p-value is for overall test for significance of our model this test the null hypothesis that all the model coefficient are 0

In conclusion one FE model with Store and Week, it differs a huge change in price and significant change in display value. If you consider price value for store and week is `-8072.8` and `-2199.4` respectively and display value for store and week is `7099.5` and `7021.8` respectively.

Two-way FE for "stores" and "week".

```
Call:
lm(formula = sales ~ price + display + store + factor(week) -
    1, data = slice)

Residuals:
   Min     1Q Median     3Q    Max
-15558  -1232    -53    996 107830

Coefficients:
                                    Estimate Std. Error t value Pr(>|t|)
price                               -8371.57     236.23 -35.438  < 2e-16 ***
display                              7274.51     663.02  10.972  < 2e-16 ***
storeALBANY,NY - PRICE CHOPPER      24104.09    1014.65  23.756  < 2e-16 ***
storeATLANTA - KROGER CO            28795.54    1020.53  28.216  < 2e-16 ***

factor(week)60                       -344.82     788.61  -0.437  0.66196
factor(week)61                      -1542.23     786.75  -1.960  0.05005 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4260 on 3478 degrees of freedom
Multiple R-squared:  0.7596,    Adjusted R-squared:  0.7513
F-statistic: 90.84 on 121 and 3478 DF,  p-value: < 2.2e-16
```

From above results, few results are exempted. We are considering now "store" and "week" as Individuals and time-period as fixed effect model consider heterogeneity across group and time. Therefore, it considers store and week as a group gives the estimated Intercept value for each respectively and P-value is less than 0.05 so our model is significant. Therefore, above figure convey the price value is in negative value i.e. `-8371.57` and display value i.e. `7274.51`, then price increases then will be a decrease in the sales. Therefore, Multiple R-squared gives approximately 76% of the variation in sales.Becks can be explained by our model (Price, Display) and F-statistics (99%) and the p-value is for an overall test for significance of our model this test the null hypothesis that all the model coefficient are 0.

In conclusion two-way FE model with Store and Week, the store intercept value are in a positive sign and week intercept values are in the combination of negative and positive sign with larger p-value and smaller p-value.

3. Estimate a FE Model with (Sales = $b_0 + b_1\ Price + b_2\ Display + ind_i + time_t + Residuals$).

Results: One-way (Individual) effect within Model **Compare price coefficients from LSDV1 and FE1 model**

```
Oneway (individual) effect Within Model

Call:
plm(formula = sales ~ price + display, data = slice, effect = "individual",
    model = "within")

Balanced Panel: n=59, T=61, N=3599

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-16300.0  -1160.0    -71.5    967.0 111000.0

Coefficients :
        Estimate Std. Error t-value  Pr(>|t|)
price   -8072.78     231.86 -34.817 < 2.2e-16 ***
display  7099.49     642.67  11.047 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    9.4003e+10
Residual Sum of Squares: 6.5119e+10
R-Squared:      0.30727
Adj. R-Squared: 0.30206
F-statistic: 784.657 on 2 and 3538 DF, p-value: < 2.22e-16
```

```
> coef(fe1)[1]
    price
-8072.785
> coef(lsdv1)[1]
    price
-8072.785
```

From the above result, we now consider `plm()` function, now we have extra information as like balanced panel, the total sum of squares and the residual sum of squares. Therefore, above figure convey the price value is in negative value i.e. `-8072.78` and display value i.e. `7099.49`, then price increases then there will be a decrease in the sales. Therefore, Multiple R-squared gives approximately 31% of the variation in sales.Becks can be explained by our model (Price, Display) and F-statistics `(784.657)` and the p-value is for an overall test for significance of our model this test the null hypothesis that all the model coefficient are 0.

In conclusion, coefficient values of the price with `lm()` and `plm()` remains same i.e. `-8072.78` but `lm()` function provides R-squared value is over all and `plm()` function provides R-squared value is partial value.

One-way (Time) FE for "week" **Compare price coefficients from LSDV2 and FE2 model**

```
Call:
plm(formula = sales ~ price + display, data = slice, effect = "time",
    model = "within")

Balanced Panel: n=59, T=61, N=3599

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
  -9520   -2870   -1010    1140  137000

Coefficients :
        Estimate Std. Error  t-value  Pr(>|t|)
price   -2199.37     219.60 -10.0154 < 2.2e-16 ***
display  7021.77     749.78   9.3651 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    1.5793e+11
Residual Sum of Squares: 1.4897e+11
R-Squared:      0.056736
Adj. R-Squared: 0.055743
F-statistic: 106.344 on 2 and 3536 DF, p-value: < 2.22e-16
```

```
> coef(fe2)[1]
   price
-2199.37
> coef(lsdv2)[1]
   price
-2199.37
```

From the above result we now consider `plm()` function, now we have extra information as like balanced panel, the total sum of squares and the residual sum of squares. Therefore, above figure convey the price value is in a negative value i.e. `-2199.37` and display value i.e. `7021.77`, then price increases then there will be a decrease in the sales. Therefore, Multiple R-squared gives approximately `0.0567` of variation, sales.Becks can be explained by our model (Price, Display) and F-statistics `(106.344)` and the p-value is for an overall test for significance of our model this test the null hypothesis that all the model coefficient are 0.

In conclusion, coefficient values of the price with lm() and plm() remains same i.e. `-2199.37`, but `lm()` function provides R-squared value is over all and `plm()` function provides R-squared value is partial value.

two-way FE for "week"and "stores" **Compare price coefficients from LSDV2 and FE2 model**

```
Twoways effects Within Model

Call:
plm(formula = sales ~ price + display, data = slice, effect = "twoways",
    model = "within")

Balanced Panel: n=59, T=61, N=3599

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-15600.0  -1230.0    -52.7    996.0 108000.0

Coefficients :
        Estimate Std. Error t-value  Pr(>|t|)
price   -8371.57     236.23 -35.438 < 2.2e-16 ***
display  7274.51     663.02  10.972 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    9.1757e+10
Residual Sum of Squares: 6.3108e+10
R-Squared:      0.31222
Adj. R-Squared: 0.30173
F-statistic: 789.438 on 2 and 3478 DF, p-value: < 2.22e-16
```

```
> coef(fe3)[1]
    price
-8371.571
> coef(lsdv3)[1]
    price
-8371.571
```

From the above result, we now consider `plm()` function, now we have extra information as like balanced panel, the total sum of squares and the residual sum of squares. Therefore, above figure convey the price value is in a negative value i.e. `-8371.571` and display value i.e. `7274.51`, then price increases then there will be a decrease in the sales. Therefore, Multiple R-squared gives approximately `31%` of the variation in sales.Becks can be explained by our model (Price, Display) and F-statistics `(789.438)` and the p-value is for an overall test for significance of our model this test the null hypothesis that all the model coefficient are 0.

In conclusion, coefficient values of the price with lm() and plm() remains same i.e. `-8371.571`, but `lm()` function provides R-squared value is over all and `plm()` function provides R-squared value is partial value.

4. Compare logarithmic transformations for 2-way FE model.

Results: **Compare models by R-square**

```
> r.squared(tw1)
[1] 0.3122236
> r.squared(tw2)
[1] 0.3507495
> r.squared(tw3)
[1] 0.5515766
> r.squared(tw4)
[1] 0.5541176
```

From the above results, we have log transformation i.e. linear- linear, linear-log,log-linear and log-log as tw1, tw2, tw3, and tw4 respectively with different r-square values which are partial values and we can say our model is a good fit because all the values are above 30%.

5. Compute price and display elasticity for best FE model.

Results: **Price elasticity**

```
> #Best
> coef(tw4)[1]
log(price)
 -2.331436
> #Second-best
> summary(coef(tw3)[1]*slice$price)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-3.746  -2.501  -2.152  -2.277  -1.976  -1.065
```

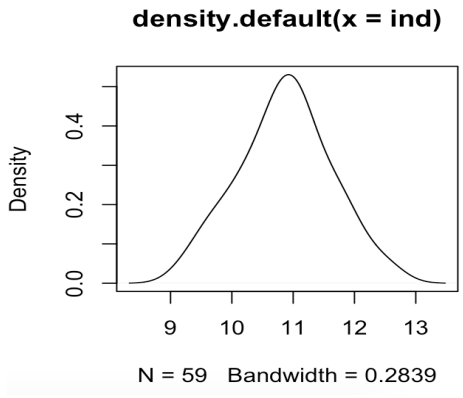
Display elasticity

```
> #Best
> coef(tw4)[2]
log(display + 0.01)
         0.07724312
> #Second-best
> summary(coef(tw3)[2]*slice$display)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.00000 0.03084 0.07898 0.10360 0.77620
```

From the above results, we have log transformation i.e. log-linear and log-log as tw3 and tw4 respectively with different best price values ranges in between `-3.746` and `-1.065` value and display values range from 0.000 to 0.77620.

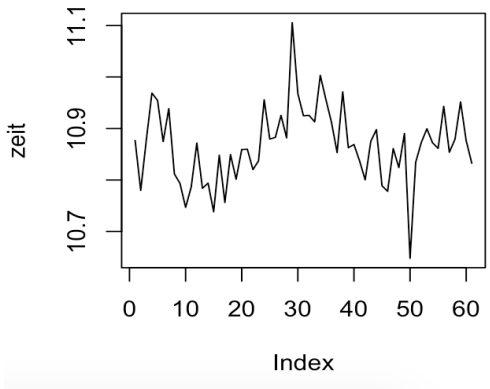
6. Plot ind_i and $time_t$ for best FE model.

Results: **Individual fixed effects and Plot hist and density of individual (store) fixed effects**



We consider **log-log** transformation for individuals for the store where percentage increase in price and display leads to the constant percentage change in sales.

Time fixed effects and Plot line (type="l") of time (week) fixed effects



We consider **log-log** transformation for individuals for a week where percentage increase in price and display leads to the constant percentage change in sales.

7. Perform specification tests for best FE model

Results:

```
        Pesaran CD test for cross-sectional dependence in panels

data:  formula
z = -2.419, p-value = 0.01557
alternative hypothesis: cross-sectional dependence
```

From above results, the p-value is less that a significance level of 0.05 and the alternative hypothesis is cross-sectional dependence then it conveys (bad!) sign.

```
        Pesaran's CIPS test for unit roots

data:  pdata.frame(slice)$sales
CIPS test = -3.5424, lag order = 2, p-value = 0.01
alternative hypothesis: Stationarity
```

From above results, the p-value is less that a significance level of 0.05 and the alternative hypothesis is Stationarity then it has no trend, which is (good!) sign.

```
        Breusch-Godfrey/Wooldridge test for serial correlation in panel models

data:  log(sales) ~ log(price) + log(display + 0.01)
chisq = 394.65, df = 61, p-value < 2.2e-16
alternative hypothesis: serial correlation in idiosyncratic errors
```

From above results, the p-value is less that a significance level of 0.05 and the alternative hypothesis is a serial correlation in idiosyncratic error. It shows Serial autocorrelation (bad!).

```
        studentized Breusch-Pagan test

data:  lsdv3
BP = 453.22, df = 120, p-value < 2.2e-16
```

From above results, the p-value is less that a significance level of 0.05. Therefore, we can reject the null hypothesis that the variance of the residuals is a constant and infer that heteroscedasticity is indeed present, which sign is (bad!).

```
                          GVIF Df GVIF^(1/(2*Df))
log(price)            1.04571  0            Inf
log(display + 0.01) 1.04571  0            Inf
```

From above results, the variance Inflation factors is less than 5 then there is no multicollinearity.

8. Estimate the best FE model with a HAC covariance matrix

Results:

```
t test of coefficients:

                     Estimate Std. Error  t value  Pr(>|t|)
log(price)         -2.3314357  0.1347881 -17.2970 < 2.2e-16 ***
log(display + 0.01)  0.0772431  0.0097392   7.9312 2.901e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

From above results log transform, the p-value for both independent variables is less that a significance level of 0.05, and the price coefficient is having negative value with increase in price then there will be a decrease in sales. The value of price/display has increased and p-value for the price is same and the display has changed if we compare with a coefficient test function.

4. Two Stage Least Squares in Fixed Effects Models:

It is widely used to estimate the parameters in systems of linear simultaneous equations and to solve the problems of omitted variables in single bias equation estimation. Basically it solves the endogeneity problem when a part of the variance in X can be endogenous which leads to correlation between X and e, the reason behind the endogeneous problem in marketing arises due to the partial reverse causation and omitted (time-varying) variables. Therefore, Consequences generates biased parameters estimates, to solve this we use 2SLS with use of instrumental variable (IV) approach and also Instrument-free approaches.

In conclusion, it applies to OLS and FE to account for the endogenous variations in X. If we chose variables Z (= instruments) that explain the endogenous variation in X where Z is not correlated with *e* and has no direct effect on Y. In the first stage we use endogenous-X as dependent variables and [z + endogenous-X] as independent variables and in second stage we use predicted-X from first stage as

explanatory variable instead of endogenous-X in main regression model. Consequently, we use fixed effects in both stages.

4.1 Exercise (With Marketing Topic):

Firstly, open the file "tuna.csv" to look into the dataset for the product classification with the columns heading names.

brand	week	sales	display	rtprice	wsprice
1	1	20347	0	0.913804551514775	0.600224162723273
1	2	44351	0	0.75328277233936	0.60022384280388
1	3	25177	0	0.753156873905878	0.600225151293283
1	4	11032	0	0.918930104093861	0.600226970578474

From above figure "tuna.csv" columns describe the distinct fields with various names to differentiate the data in the "tuna.csv" file.

1. Estimate best 2-way FE model with Sales = $b_0 + b_1\ display + b_2\ rtprice + ind_i + time_t + residuals$.
Results:

```
Call:
plm(formula = log(sales) ~ log(display + 0.01) + log(rtprice),
    data = tuna, effect = "twoways", model = "within")

Balanced Panel: n=7, T=338, N=2366

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
 -4.420  -0.237   0.015   0.233   3.090

Coefficients :
                     Estimate Std. Error  t-value  Pr(>|t|)
log(display + 0.01)  0.0687816  0.0073001   9.4221 < 2.2e-16 ***
log(rtprice)        -3.8785638  0.1423821 -27.2405 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1060.5
Residual Sum of Squares: 616.77
R-Squared:      0.41842
Adj. R-Squared: 0.35723
F-statistic: 726.637 on 2 and 2020 DF, p-value: < 2.22e-16
> coef(feno)[2]
log(rtprice)
   -3.878564
```

From above results log transform, the p-value for the independent variable and overall is less that a significance level of 0.05 and the price coefficient is having negative value with increase in price then there will be a decrease in sales and R-squared value is partial value.

2. Estimate best 2SLS-2-way FE model with Sales = $b_0 + b_1\ display + b_2\ rtprice + ind_i + time_t + residuals$, $rtprice_p = b_0 + b_1\ display + b_2\ wsprice + ind_i + time_t + residuals$.
Results:

```
Twoways effects Within Model
Instrumental variable estimation
   (Balestra-Varadharajan-Krishnakumar's transformation)

Call:
plm(formula = log(sales) ~ log(display + 0.01) + log(rtprice) |
    log(display + 0.01) + log(wsprice), data = tuna, effect = "twoways",
    model = "within")

Balanced Panel: n=7, T=338, N=2366

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-4.4200 -0.2360  0.0144  0.2320  3.0900

Coefficients :
                   Estimate Std. Error t-value  Pr(>|t|)
log(display + 0.01)  0.070139   0.012886  5.4432 5.868e-08 ***
log(rtprice)        -3.825729   0.437005 -8.7544 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1060.5
Residual Sum of Squares: 616.81
R-Squared:      0.4184
Adj. R-Squared: 0.35721
F-statistic: 726.519 on 2 and 2020 DF, p-value: < 2.22e-16
> ###Price elasticity
> coef(feiv)[2]
log(rtprice)
   -3.825729
```

```
Twoways effects Within Model

Call:
plm(formula = log(sales) ~ log(display + 0.01) + log(rtprice) +
    log(wsprice), data = tuna, effect = "twoways", model = "within")

Balanced Panel: n=7, T=338, N=2366

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
 -4.420  -0.237   0.015   0.233   3.090

Coefficients :
                   Estimate Std. Error  t-value Pr(>|t|)
log(display + 0.01)  0.0686729  0.0073511   9.3418   <2e-16 ***
log(rtprice)        -3.8848391  0.1506370 -25.7894   <2e-16 ***
log(wsprice)         0.0076362  0.0597260   0.1279   0.8983
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1060.5
Residual Sum of Squares: 616.76
R-Squared:      0.41842
Adj. R-Squared: 0.35705
F-statistic: 484.194 on 3 and 2019 DF, p-value: < 2.22e-16
> vif(feer)
                    GVIF Df GVIF^(1/(2*Df))
log(display + 0.01) 1.514448  0            Inf
log(rtprice)        1.514448  0            Inf
log(wsprice)        1.514448  0            Inf
```

From above results log transform, the p-value for the independent variable and overall is less that a significance level of 0.05 for both results and the price coefficient is having negative value in both results, with an increase in price then there will be a decrease in sales and R-squared value is partial value. Therefore, the variance Inflation factors are less than 5 then there is no multicollinearity in second results where it considers three independent variables.

3. Compute difference in parameter estimates.

a) with Vs. without IV

Results:

```
> coef(feno)[1]-coef(feiv)[1]
log(display + 0.01)
       -0.001357856
> #Display
> coef(feno)[1]-coef(feiv)[1]
log(display + 0.01)
       -0.001357856
> #Retail price
> coef(feno)[2]-coef(feiv)[2]
log(rtprice)
 -0.05283515
> phtest(fe1,fe2)

        Hausman Test

data:  sales ~ price + display
chisq = 5213, df = 2, p-value < 2.2e-16
alternative hypothesis: one model is inconsistent

> cor(feno$residuals,tuna$rtprice)
[1] 0.003027503
> cor(feiv$residuals,tuna$rtprice)
[1] 0.002505018
```

From the above results log transform, the p-value overall is less that a significance level of 0.05 and the price/display coefficient is having negative value, with an increase in price/display then there will be a decrease in sales and R-squared value is partial value. Therefore, the alternative hypothesis is inconsistent with one model, its best for fixed effect model and both correlation values are positive which indicates the extent to price and display variables increase or decrease in parallel.

b) with Vs. without omitted variable display.

Results:

```
Twoways effects Within Model
Instrumental variable estimation
   (Balestra-Varadharajan-Krishnakumar's transformation)

Call:
plm(formula = log(sales) ~ log(rtprice) | log(wsprice), data = tuna,
    effect = "twoways", model = "within")

Balanced Panel: n=7, T=338, N=2366

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-4.51000 -0.23600  0.00873  0.23700  3.18000

Coefficients :
             Estimate Std. Error t-value  Pr(>|t|)
log(rtprice) -4.09495    0.40441 -10.126 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1060.5
Residual Sum of Squares: 648.05
R-Squared:      0.39286
Adj. R-Squared: 0.33557
F-statistic: 1286.21 on 1 and 2021 DF, p-value: < 2.22e-16
> #Retail Price
> coef(feiv)[2]-coef(feod)[1]
log(rtprice)
   0.2692212
> cor(feiv$residuals,tuna$rtprice)
[1] 0.002505018
> cor(feod$residuals,tuna$rtprice)
[1] -0.006318297
```

From above results log transform, the p-value overall is less that a significance level of 0.05 and the price coefficient is having negative value, with an increase in price then there will be a decrease in sales and R-squared value is partial value. Therefore, one correlation values are positive which indicates the extent to price and display variables increase or decrease in parallel. Another correlation value is negative indicates the extent to one price/display variable increases as the other decreases.

5. Random Effects Models:

This model examines how individuals and/or time effect error variances and to avoid loss of degrees of freedom then ind_i and $time_t$ can be assumed as random, it is appropriate for N individuals drawn randomly from large population. To explain in the statistical way it considers observations randomly distributed from individuals (index i) over several time periods (index t). This model is adapted when N is large and FE model would point to an excessive loss of degrees of freedom. Its assumes that the individual specific effect will be uncorrelated with the independent variables.

The Mathematical Equation : $$Y_{it} = b_0 + b_1 X_{it} + ind_i + time_t + u_{it} \ldots\ldots\ldots\ldots (1)$$

From above equation individuals i = 1,2…..N and time periods t = 1,2…...T. It assumes ind_i and $time_t$ are random disturbances drawn from specified distribution there are represented below.

$$ind_i \sim IID(0, s^2_i)$$
$$time_t \sim IID(0, s^2_t)$$
$$u_{it} \sim IID(0, s^2_{it})$$

From above equations ind_i and $time_t$ are independent of u_{it}. $IID(m, s^2)$ represents Independent and Identically Distributed random variables with mean = m and variance s^2. This model account for the time-invariant heterogeneity of individuals or time-periods (one-way FE model) or both individuals and time-periods (two-way FE model). It partially demeaning the data and it allows explanatory variables that are invariant across individuals or time periods will be an advantage compared over FE model. Therefore, It has symmetric distribution for ind_i and $time_t$ and it provide no correlation between X and ind_i / $time_t$ will be an disadvantage of this model.

5.1 One-way Random Effect Model:

This model account for time-invariant heterogeneity of individuals or time-periods. The specification for the one-way random model is individual random effect.

The Mathematical Equation :$(Y_{it} - \boldsymbol{r}_i Y_{mi}) = b_0(1- \boldsymbol{r}_i 1) + b_1 (X_{it} - \boldsymbol{r}_i X_{mi}) + (e_{it} - \boldsymbol{r}_i e_{mi}) + time_t$ --------- (2)

From above equation Y_{mi} and X_{mi} are the average value within observations of individual i. The individual random effects remove the effect of ind_i. if $ind_i \sim IID(0, s^2_i)$ with $r_i = 1-[s_{it} / (Ts^2_i + s^2_{it})^{0.5}]$ where if $r_i = 0$ then OLS, if $r_i = 1$ then 1-way FE model.

The specification for the one-way random model is Time random effect.

The Mathematical Equation :$(Y_{it} - \boldsymbol{r}_t Y_{mt}) = b_0(1-\boldsymbol{r}_t 1) + b_1 (X_{it} - \boldsymbol{r}_t X_{mt}) + (e_{it} - \boldsymbol{r}_t e_{mt}) + ind_i$ --------- (3)

From above equation Y_{mt} and X_{mt} are the average value within observations of time-period t. The time random effects remove the effect $time_t$. if $time_t \sim IID(0, s^2_t)$ with $r_i = 1-[s_{it} / (Ns^2_t + s^2_{it})^{0.5}]$ where if $r_t = 0$ then OLS, if $r_t = 1$ then 1-way FE model.

5.2 Two-way Random Effect Model:

This model account for time-invariant heterogeneity of individuals and time-periods. The specification for the two-way random model is individual and time random effect.

The Mathematical Eq :$(Y_{it} - \boldsymbol{r}_i Y_{mi} - \boldsymbol{r}_t Y_{mt} + \boldsymbol{r}_{it} Y_m) = b_1 (X_{it} - \boldsymbol{r}_i X_{mi} - \boldsymbol{r}_t X_{mt} + \boldsymbol{r}_{it} X_m) + (e_{it} - \boldsymbol{r}_i e_{mi} - \boldsymbol{r}_t e_{mt} + \boldsymbol{r}_{it} e_m)$------(4)

From above equation Y_{mi} and X_{mi} are the average value within observations of individual i, Y_{mt} and X_{mt} are the average value within observations of Time-period t and Y_m and X_m are the average overall value. The individual and time random effects remove the effect ind_i and $time_t$ if $ind_i \sim IID(0, s^2_i)$ and $time_t \sim IID(0, s^2_t)$ with $r_{it} = r_i + r_t + [s_{it} / (Ts^2_i + Ns^2_t + s^2_{it})^{0.5}]$ where if $r_i / r_t / r_{it} = 0$ then OLS, if $r_i / r_t / r_{it} = 1$ then 1-way FE model.

The specification for the two-way random model is how to estimate $s^2_n / s^2_t / s^2_{it}$. Therefore, for this approach it requires different methods or procedures for estimating $s^2_n / s^2_t / s^2_{it}$. Although, in R software *plm* function has two methods they are "Swamy and Arora (1972), Amemiya (1971), wallace and hussain (1969), and Nerlove (1971)" among all methods Swamy and Arora (1971) is default in R software.

5.3 Model diagnostics for Random Effect model:

Before choosing FE or RE model is quite difficult so the Hausman specification test (H) will provide an advantage to examines if the individual effects are uncorrelated with the other regression models. Since computation is complicated. Therefore, few assumptions will give an clarification how we can understand which model is good among FE versus RE model. If H_0 : FE = RE then (H_0 good!) but if FE = RE the same specification test as FE model is applicable for RE model i'e. HAC covariance matrix and 2SLS.

5.4 Exercise (With Marketing Topic):

Firstly, open the file "orange.csv" to look into the dataset for the product classification with the columns heading names.

store	week	sales	price	display	educ	income
1	1	6463,99999999398	4,984375	0	0,28015016424214	10,745377961834
1	2	4032,00000019675	4,984375	0	0,28015016424214	10,745377961834
1	3	7040,0000003373	4,984375	0	0,28015016424214	10,745377961834
1	4	5503,99999992746	4,984375	0	0,28015016424214	10,745377961834

From above figure "orange.csv" columns describe the distinct fields with various names to differentiate the data in the "orange.csv" file.

1. Compare OLS, 2-way-FE and 2-way-RE model using with log(Sales) = b_0 + b_1 *log(price)* + b_2 *log(display)* + b_3 *educ*+ b_4 *income*+ *residuals*.

Results:

```
> #Estimate OLS model (use summary() and save as object "ols" simultaneously)
> summary(ols1<-lm(e1,data=juice))

Call:
lm(formula = e1, data = juice)

Residuals:
     Min       1Q   Median       3Q      Max
-2.02368 -0.42933 -0.01285  0.45284  1.97533

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       13.10347    0.12214  107.28   <2e-16 ***
log(price)        -2.30608    0.08846  -26.07   <2e-16 ***
log(display + 0.01)  0.13516    0.01080   12.52   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6809 on 1663 degrees of freedom
Multiple R-squared:  0.4466,    Adjusted R-squared:  0.446
F-statistic: 671.1 on 2 and 1663 DF,  p-value: < 2.2e-16
```

```
> summary(ols2<-lm(e2,data=juice))

Call:
lm(formula = e2, data = juice)

Residuals:
     Min       1Q   Median       3Q      Max
-1.91133 -0.39215 -0.00025  0.41320  1.74819

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       26.947627   0.977627   27.56   <2e-16 ***
log(price)        -2.424090   0.078480  -30.89   <2e-16 ***
log(display + 0.01)  0.129002   0.009558   13.50   <2e-16 ***
educ               5.159052   0.253793   20.33   <2e-16 ***
income            -1.403071   0.096127  -14.60   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6026 on 1661 degrees of freedom
Multiple R-squared:  0.5671,    Adjusted R-squared:  0.566
F-statistic: 543.9 on 4 and 1661 DF,  p-value: < 2.2e-16
```

```
> #Estimate FE model (use summary() and save as object "fe" simultaneously)
> summary(fe1<-plm(e1,data=juice,effect="two",model="within"))
Twoways effects Within Model

Call:
plm(formula = e1, data = juice, effect = "two", model = "within")

Balanced Panel: n=14, T=119, N=1666

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-1.43000 -0.18300  0.00287  0.18400  1.30000

Coefficients :
                   Estimate Std. Error t-value  Pr(>|t|)
log(price)         -2.7183582  0.0682936 -39.804 < 2.2e-16 ***
log(display + 0.01)  0.1179574  0.0076348  15.450 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    527.24
Residual Sum of Squares: 177.42
R-Squared:      0.6635
Adj. R-Squared: 0.61013
F-statistic: 1510.36 on 2 and 1532 DF, p-value: < 2.22e-16
> summary(fe2<-plm(e2,data=juice,effect="two",model="within"))
Error in crossprod(t(X), beta) : non-conformable arguments
```

The above error is due to collinearity among variables. PLM does not automatically place NAs in the beta vector for variables that were not estimated due to collinearity.

```
> #Estimate RE model (use summary() and save as object "re" simultaneousl
> summary(re1<- plm(e1,data=juice,effect="two",model="random"))
Twoways effects Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = e1, data = juice, effect = "two", model = "random")

Balanced Panel: n=14, T=119, N=1666

Effects:
                var std.dev share
idiosyncratic 0.11581 0.34031 0.298
individual    0.21015 0.45842 0.541
time          0.06216 0.24931 0.160
theta  : 0.9321 (id) 0.6573 (time) 0.6561 (total)

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-1.49000 -0.20100 -0.00688  0.21200  1.27000

Coefficients :
                   Estimate Std. Error t-value  Pr(>|t|)
(Intercept)        13.5974162  0.1520041  89.454 < 2.2e-16 ***
log(price)         -2.6775468  0.0636000 -42.100 < 2.2e-16 ***
log(display + 0.01)  0.1191618  0.0072348  16.471 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    582.2
Residual Sum of Squares: 193.38
R-Squared:      0.66785
Adj. R-Squared: 0.66665
F-statistic: 1671.87 on 2 and 1663 DF, p-value: < 2.22e-16
```

```
> summary(re2<- plm(e2,data=juice,effect="two",model="random"))
Twoways effects Random Effect Model
   (Swamy-Arora's transformation)

Call:
plm(formula = e2, data = juice, effect = "two", model = "random")

Balanced Panel: n=14, T=119, N=1666

Effects:
                var std.dev share
idiosyncratic 0.11581 0.34031 0.296
individual    0.21382 0.46241 0.546
time          0.06216 0.24931 0.159
theta  : 0.9327 (id) 0.6573 (time) 0.6562 (total)

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-1.53000 -0.20000 -0.00326  0.20400  1.29000

Coefficients :
                   Estimate Std. Error  t-value  Pr(>|t|)
(Intercept)        27.4487289  8.1047256   3.3868 0.0007237 ***
log(price)         -2.6787016  0.0635050 -42.1810 < 2.2e-16 ***
log(display + 0.01)  0.1191014  0.0072238  16.4873 < 2.2e-16 ***
educ                5.2150462  2.1255647   2.4535 0.0142500 *
income             -1.4192312  0.8058566  -1.7611 0.0783976 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    582.17
Residual Sum of Squares: 192.56
R-Squared:      0.66924
Adj. R-Squared: 0.66723
F-statistic: 840.175 on 4 and 1661 DF, p-value: < 2.22e-16
```

From above results if we compare OLS and FE and RE model outputs we can see the slight difference in the price/display value with respect to R-squared value and same p-value for all three models. The major difference is OLS with two and four independent variables it has different price/display value alone with R-squared value is higher when more independent variables are added. Whereas RE model has significantly similar price/display/R-squared value even with two and four variables, it gives information about the variance of the components of the errors and FE model has an error with four independent variables.

2. Test for H_0 : FE = RE

Results:

```
> ##Hausmann test
> phtest(fe1,re1)

        Hausman Test

data:  e1
chisq = 2.8798, df = 2, p-value = 0.237
alternative hypothesis: one model is inconsistent

> phtest(fe2,re2)

        Hausman Test

data:  e2
chisq = 2.6758, df = 2, p-value = 0.2624
alternative hypothesis: one model is inconsistent
```

From above results, `phtest(fe1,re1), phtest(fe2,re2)` the p-value is greater that of the significance level of 0.05, therefore, random effect model would be appropriate as alternative hypothesis conveys one model is inconsistent with fixed effect model and random effect model.

3. Compare price and display elasticities of FE and RE model.

Results:

```
> ##Price and Display elasticities for FE model
> coef(fe1)
        log(price) log(display + 0.01)
        -2.7183582          0.1179574
> coef(fe2)
        log(price) log(display + 0.01)
        -2.7183582          0.1179574
> ##Price and Display elasticities for RE model
> coef(re1)[2:3]
        log(price) log(display + 0.01)
        -2.6775468          0.1191618
> coef(re2)[2:3]
        log(price) log(display + 0.01)
        -2.6787016          0.1191014
```

If we compare above FE and RE models the price elasticities are in negative sign but it has minimal difference in price value due to FE includes dummy variables whereas RE omits dummy variables and even it compares to display elasticities is in positive but it has minimal difference for both models.

4. Compare HAC covariance matrix for FE and RE model.

Results:

```
> ##for FE model
> coeftest(fe1,vcovHC(fe1,method="arellano",cluster = "group"))

t test of coefficients:

                      Estimate Std. Error t value  Pr(>|t|)
log(price)          -2.7183582  0.0803159 -33.846 < 2.2e-16 ***
log(display + 0.01)  0.1179574  0.0057225  20.613 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> coeftest(fe2,vcovHC(fe2,method="arellano",cluster = "group"))
Error in solve.default(crossprod(demX)) :
  Lapack routine dgesv: system is exactly singular: U[3,3] = 0
> ##for RE model
> coeftest(re1,vcovHC(re1,method="arellano",cluster = "group"))

t test of coefficients:

                      Estimate Std. Error t value  Pr(>|t|)
(Intercept)         13.5974162  0.1736288  78.313 < 2.2e-16 ***
log(price)          -2.6775468  0.1181572 -22.661 < 2.2e-16 ***
log(display + 0.01)  0.1191618  0.0064736  18.407 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> coeftest(re2,vcovHC(re2,method="arellano",cluster = "group"))

t test of coefficients:

                      Estimate Std. Error  t value  Pr(>|t|)
(Intercept)         27.4487289  6.3879585   4.2969 1.832e-05 ***
log(price)          -2.6787016  0.1179485 -22.7108 < 2.2e-16 ***
log(display + 0.01)  0.1191014  0.0064563  18.4473 < 2.2e-16 ***
educ                 5.2150462  1.8391645   2.8356   0.00463 **
income              -1.4192312  0.6443937  -2.2024   0.02777 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From above results we can view FE does not have Intercept value whereas RE model has Intercept value and price/display value slightly shows the difference but the p-value is same for both models. Therefore, there is a difference in the T-value and standard error for both models. As we can see an error for FE model for four independent variables.

6. Dynamic Panel Models:

This model explain in the sense that it contains a lagged dependent variables. For simplicity let us consider equation with static panel model.

$$\text{The Mathematical Equation : } Y_{it} = b_0 + b_1 X_{it} + ind_i + time_t + u_{it} \ldots\ldots\ldots\ldots (1)$$

From above equation individuals i = 1,2…...N and time periods t = 1,2…...T. The dynamic panel model when lagged Y_{it} and X_{it} are used.

$$Y_{it} = b_0 + b_1 X_{it} + b_2 Y_{i(t-1)} + ind_i + time_t + u_{it} \ldots\ldots\ldots\ldots (2)$$
$$Y_{it} = b_0 + b_1 X_{it} + b_2 Y_{i(t-1)} + b_3 X_{i(t-1)} + ind_i + time_t + u_{it} \ldots\ldots\ldots\ldots (3)$$

From, above equations, only problem arises correlation between ind_i and $Y_{i(t-1)}$. It shows bias small for large T (>100) but severe in Large N, small T setting (*Nickell 1981*).

This model account for time-invariant heterogeneity of individuals by first differencing all variables (Δ) and use of LSDV-approach for time-periods where $time_t$: change to 1 in t and -1 in t+1 by Δ .

$$Y_{it} - Y_{i(t-1)} = b_1 [X_{it} - X_{i(t-1)}] + b_2 [Y_{i(t-1)} - Y_{i(t-2)} + time_t + u_{i(t-1)}$$
$$\Delta Y_{it} = b_1 \Delta X_{it} + b_2 \Delta Y_{i(t-1)} + \Delta time_t + \Delta u_{it}$$

From, above equation now the correlation exists between Δu_{it} and $\Delta Y_{i(t-1)}$. The first attempt is 2SLS with $Y_{i(t-2)}$ as instrument for $\Delta Y_{i(t-1)}$ *(Anderson and Hsiao 1981)* . The most prominent extension by Arellano and Bond (1991) makes use of more instruments (i.e. # of instruments is growing with number of time-periods: *t=3:* Y_1 */ t=4 :* Y_1 , Y_2 */ t=5 :* Y_1 , Y_2 , Y_3 */* ……. . Let us consider dynamic panel model by Arellano and Bond (1991) where we perform first-differencing all variables (Δ) to remove ind_i and use of LSDV-approach for time-periods ($\Delta time_t$).

$$\Delta Y_{it} = b_1 \Delta X_{it} + b_2 \Delta Y_{i(t-1)} + \Delta time_t + \Delta u_{it}$$

For individuals i, the matrix of instruments becomes

$$W_i = \begin{pmatrix} y_1 & 0 & 0 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & x_{i3} \\ 0 & y_1 & y_2 & 0 & 0 & 0 & \dots & 0 & 0 & 0 & 0 & x_{i4} \\ 0 & 0 & 0 & y_1 & y_2 & y_3 & \dots & 0 & 0 & 0 & 0 & x_{i5} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & \dots & \dots & y_1 & y_2 & \dots & y_{t-2} & x_{iT-2} \end{pmatrix}$$

Large number of instruments makes 2SLS inefficient for this solution is Generalized Method of Moments (GMM) for an example simple linear Regression with OLS versus GMM. *OLS:* $E[u_i, u_i] = E[(Y_i - b_1 X_i), (Y_i - b_1 X_i)] = 0$ *(i.e.* variance of u_i should be zero = minimize to achieve least squares*). GMM:* $E[X_i, u_i] = E[X_i, (Y_i - b_1 X_i)] = 0$ *(i.e.* covariance between X_i and u_i should be zero = minimize to achieve moment condition*).* Therefore, solving by b_1 leads in both cases to $b_1 = (X'X)^{-1}(X'Y)$. The *2SLS :* $E[Z_i, u_i] = E[X_i, (Z_i - b_1 X_i)] = 0$ and solving by b_1 leads to $b_1 = (Z'X)^{-1}(Z'Y)$. Consequently, moment condition in dynamic panel model is $E[A W_i, A \Delta u_i] = 0$ where W_i stands for matrix of GMM instruments for individual *i,* Δu_i stands for residuals for individual *i* $[= \Delta Y_{it} - b_1 \Delta X_{it} + b_2 \Delta Y_{i(t-1)} + time_t)]$ and A stands for weighting matrix (one step or two step) $A = (\Sigma_i W_i' H W_i)^{-1}$.

One-step H = $\begin{pmatrix} 2 & -1 & 0 & \dots & 0 \\ -1 & 2 & -1 & \dots & 0 \\ 0 & -1 & 2 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & -1 & 2 \end{pmatrix}$ and Two-step H = $\Sigma_i \Delta u_i \Delta u_i'$ with Δu_i as residuals from one-step estimation and solving the moment condition by b (b_1 , b_2) leads to $b_{gmm} = (\Delta X' W A W' \Delta X)^{-1} (\Delta X^{-1} W A W' \Delta Y)$. Now, Compared to OLS $b_{ols} = (X'X)^{-1}(X'Y)$ or GLS $b_{gls} = (X'_{itm} X_{itm})^{-1}(X'_{itm} Y_{itm})$ with X_{itm} and Y_{itm} as demeaned data matrix according to FE and RE specification.

6.1 Model Diagnostic for Dynamic Panel Models:

It required to run few test they are

1. Sargan Test : if H_0 : GMM instruments are valid/not-overidentified (H_0 good!).
2. Autocorrelation test :if H_0 : No autocorrelation (H_0 good!).
3. Wald Test (for coefficients and time-dummies)
   If H_0 : All b = 0 (H_0 bad!).

If H_0: All $time_t = 0$ (H_0 "bad", i.e. time dummies unnecessary).

4. Technically, R^2 not valid but computable by comparing predicted values and actual values.

6.2 Exercise (With Marketing Topic):

Firstly, open the file "employment.csv" to look into the dataset for the product classification with the columns heading names.

firm	year	emp	wage	capital	output
1	1977	5,0409999	13,1516	0,58939999	95,707199
1	1978	5,5999999	12,3018	0,6318	97,356903
1	1979	5,0149999	12,8395	0,6771	99,608299
1	1980	4,7150002	13,8039	0,6171	100,5501

From above figure "employment.csv" columns describe the distinct fields with various names to differentiate the data in the "employment.csv" file.

1. Estimate Arellano and Bond(1991) equation using OLS, 2-way-FE and 2-way 2-step GMM $\log(emp_{it}) = log(emp_{it-1}) + log(emp_{it-2}) + log(wage_{it}) + \log(wage_{it-1}) + \log(capital_{it}) + \log(output_{it}) + \log(output_{it-1})$

Results:

```
Oneway (individual) effect Pooling Model

Call:
plm(formula = log(emp) ~ lag(log(emp), 1:2) + lag(log(wage),
    0:1) + log(capital) + lag(log(output), 0:1), data = work,
    effect = NULL, model = "pooling")

Unbalanced Panel: n=140, T=5-7, N=751

Residuals :
   Min. 1st Qu.  Median 3rd Qu.    Max.
-0.6330 -0.0463  0.0038  0.0514  0.7360

Coefficients :
                         Estimate Std. Error  t-value  Pr(>|t|)
(Intercept)           -0.5339419  0.2527557  -2.1125   0.03498 *
lag(log(emp), 1:2)1    1.1434908  0.0336513  33.9806 < 2.2e-16 ***
lag(log(emp), 1:2)2   -0.1945458  0.0318698  -6.1044 1.662e-09 ***
lag(log(wage), 0:1)0  -0.5384417  0.0508817 -10.5822 < 2.2e-16 ***
lag(log(wage), 0:1)1   0.4856708  0.0514256   9.4441 < 2.2e-16 ***
log(capital)           0.0478119  0.0065705   7.2768 8.708e-13 ***
lag(log(output), 0:1)0 0.6874070  0.0862645   7.9686 6.036e-15 ***
lag(log(output), 0:1)1 -0.5244157  0.0829005  -6.3258 4.348e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares:    1350.9
Residual Sum of Squares: 9.5224
R-Squared:      0.99295
Adj. R-Squared: 0.98237
F-statistic: 14951.7 on 7 and 743 DF, p-value: < 2.22e-16
```

From above results, the coefficients named as "lag" with positive and negative values for emp and wage and output and the p-value is less that of the significance level of 0.05 all independent variables are significant with OLS with R-squared value is 99% is perfect to explain our model.

```
> summary(fe)
Twoways effects Within Model

Call:
plm(formula = log(emp) ~ lag(log(emp), 1:2) + lag(log(wage),
    0:1) + log(capital) + lag(log(output), 0:1), data = work,
    effect = "two", model = "within")

Unbalanced Panel: n=140, T=5-7, N=751

Residuals :
    Min.  1st Qu.   Median  3rd Qu.     Max.
-0.47500 -0.04350  0.00131  0.04460  0.45200

Coefficients :
                       Estimate Std. Error t-value  Pr(>|t|)
lag(log(emp), 1:2)1     0.700274   0.037557 18.6458 < 2.2e-16 ***
lag(log(emp), 1:2)2    -0.169027   0.035413 -4.7731 2.284e-06 ***
lag(log(wage), 0:1)0   -0.559251   0.056941 -9.8215 < 2.2e-16 ***
lag(log(wage), 0:1)1    0.292637   0.059360  4.9299 1.067e-06 ***
log(capital)            0.348173   0.026848 12.9682 < 2.2e-16 ***
lag(log(output), 0:1)0  0.490632   0.123499  3.9728 7.971e-05 ***
lag(log(output), 0:1)1 -0.607329   0.122101 -4.9740 8.584e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Total Sum of Squares:    15.748
Residual Sum of Squares: 5.3411
R-Squared:      0.66085
Adj. R-Squared: 0.52622
F-statistic: 166.463 on 7 and 598 DF, p-value: < 2.22e-16
```

From above results coefficients named as “lag” with positive and negative values for emp and wage and output and the p-value is less that of a significance level of 0.05 all independent variables are significant with FE with R-squared value is 66% is moderate to explain our model.

```
> summary(ab)
Twoways effects Two steps model

Call:
pgmm(formula = log(emp) ~ lag(log(emp), 1:2) + lag(log(wage),
    0:1) + log(capital) + lag(log(output), 0:1) | lag(log(emp),
    2:9), data = work, effect = "twoways", model = "twostep")

Unbalanced Panel: n=140, T=7-9, N=1031

Number of Observations Used:  611

Residuals
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.6191000 -0.0255700  0.0000000 -0.0001339  0.0332000  0.6410000

Coefficients
                       Estimate Std. Error z-value  Pr(>|z|)
lag(log(emp), 1:2)1     0.474151   0.185398  2.5575 0.0105437 *
lag(log(emp), 1:2)2    -0.052967   0.051749 -1.0235 0.3060506
lag(log(wage), 0:1)0   -0.513205   0.145565 -3.5256 0.0004225 ***
lag(log(wage), 0:1)1    0.224640   0.141950  1.5825 0.1135279
log(capital)            0.292723   0.062627  4.6741 2.953e-06 ***
lag(log(output), 0:1)0  0.609775   0.156263  3.9022 9.530e-05 ***
lag(log(output), 0:1)1 -0.446373   0.217302 -2.0542 0.0399605 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sargan Test: chisq(25) = 30.11247 (p.value=0.22011)
Autocorrelation test (1): normal = -1.53845 (p.value=0.12394)
Autocorrelation test (2): normal = -0.2796829 (p.value=0.77972)
Wald test for coefficients: chisq(7) = 142.0353 (p.value=< 2.22e-16)
Wald test for time dummies: chisq(6) = 16.97046 (p.value=0.0093924)
```

From above results, the coefficients named as “lag” with positive and negative values for emp and wage and output and the p-value are less that of a significance level of 0.05 for the few variables are significant with GMM. In conclusion, it has different p-values for different test methods. It also mentions that it is an unbalanced panel.

2. Compare the parameters estimates (i.e. elasticities) of OLS, FE and GMM model

Results:

	OLS	FE	GMM1	GMM2
lag(log(emp), 1:2)1	1.1434908	0.700273674	0.53461362	0.47415060
lag(log(emp), 1:2)2	-0.1945458	-0.169027073	-0.07506919	-0.05296749
lag(log(wage), 0:1)0	-0.5384417	-0.559250938	-0.59157311	-0.51320478
lag(log(wage), 0:1)1	0.4856708	0.292637302	0.29150961	0.22463981
log(capital)	0.0478119	0.348172571	0.35850245	0.29272309
lag(log(output), 0:1)0	0.6874070	0.490632101	0.59719848	0.60977482
lag(log(output), 0:1)1	-0.5244157	-0.607328613	-0.61170445	-0.44637259
1979	0.0000000	-0.001666133	0.00542719	0.01050897
1980	0.0000000	0.004244105	0.01646207	0.02465118
1981	0.0000000	-0.032347631	-0.01641563	-0.01580193
1982	0.0000000	-0.043549420	-0.03877363	-0.03744198
1983	0.0000000	-0.061715410	-0.04019665	-0.03928881
1984	0.0000000	-0.160093699	-0.02845569	-0.04950935

From above results, if the estimate onestep and twostep test method where independent variables are positive then OLS model will have greater value if it is negative value then GMM model will have a higher value.

3. Check the model diagnostics for the GMM model and check the elements of the `pgmm()` -object

Results:

```
> summary(ab,time.dummies=TRUE)
Twoways effects Two steps model

Call:
pgmm(formula = log(emp) ~ lag(log(emp), 1:2) + lag(log(wage),
    0:1) + log(capital) + lag(log(output), 0:1) | lag(log(emp),
    2:9), data = work, effect = "twoways", model = "twostep")

Unbalanced Panel: n=140, T=7-9, N=1031

Number of Observations Used:  611

Residuals
      Min.    1st Qu.     Median       Mean    3rd Qu.       Max.
-0.6191000 -0.0255700  0.0000000 -0.0001339  0.0332000  0.6410000

Coefficients
                        Estimate Std. Error z-value  Pr(>|z|)
lag(log(emp), 1:2)1     0.4741506  0.1853985  2.5575 0.0105437 *
lag(log(emp), 1:2)2    -0.0529675  0.0517491 -1.0235 0.3060506
lag(log(wage), 0:1)0   -0.5132048  0.1455653 -3.5256 0.0004225 ***
lag(log(wage), 0:1)1    0.2246398  0.1419495  1.5825 0.1135279
log(capital)            0.2927231  0.0626271  4.6741 2.953e-06 ***
lag(log(output), 0:1)0  0.6097748  0.1562625  3.9022 9.530e-05 ***
lag(log(output), 0:1)1 -0.4463726  0.2173020 -2.0542 0.0399605 *
1979                    0.0105090  0.0099019  1.0613 0.2885484
1980                    0.0246512  0.0157698  1.5632 0.1180087
1981                   -0.0158019  0.0267313 -0.5911 0.5544275
1982                   -0.0374420  0.0299934 -1.2483 0.2119056
1983                   -0.0392888  0.0346649 -1.1334 0.2570509
1984                   -0.0495094  0.0348578 -1.4203 0.1555141
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Sargan Test: chisq(25) = 30.11247 (p.value=0.22011)
Autocorrelation test (1): normal = -1.53845 (p.value=0.12394)
Autocorrelation test (2): normal = -0.2796829 (p.value=0.77972)
Wald test for coefficients: chisq(7) = 142.0353 (p.value=< 2.22e-16)
Wald test for time dummies: chisq(6) = 16.97046 (p.value=0.0093924)
```

From above results, sargan test gives the p-value is greater than 0.05 then it is valid (H_0 = good!), Autocorrelation test 1 and 2 gives the p-value is greater than 0.05 then no autocorrelation (H_0 = good!). The wald test for coefficient and time dummies the p-value is less that of the significance level of 0.05 we can reject the null hypothesis so it is not equal to zero.

pgmm():

```
> ab$residuals
$`1`
       1979         1980         1981         1982         1983         1984
 0.000000000  0.041244258 -0.008706737 -0.103947845  0.001997046  0.000000000

$`2`
       1979         1980         1981         1982         1983         1984
 0.000000000  0.006602207  0.097298191 -0.008137535 -0.070165151  0.000000000

  ...  ...........  ...............  ...............  ...............  ...........  ......

 $`139`
       1979        1980        1981        1982        1983        1984
 0.03429461 -0.09399989  0.08906512  0.06027159  0.18436481 -0.14110082

 $`140`
        1979         1980         1981         1982         1983         1984
 0.022742342  0.040764077  0.064547248 -0.044925652 -0.001889612 -0.013009632
```

```
> ab$model
$`1`

  ...  ... • •  ......  ... • •  ......  ... • •  ......  ......  ......  ......  ......  ... • •

$`140`
      log(emp) lag(log(emp), 1:2)1 lag(log(emp), 1:2)2 lag(log(wage), 0:1)0 lag(log(wage), 0:1)1
1979  0.009081446        -0.015320634        -0.062939109          0.050443058          0.070972722
1980 -0.026066992         0.009081446        -0.015320634         -0.025753630          0.050443058
1981 -0.052007325        -0.026066992         0.009081446          0.073287690         -0.025753630
1982 -0.109546374        -0.052007325        -0.026066992          0.004233845          0.073287690
1983 -0.048119248        -0.109546374        -0.052007325          0.010081098          0.004233845
1984 -0.038587020        -0.048119248        -0.109546374         -0.007215116          0.010081098
     log(capital) lag(log(output), 0:1)0 lag(log(output), 0:1)1
1979  -0.06477344            0.023951845           0.013306240  1  0  0  0  0 0
1980  -0.11358141           -0.109384309           0.023951845 -1  1  0  0  0 0
1981  -0.22204910           -0.006053284          -0.109384309  0 -1  1  0  0 0
1982  -0.12318779           -0.001040840          -0.006053284  0  0 -1  1  0 0
1983  -0.12827687            0.075620620          -0.001040840  0  0  0 -1  1 0
1984  -0.01966739            0.057728110           0.075620620  0  0  0  0 -1 1
```

```
> ab$W
$`1`

...  ... • •  ......  ... • •  ......  ... • •  ......  ......  ......  ......  ......  ... • •

$`140`

1979 0.4324316 0.3694924 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
1980 0.0000000 0.0000000 0.4324316 0.3694924 0.3541718 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
1981 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.4324316 0.3694924 0.3541718 0.3632533 0.0000000
1982 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.4324316
1983 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
1984 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000

1979 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
1980 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
1981 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
1982 0.3694924 0.3541718 0.3632533 0.3371863 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
1983 0.0000000 0.0000000 0.0000000 0.0000000 0.4324316 0.3694924 0.3541718 0.3632533 0.3371863 0.2851789
1984 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000
                                                                                                 log(capital)
1979 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  1  0  0  0  0 0  -0.06477344
1980 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 -1  1  0  0  0 0  -0.11358141
1981 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0 -1  1  0  0 0  -0.22204910
1982 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0 -1  1  0 0  -0.12318779
1983 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000 0.0000000  0  0  0 -1  1 0  -0.12827687
1984 0.4324316 0.3694924 0.3541718 0.3632533 0.3371863 0.2851789 0.1756326  0  0  0  0 -1 1  -0.01966739
     log(output) lag(log(output), 1)    log(wage) lag(log(wage), 1)
1979  0.023951845          0.013306240  0.050443058       0.070972722
1980 -0.109384309          0.023951845 -0.025753630       0.050443058
1981 -0.006053284         -0.109384309  0.073287690      -0.025753630
1982 -0.001040840         -0.006053284  0.004233845       0.073287690
1983  0.075620620         -0.001040840  0.010081098       0.004233845
1984  0.057728110          0.075620620 -0.007215116       0.010081098
```

```
> ab$A1
  Output is too large to capture
> dim(ab$A2)
  [1] 38 38
```

From the above results, pgmm() function generate 140 sample outputs in detail for the independent variables values.

4. Compare the R^2 for the OLS, FE and GMM model.

Results:

```
> ###Functions to compute R2
> r.squared(ols)
[1] 0.992951
> r.squared(fe)
[1] 0.6608513
> gmmR2(ab)
     GMM_R2
1 0.3999345
```

From above results, the R-squared value is different for each model whereas OLS model computes 99% and FE model computes 66% and GMM computes 40%. Which explains that our model is a good fit or not and with the OLS model conveys our model is perfect and FE model conveys our model is moderate and GMM model conveys our model is average.

Bibliology:

1. Baltagi, B (2013). *Econometric Analysis of Panel Data*. 5th Edition. John Wiley & Sons Ltd.

2. Croissant Y, Millo G (2008). *Panel Data Econometrics in R*: The plm Package. Journal of Statistical Software 27(2), 1-43.

3. Arellano, M. (1987). Computing robust standard errors for within group estimators. *Oxford Bulletin of Economics and Statistics*, 49, 431–434.

4. Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, 58, 227–297.

5. Bijmolt, T. H., Heerde, H. J. V., and Pieters, R. G. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research*, 42, 141-156.

6. Rossi, P. E. (2014). Even the Rich Can Make Themselves Poor: A Critical Examination of IV Methods in Marketing Applications. *Marketing Science*, 33, 655-672.

7. Swamy, P. and Arora, S.S. (1972). The exact finite sample properties of the estimators of coefficients in the error components regression models. *Econometrica*, 40, 261–275.

8. Amemiya, T. (1971) The estimation of the variances in a variance–components model. *International Economic Review*, 12, 1–13.

9. Anderson T, and Hsiao C (1981). Estimation of Dynamic Models With Error Components. Journal of the American Statistical Association, 76, 598-606.

10. Ebbes, P., Wedel, M., Böckenholt, U., and Steerneman, A. G. M. (2005). Solving and Testing for Regressor-Error (in)Dependence When no Instrumental Variables are Available: With New Evidence for the Effect of Education on Income. *Quantitative Marketing and Economics*, 3, 365–392.

11. Kim, J. S., and Frees, E. W. (2007). Multilevel Modeling with Correlated Effects. *Psychometrika 72*, 505-533.

12. Lewbel, A. (1997). Constructing Instruments for Regressions with Measurement Error when No Additional Data Are Available, with An Application to Patents and R&D. *Econometrica*, 65, 1201-1213.

13. Nerlove, M. (1971) Further evidence on the estimation of dynamic economic relations from a time–series of cross–sections. *Econometrica*, 39, 359–382.

14. Nickell, S (1981) Biases in Dynamic Models with Fixed Effects. Econometrica, 49, 1417-1426 Park, S., and Gupta, S. (2012). Handling Endogenous Regressors by Joint Estimation using Copulas. *Marketing Science*, 31, 567-586.

15. Wallace, T.D. and Hussain, A. (1969) The use of error components models in combining cross section with time series data. *Econometrica*, 37, 55–72.