**Fall 2016 –** **OPIM 5503 Data Analytics Using R**

**Instructor: Ram Gopal**



**Team- Random R**

**Topic – Panel Data Analysis**

**Ashish Gupta, Krishna Muddodi, Nachiket Garge, Rajit Podder, Sai Pramod Kotla**

**Executive Summary:**

Through this report we aim to display panel data modelling using PLM package in R. We as team predicted sales for Rossmann stores for 1115 stores using 2013-2015 historic sales data. Store sales is typical panel problem, frequently observed when different stores are different in terms of store characteristics and we have sales data over time. It is a combination of cross-sectional and time series data. Some other examples of panel data can be census data over years, user rating for same set of users over year etc. Panel data is more difficult to collect as, we need to collect same observations over different periods, but it is more effective way than simple linear regression as it removes individual ability caused by separate entities.

In this report we have covered exploratory analysis of Rossmann Sales data, where we have visualized trends of store sales and other interesting patterns observed in the data. Later we have prepared data for panel data regression and used five estimation techniques to predict sales. Finally, we have compared all the techniques and modelled using most appropriate estimation technique to predict Rossmann sales.

**Business Objective:**

Rossmann operates over 3,000 drug stores in 7 European countries. Daily thousands of customers with different demands and behavior visit Rossmann. Store sales are influenced by many factors, including promotions, competition, school and state holidays, seasonality, and locality. Reliable sales forecasts enable store managers to create effective staff schedules that increase productivity and motivation.

The business objective of carrying out this exercise was to predict the sales across all the stores for the panel data. Our aim is to create robust models that account seasonal effects as well as individual store effects in order to predict the sales. Accurate prediction of sales will help managers to stay focused on what’s most important to them: their customers and their teams!

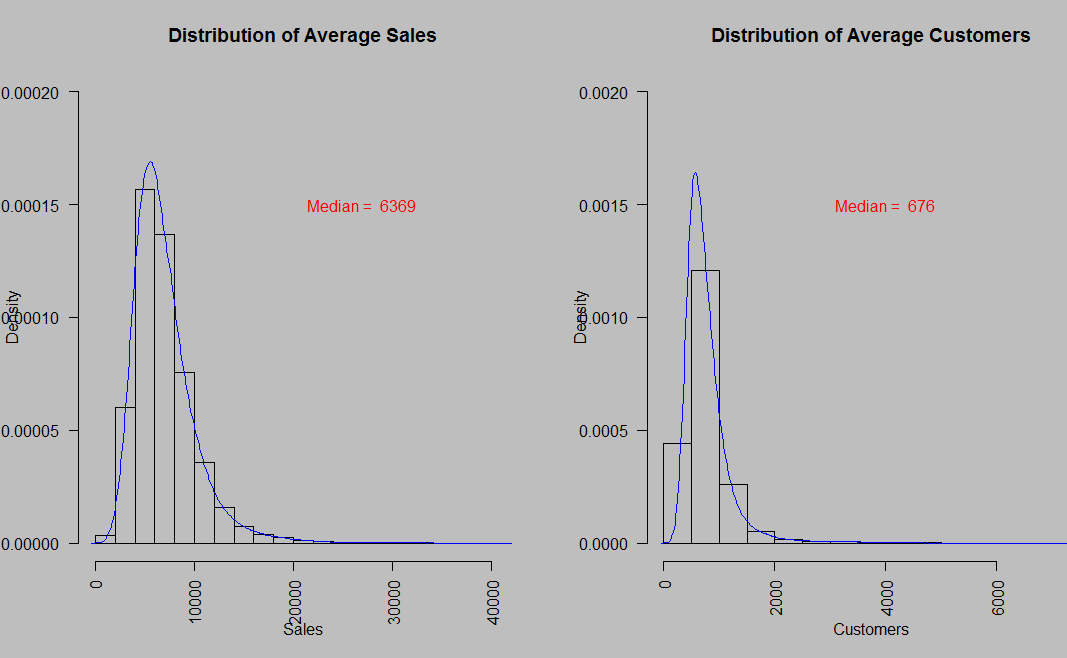
**Data:**

The dataset contains historical sales data for 1,115 Rossmann stores over a period from (01-01-2013 to 07-31-2015) with total of 1,017,209 observations. We had two csv files. One containing store related data like Store Establishment Year, Assortment etc. and one containing sales data over time for different stores.

The following are descriptions for the fields in data:

1. Id - an Id that represents a (Store, Date) tuple within the test set
2. Store - a unique Id for each store
3. Sales - the turnover for any given day (this is what you are predicting)
4. Customers - the number of customers on a given day
5. Open - an indicator for whether the store was open: 0 = closed, 1 = open
6. StateHoliday - indicates a state holiday. Normally all stores, with few exceptions, are closed on state holidays. Note that all schools are closed on public holidays and weekends. a = Public holiday, b = Easter holiday, c = Christmas, 0 = None
7. SchoolHoliday - indicates if the (Store, Date) was affected by the closure of public schools
8. StoreType - differentiates between 4 different store models: a, b, c, d
9. Assortment - describes an assortment level: a = basic, b = extra, c = extended
10. CompetitionDistance - distance in meters to the nearest competitor store
11. CompetitionOpenSince[Month/Year] - gives the approximate year and month of the time the nearest competitor was opened
12. Promo - indicates whether a store is running a promo on that day
13. Promo2 - Promo2 is a continuing and consecutive promotion for some stores: 0 = store is not participating, 1 = store is participating
14. Promo2Since[Year/Week] - describes the year and the calendar week when the store started participating in Promo2
15. PromoInterval - describes the consecutive intervals Promo2 is started, naming the months the promotion is started anew. E.g. "Feb,May,Aug,Nov" means each round starts in February, May, August, November of any given year for that store

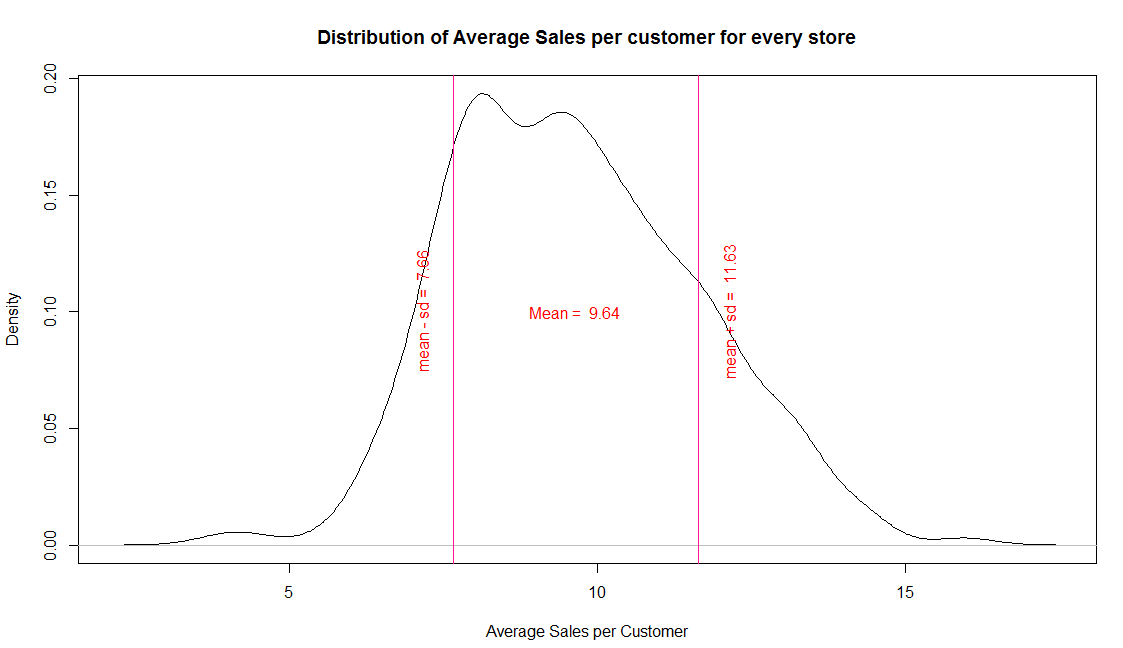
**Exploratory Analysis:**



The plot on the left shows the density plot for the distribution of average sales. The median of the sales across all stores is marked on the plot. It can be observed that the distribution of average sales is right skewed. Most of stores has average sales just above 5000.

The plot on the right shows the density plot *for* the distribution of average customers on all the stores. The median of the number of customers across all stores is marked on the plot. It can be observed that the distribution of average customers is also right skewed. Most of stores has average customers around 800.

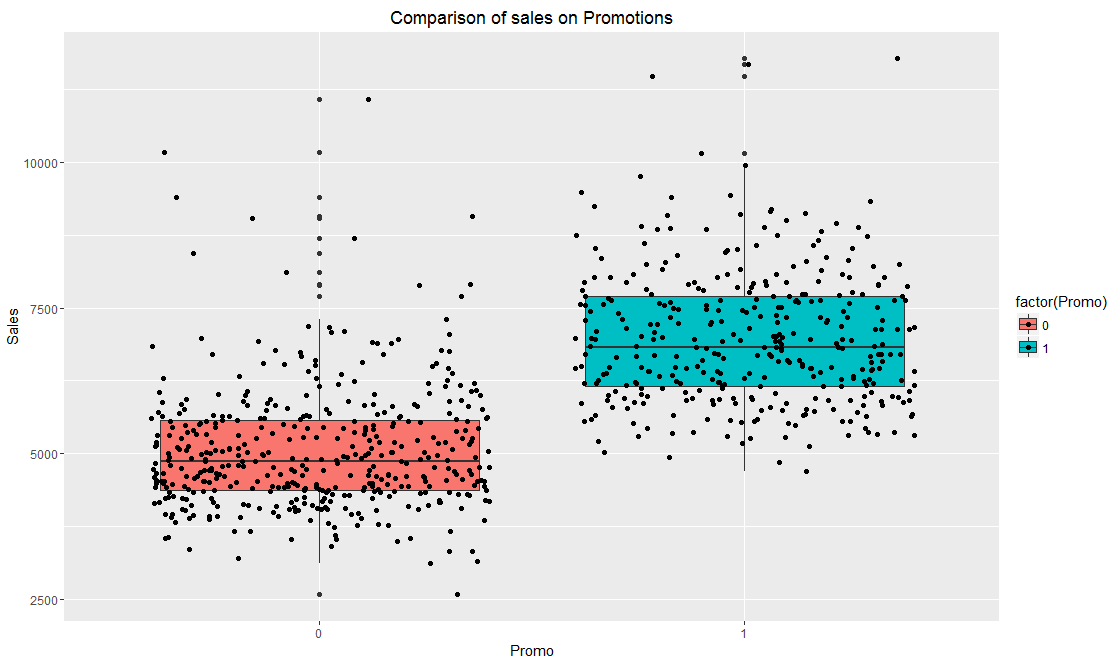
It can be calculated from the two medians that the average customer shops for around $9.421 from a store.



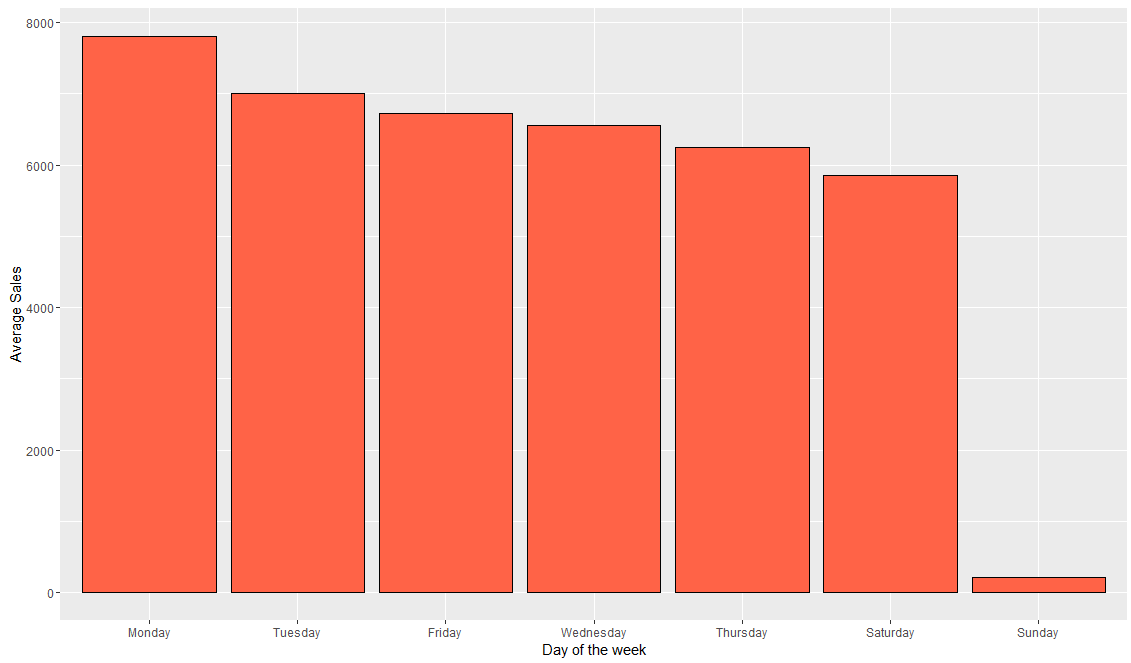
The plot above shows the distribution of average sales per customer across every store. The mean of average sales per customer on any store is around $9.64. A line has been drawn to mark the average sales per customer one standard deviation on the left and right of the mean. It can be inferred from the plot that most of the of stores has an average sale per customer between 7.65 and 11.63.



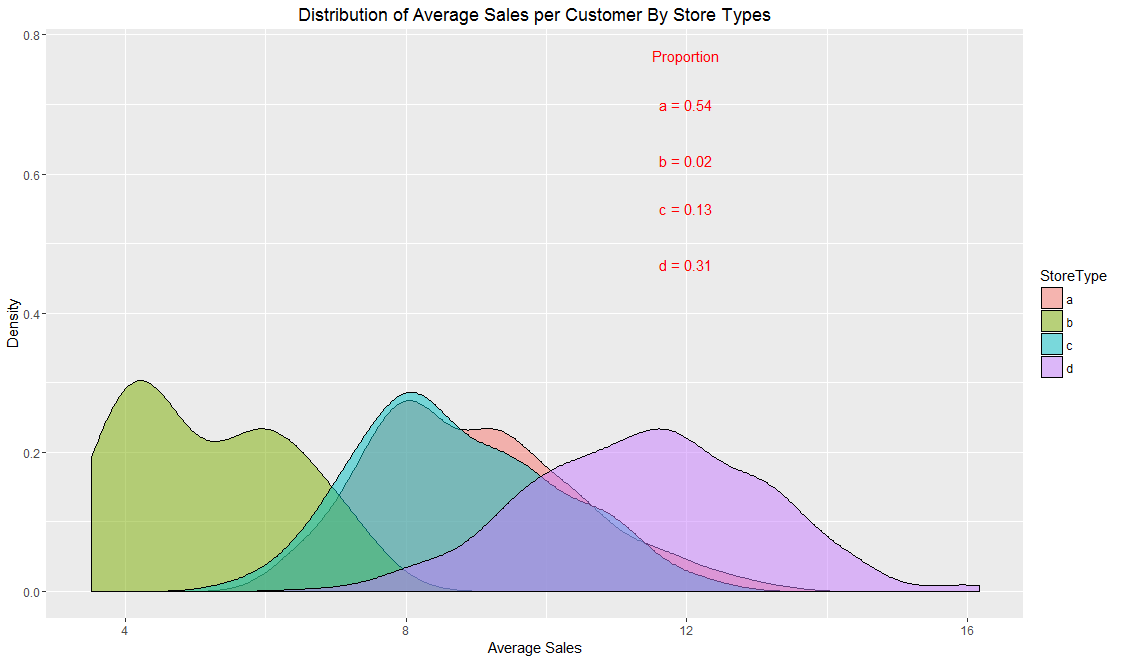
The boxplot above shows the comparison between sales on any randomly sampled store when there was a school holiday and when there was not. It can be observed that the median sales did not differ much depending on whether there was a school holiday or not. When another random sample was run, the outcome was same every time which proved that school holiday did not affect sales on all stores.



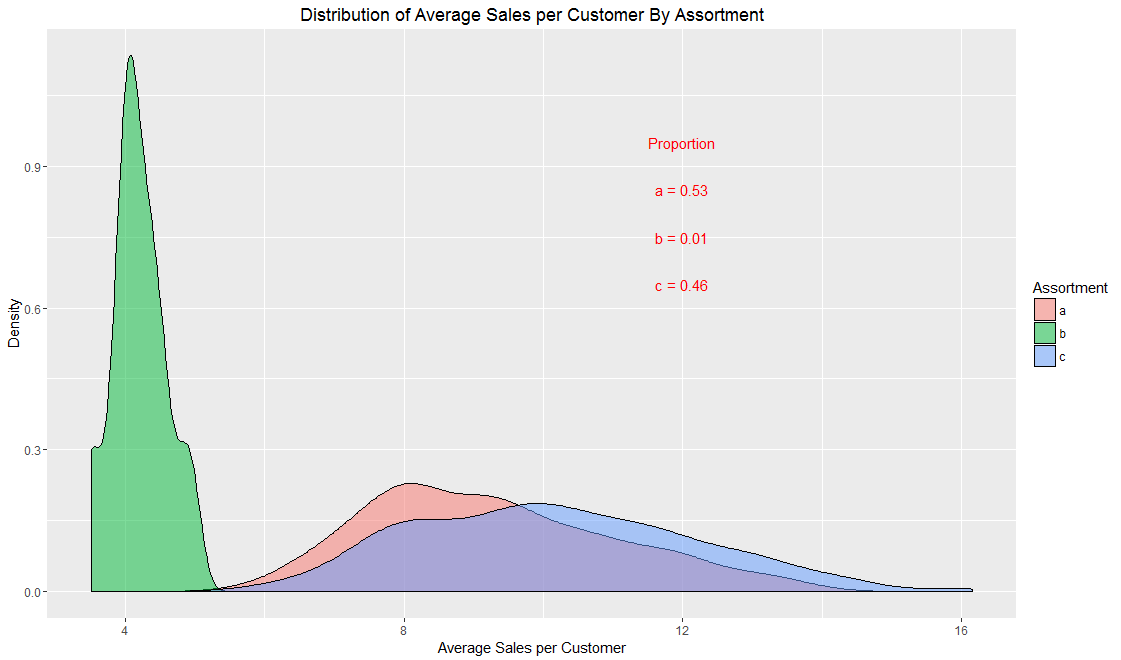
The boxplot above shows the comparison between sales on any randomly sampled store when it had promotions on offer and when it did not. It can be clearly observed that when a store had promotions running, the median sales was much higher than when it did not have promotions. When another random sample was run, the outcome was same every time which proved that sales were affected by promotions across all stores.



The bar plot above shows the comparison of average sales across the days of the week. The days of the week has been arranged from the highest average sales to the lowest average sales. It can be inferred from the plot above that average sales across all stores was most on Monday and Sunday had drastically low average sales compared to the rest of the week.



The plot above shows the distribution of average sales per customer of every store based on store type. The plot also displays the proportion of each store type a, b, c and d. It can be inferred from the plot that assortment type ‘b’ has the lowest average sales per customer, whereas assortment type ‘d’ has the highest average sales per customer. It should also be noted that store type ‘b’ are very few in number.



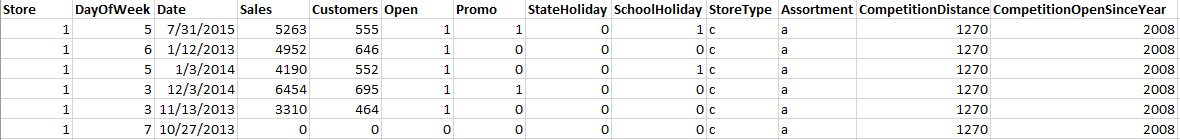
The plot above shows the distribution of average sales per customer of every store based on assortment type. The plot also displays the proportion of each assortment type a, b and c. It can be inferred from the plot that assortment type ‘b’ has the lowest average sales per customer, whereas assortment type ‘c’ has the highest average sales per customer. It should also be noted that assortment type ‘b’ are very few in number across all stores.

**Panel Data:**

Cross-sectional data, or a cross section of a study population, in statistics and econometrics is a type of data collected by observing many subjects (such as individuals, firms, countries, or regions) at the same point of time, or without regard to differences in time.

**Panel data** (also known as longitudinal or cross- sectional time-series data) is a dataset in which the behavior of entities are observed across time. These entities could be states, companies, individuals, countries, etc. In statistics and econometrics, the term panel data refers to multi-dimensional data frequently involving measurements over time. Panel data contain observations of multiple phenomena obtained over multiple time periods for the same firms or individuals. So basically they are cross-sectional and time series data.

Eg. In our data we have historical daily sales data for 1,115 Rossmann stores over 2.5 years of time span.



The above figure shows snapshot of data used, where we can observe data for Store = 1, similarly we have data for 1,115 stores repeated over time.

**Why to perform panel data modelling when we can perform linear regression on the same data?**

The basic assumptions of linear regression are:

1. Mean of response variable is linearly related to predictor variables
2. Error terms are normally distributed
3. Errors have equal variance (homoscedastic)
4. Errors are independent at each predictor value

Also for stability of our model we don’t want any correlated predictors in our model. If these assumptions don’t hold good the OLS estimates are BIASED and/or INEFFICIENT.

Biased - Expected value of parameter estimate is different from true.  
o If an estimator is unbiased, or if the bias shrinks as the sample size increases, we say it is **consistent.**

Inefficient - Estimator is less accurate as sample size increases than an alternative estimator.  
o Estimators that take full advantage of information more **efficient**

General equation for Panel Data Model is given by:



Here the error term consists of two components, an “idiosyncratic” component u and an “unobserved heterogeneity” component c, which is time-invariant and explains unobserved variability between individuals.

If the unobserved heterogeneity i.e. c is correlated with one or more of the explanatory variables, OLS parameter estimates are biased and inconsistent. While on the other hand, Panel data modelling estimation methods take care of “unobserved heterogeneity” by either first differencing or taking de-mean values with time for individuals.

If we have more than one observation on any individual, the errors will be correlated and OLS estimates will be inefficient

Using fixed and random panel models (explained later) addresses the issues related to “unobserved heterogeneity” and thus we use panel modelling techniques over simple linear regression.

**Advantages of panel data:**

1. Since panel data relate to individuals, firms, states, countries, etc., over time, there is bound to be heterogeneity in these units. The techniques of panel data estimation can take such heterogeneity explicitly into account by allowing for individual-specific variables, such as individuals, firms, states, and countries.
2. By combining time series of cross-section observations, panel data give “more informative data, more variability, less collinearity among variables, more degrees of freedom and more efficiency.
3. By studying the repeated cross section of observations, panel data are better suited to study the dynamics of change. Spells of unemployment, job turnover, and labor mobility are better studied with panel data.
4. Panel data can better detect and measure effects that simply cannot be observed in pure cross-section or pure time series data. For example, the effects of minimum wage laws on employment and earnings can be better studied if we include successive waves of minimum wage increases in the federal and/or state minimum wages.
5. Panel data enables us to study more complicated behavioral models. For example, phenomena such as economies of scale and technological change can be better handled by panel data than by pure cross-section or pure time series data.
6. By making data available for several thousand units, panel data can minimize the bias that might result if we aggregate individuals or firms into broad aggregates.

**Type of Regressors in Panel Data**

1. **Varying regressor** Xit – These are the regressors which depend on both i (in this case store) and time. e.g. Number of customers – They vary with time for the same stores that is they show within group variability. Also they show between group variability i.e. the value for customers in different stores for same time is different.
2. **Time-invariant regressors** Xit = Xi for all t – These are regressors which are time independent e.g. In this case all attributes corresponding to stores like Assortment type, Store type, Competition Distance and Competition Open Since Year all are time invariant. These variables have within group variation equals to 0 i.e. for the same store they don’t change over time.
3. **Individual-Invariant Regressors** Xit = Xt for all i – These variables do not depend on individual ids (in this case stores), they depend only on time, i.e. these variables have within group variation for a given time equals to 0. e.g. School Holiday, State Holiday depends on the date but not on individual store.

* Please refer appendix to learn more on between group and within group variation.

**Panel Data Models**

Panel data models describe the individual behavior both across time and across individuals. There are three types of models: the pooled model, the fixed effects model, and the random effects model.

1. **Pooled Model**: The pooled model specifies constant coefficients, the usual assumptions for cross-sectional analysis. Mathematically,



**Individual-specific effects model**

We assume that there is unobserved heterogeneity across individuals captured by alpha. Example: unobserved ability of an individual that affects sales of stores. The main question is whether the individual-specific effects alpha is correlated with the regressors. If they are correlated, we have the fixed effects model. If they are not correlated, we have the random effects model.

1. **Fixed effects model (FE):** The FE model allows the individual-specific effects alpha to be correlated with the regressors x. We include alpha as intercepts. Each individual has a different intercept term and the same slope parameters.



We can recover the individual specific effects after estimation as:



In other words, the individual-specific effects are the leftover variation in the dependent variable that cannot be explained by the regressors.

1. **Random effects model (RE):** The RE model assumes that the individual-specific effects alpha are distributed independently of the regressors.

We include alpha in the error term.





To estimate these models, we have different techniques.

**Panel Data Estimators**

Panel data models can be estimated with several estimators. The estimators differ on whether they consider between or within variation in the data.

We prefer estimators that are consistent and efficient. Different estimators used are:

1. Pooled OLS Estimator
2. Between Estimator
3. First Differences Estimator
4. Fixed Effects or Within Estimator
5. Random Effects Estimator

**Pooled OLS Estimator**

Here we pool all observations together and run regression model, neglecting the cross-section and time series nature of the data. In our example, by Combining all 1,115 stores for entire 2.5 years we deny the heterogeneity or individuality of the stores. In simpler words we are assuming that all the stores are same.

Basically here we completely ignore the fact that we are using panel data. If the pooled model has uncorrelated errors, then in that case model is consistent otherwise it will not be consistent.

The pooled OLS estimator is obtained by stacking the data over i and t into one long regression with NT observations and estimating it by OLS. The equation for pooled estimators will be,



**Between Estimator**

In between estimation technique we average all the rows over time and thus removing all the time information we have. So it considers variation that are there only “between” the cross-sectional data, in this case between different stores.

Thus, in between estimators we have, OLS estimation of the time-averaged dependent variable on the time-averaged regressors for each individual.



This model is seldom used, as it averages out the time component, and lot of essential data is lost. We miss to model the seasonal changes. The Random Effect and pooled estimators depict the between variations along with time variations and thus are more efficient than this model.

**Within Estimator or Fixed Effect Estimator**

In this method of estimating we observe the “within” group effects over time. This estimator uses all the NT rows for estimation. It uses time-demeaned variables (the individual-specific deviations of variables from their time-averaged values). So basically if we take number of customers of an individual store on a particular day and subtract it from average number of customer of that store over time, I can time-demeaned value for the number of customers for that particular store.

Mathematically,

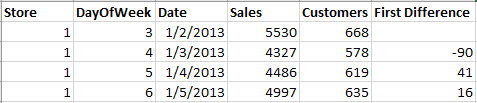


Taking time demeaned values removes the individual specific time invariant component alpha observed in pooled OLS estimates and thus provides better estimation.

The limitation of this estimation technique is that, time-invariant variables are dropped off and their coefficients cannot be measured. For example, in our case store assortment type and store type are same over time for a particular store so their average value will remain same and cancel out from the equation. If there are many time-invariant variables which are critical we need to use different estimation techniques like pooled OLS or between estimation.

**First-Differences Estimator**

The first difference estimator uses the one-period changes for each individual. It uses first-differenced variables (the individual-specific one-period changes for each individual). For example, if we have 4 rows for a store as shown below,



The first difference is null, -90, 41 and 16. That is it takes the difference of number of customers from the previous day value. This is an OLS estimation of the one-period changes of the dependent variable on the one-period changes in the regressors.

Mathematically,

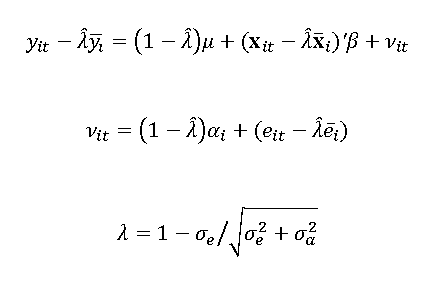


As individual specific effect alpha does not vary over time it gets cancelled out every time we subtract from previous time value. The limitation of first differences estimator is same as fixed estimators. It does not account for the variation due to time invariant variables like store type.

**Random Effects Estimator**

It is a transformed model which is combination of pooled OLS and fixed effect model. It incorporates both between variation and within variation.

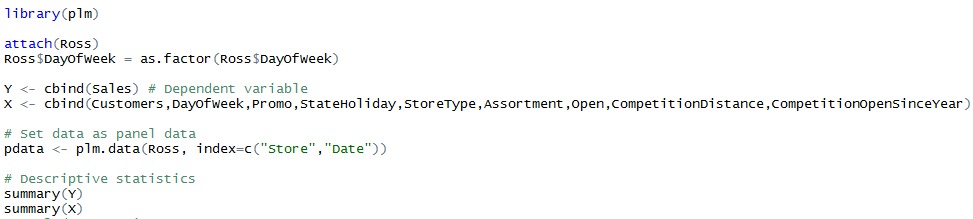
Mathematically,



We can observe the individual-specific effects alpha in the error term. As the value of lambda increase this individual effect will decrease. As lambda approaches 0, our estimation becomes a pooled OLS estimation, and as lambda approaches 1 our estimation acts like a fixed effect estimation. The random effects estimates are a weighted average of the between and within estimates. Random Effect Estimators are inconsistent when we have a fixed effect model and thus cannot be used in all cases.

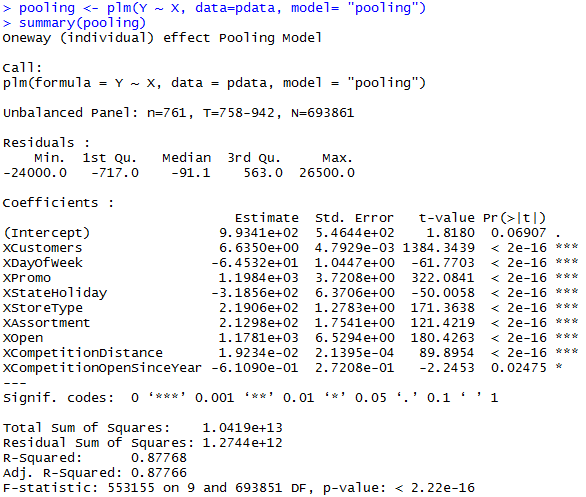
**R Code and Results:**

* To start with we loaded plm package and added plm library.
* Dependent and Independent variables are identified
* Data was converted to panel data with id as store and t as Date.
* After this various estimation techniques were applied.



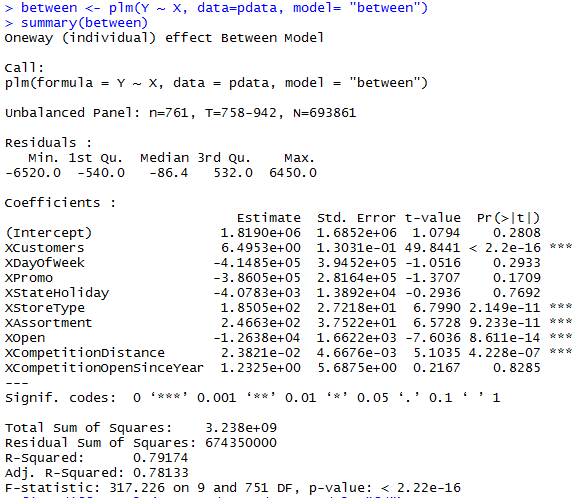
We employed the following estimations for our model building.

1. **Pooling Estimating**



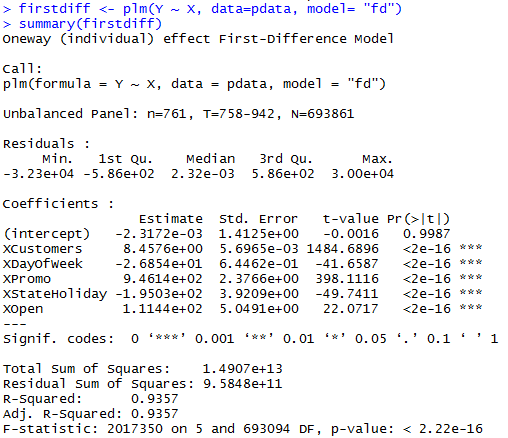
p-value of the overall model is less than 0.05 indicating that there is at least one variable that significantly influences sales. The simple pooling model gives us R square of 87.76% i.e. our model can predict about 87% of variation with the help of above model. Most of the variables are significant contributors to predict sales. If customers visiting store increases by 1 sales will increase by 6.635, like this we can interpret other results.

1. **Between Estimating**



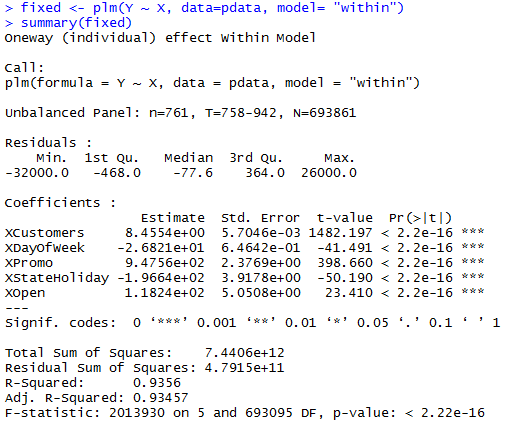
The between model as mentioned before collapses the time variable and thus gives a poor performance compare to pooled model observed before. The R square value is less and residual sum of square is more clearly indicating poor model performance.

1. **First Difference**



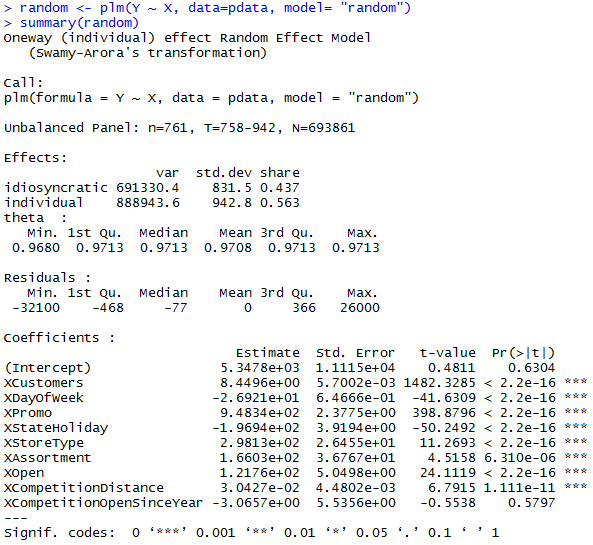
First difference model predicts the sale well and overall model is significant. With overall R square of 0.9357, model performance is better than pooled OLS estimator and between estimator observed earlier. Further error tests and consistency checks should be done before using this model for prediction. The parameters indicate that when store is on promotion sales increase by approximately 1000. This number is high and should be checked before use.

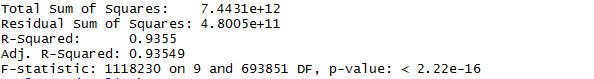
1. **Within or Fixed Estimation**



Fixed estimators also provide a decent model with high R square value of 0.93, the model explains about 93% of variation in the data.

1. **Random Estimation:**





Random model also shows high value of R square and low Residual Sum Squares. As mentioned earlier we need to use estimators which are consistent and efficient. To check and compare different models we ran various statistical tests.

**Which is the Best Model?**

R square and RMSE values are very close for all models. But, coefficients for OLS (Pooled) and Individual-specific effects model are very different from each other, which indicates that unobserved heterogeneity is important and OLS is inconsistent and biased. Therefore, we need an estimation technique that is consistent and efficient.

To compare different models, we have specific tests as shown below.

1. **plmtest** – This test is compare pooled test and random effect results.

Null hypothesis: Pooled estimations are appropriate

Alternative hypothesis: Random Effect estimations are appropriate

Result: p value < 0.05 so we reject the null hypothesis, and we can say that Random Effect estimation is better than pooled OLS estimation.

1. **pFtest** - This test is compare pooled OLS estimation and Fixed effect results.

Null hypothesis: Pooled estimations are appropriate

Alternative hypothesis: Fixed Effect estimations are appropriate

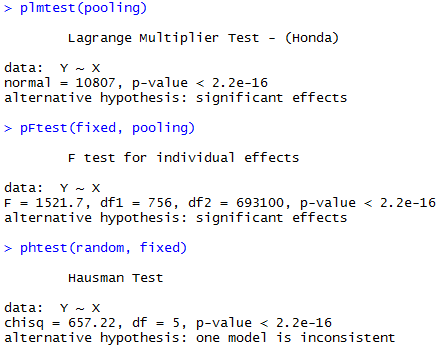
Result: p value < 0.05 so we reject the null hypothesis, and we can say that Fixed Effect estimation is better than pooled OLS estimation.

1. **phtest** - This test is compare Fixed Effect estimation and random effect results.

Null hypothesis: Random Effect estimations are appropriate

Alternative hypothesis: Fixed Effect estimations are appropriate

Result: p value < 0.05 so we reject the null hypothesis, and we can say that Fixed Effect estimation is better than Random Effect Model and Random effect model is inconsistent.

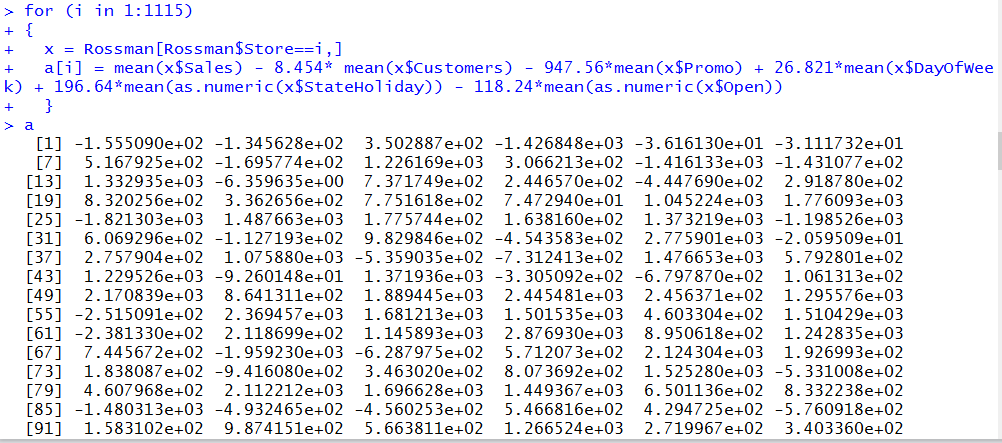


**The Model:**

All above tests indicate that Fixed Effects Model would be the best choice. Also, we choose within estimation technique for model parameter estimation.

**it = αi it**Here αi (individual specific effect) captures the unobserved variation for an individual and it is “idiosyncratic error”. After estimation, αi can be recovered as:





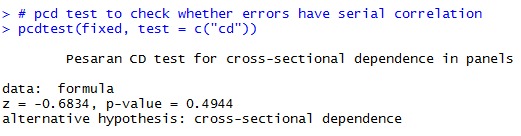
αi is calculated by averaging out the within variation. Thus for every individual, we’ll have different αi (constant) that represents effect of variables pertaining to stores only (individual specific effect) and capture the leftover variation in the sales.

To check the behavior of errors we performed **pcdtest**.

Null hypothesis: No serial Correlation

Alternate hypothesis: There exists serial correlation.

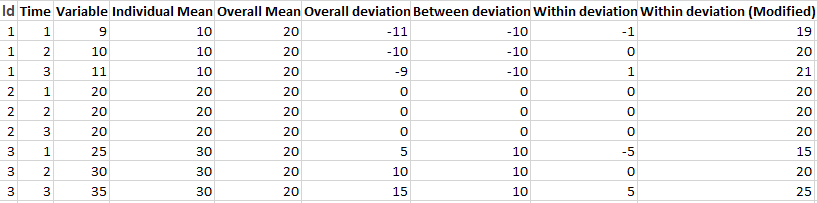
Observing the results, we see that p>0.05, thus we fail to reject the null hypothesis that errors have serial correlation.



**Appendix:**

Variation for the dependent variable and regressors

1. **Overall variation**: variation over time and individuals.
2. **Between variation**: variation between individuals.
3. **Within variation**: variation within individuals (over time).



**References:**

[1] <https://en.wikipedia.org/wiki/Panel_data>

[2] <https://sites.google.com/site/econometricsacademy/econometrics-models/panel-data-models>

[3] <http://documentslide.com/documents/panel-data-565f38d16381a.html>

[4]<https://jrvargas.files.wordpress.com/2011/01/wooldridge_j-2002_econometric_analysis_of_cross_section_and_panel_data.pdf>

[5] <http://www.indiana.edu/~wim/docs/10_7_2011_slides.pdf>

[6] <http://statmath.wu.ac.at/~hauser/LVs/FinEtricsQF/FEtrics_Chp5.pdf>