**ECONOMETRICS TIME SERIES KINGDOM OF SAUDI ARABIA**

**By**

**Ravichandra**

**Batch: 23**

**Email:** [**ravi.metrockz@gmail.com**](mailto:ravi.metrockz@gmail.com)

Contents

[Purpose 3](#_Toc479924137)

[1.1 Problem Description 3](#_Toc479924138)

[1.2 Project scope 3](#_Toc479924140)

[1.3 Constraints 3](#_Toc479924141)

[2. State-of-the-art 3](#_Toc479924142)

[2.1 Introduction to FDIs 3](#_Toc479924143)

[2.2 Trending Solutions 5](#_Toc479924144)

[Variable importance in Random Forests: 6](#_Toc479924145)

[Time Series Regression: Lagged Variables 14](#_Toc479924146)

[Vector Autoregressive models VAR(p) models 16](#_Toc479924147)

[3. Method 17](#_Toc479924148)

[3.1 Possibilities 17](#_Toc479924149)

[3.2 Action points 19](#_Toc479924150)

[3.3 Assumptions 19](#_Toc479924151)

[4. Data 19](#_Toc479924152)

[4.1 Data Processing 20](#_Toc479924153)

[5. Results 20](#_Toc479924154)

[5.1 Multiple Linear Regression values 20](#_Toc479924155)

[5.2 stepAIC 22](#_Toc479924156)

[5.3 Below plot showing the top 30 variables obtained from RF 26](#_Toc479924157)

[5.4 Linear Regression assumption plots 27](#_Toc479924158)

[5. 4.1 Below plot showing multivariate Normality 27](#_Toc479924159)

[5.4.2 Below plot showing Residuals vs Fitted values 28](#_Toc479924160)

[5.4.3 Below plot showing sqrt of Standardized Residuals vs Fitted values 29](#_Toc479924161)

[5.4.4 Below plot showing Homoscedacity 30](#_Toc479924162)

[5.5 Time series lags results 31](#_Toc479924163)

[6. Analysis 34](#_Toc479924164)

[6.1 Decision point 34](#_Toc479924165)

[6.2 Scope of enhancements 34](#_Toc479924166)

[7. Appendices 34](#_Toc479924167)

# Purpose

## 1.1 Problem Description

This project focuses on the role of Foreign Direct investment in SaudiArabia’s economic development and government policies towards FDI. Attracting inward investment is a key component in its economic development initiatives, for the potential knowledge transfer and cross fertilization to domestic operators which it can provide. The project has the following objectives: To study the trends and pattern of flow of FDI and to assess the determinants of FDI inflows by studying relationships among the econometrics variables such as GDP, exports, and Political stability index, etc.



## Project scope

This project is to study closely the factors which are positively affecting the net inflow of Foreign Direct Investments (FDIs) into SaudiArabia’s economy and each factor will be taking how many years to boost the FDI inflows. To predict the type of factors which are affecting the FDIs, Random Forest & Multiple linear regression analysis have been used and to predict the number of years for each factor to make a positive impact on attracting the FDIs, Timeseries analysis technique has been used.

Section 5 deals with some empirical aspects related to causality flows between FDI and related macroeconomic variables and to the determinants of FDI. In this context the top 30 variables which attracting FDI are investigated. A final section of the project then concludes the study.

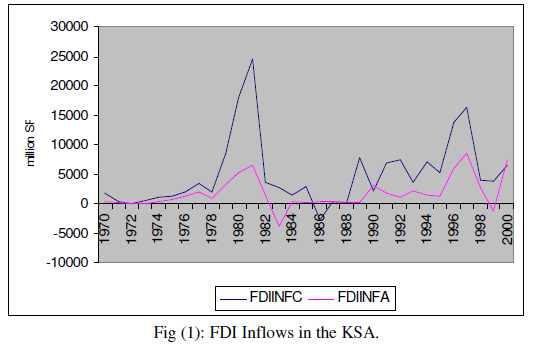
## Constraints

We must get the most significant attributes which are implementable and actionable in nature after applying the model on the data

# State-of-the-art

## Introduction to FDIs

Foreign Direct Investment (FDI) inflow in a country is considered to be an important waterway for the transmission of new ideas, technologies and business skills among different nations. It can have better impact on the developmental efforts of a country in transition. The prospects for growth can get better with an increase in the total level of capital investment in the economy and by introducing more productive technology and techniques. An essential source of new physical capital is being provided by the foreign investors helping in channelizing the domestic resources into investment. Some countries face shortages of foreign exchange to import foreign products and equipments necessary to put an impetus to the development process. This shortage also slogs the ability of a nation to promote trade links with other nations in the process of globalization. Foreign companies not only make investment in the host country but also make limited trade with parent country, making an increase in the host countries trade, helping it out to integrate with the world economy. The use of FDI as a tool for entering a global market is often considered to be most essential. An intensive growth of the technology in the specific industry also motivates FDI. In addition, many FDIs are motivated by defensive tactics such as responding to a saturated home market or reacting to problems in the homemarket.  
Since the first oil shock in 1973, FDI has been entered into Saudi economy as a minor capital source. Most foreign participation was seen in building productive facilities and providing services under contract, rather than through FDI. In January 1999, Saudi Arabian Government created the Supreme Petroleum and Mineral Affairs Council, separate entity from the Saudi Arabian Oil Company, (ARAMCO), to direct post-production energy policy. In 2000, the Government established a new investment regulatory body, the General Investment Authority. Subject to all appropriate documents being lodged, it processes applications within 30 days. Until April 2000, foreign equity in the industrial sector was limited to 49 per cent. While in principle full foreign ownership was allowed in other sectors (except oil).Wholly foreign owned firms could not bid for government contracts or access cheap credit or tax concessions Foreign companies can now access tax holidays and concessional finance in other sectors, making full foreign ownership more feasible.



As can be seen from the data and figures, contracted FDI inflows (FDIINFC)

had more pronounced fluctuations as compared to actual FDI (FDIINFA). A high

peak of contracted FDI is detected in 1981 preceding the massive investments in the

petrochemical industry and the establishment of the giant conglomerate Saudi Arabia

Basic Industries Company (SABIC) that occurred in the early eighties. Actual inflows

normally fell short of the contracted FDIs and were generally more stable since they

were spread over a number of years. A slight upward trend in FDI inflows could also

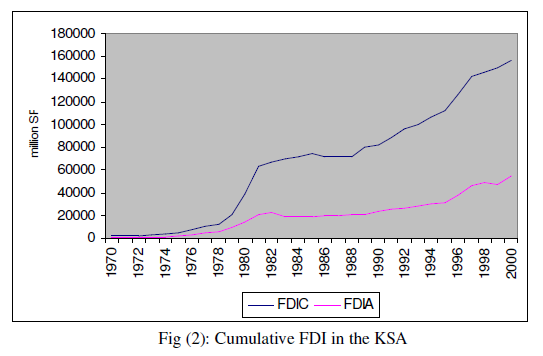
be observed to be occurring since 1984 possibly due to the liberalization and openingup

efforts of the country. Table (A2) in the appendix provides data on cumulative

contracted and realized FDIs whereas Fig (2) below illustrates graphically the

performance of both contracted (FDIC) and realized (FDIA) cumulative FDI through

the sample period.



Positive trends are clearly seen in both series with the common jump that have

occurred in the early eighties because of the infusion of massive FDI into the

petrochemicals sub-sector and the gradual increase throughout the nineties with some

acceleration towards the end of the period.

Here, the most commonly used techniques for solving the given problem are listed. Near all of them are heuristics and meta heuristics because no exact algorithm can be guaranteed to find significant attributes which affects the FDIs in any country

## Trending Solutions

**Random Forests :** To extract the top features from the data which are favoring the FDI inflows  
Random forests or random decision forests are an [ensemble learning](https://en.wikipedia.org/wiki/Ensemble_learning) method for [classification](https://en.wikipedia.org/wiki/Statistical_classification), [regression](https://en.wikipedia.org/wiki/Regression_analysis) and other tasks, that operate by constructing a multitude of [decision trees](https://en.wikipedia.org/wiki/Decision_tree_learning) at training time and outputting the class that is the [mode](https://en.wikipedia.org/wiki/Mode_(statistics)) of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of [overfitting](https://en.wikipedia.org/wiki/Overfitting" \o "Overfitting) to their training set

The first algorithm for random decision forests was created by Tin Kam Ho using the [random subspace method](https://en.wikipedia.org/wiki/Random_subspace_method), which, in Ho's formulation, is a way to implement the "stochastic discrimination" approach to classification proposed by Eugene Kleinberg

The general method of random decision forests was first proposed by Ho in 1995, who established that forests of trees splitting with [oblique hyperplanes](https://en.wikipedia.org/w/index.php?title=Oblique_hyperplanes&action=edit&redlink=1), if randomly restricted to be sensitive to only selected feature dimensions, can gain accuracy as they grow without suffering from overtraining. A subsequent work along the same lines concluded that other splitting methods, as long as they are randomly forced to be insensitive to some feature dimensions, behave similarly. Note that this observation of a more complex classifier (a larger forest) getting more accurate nearly monotonically is in sharp contrast to the common belief that the complexity of a classifier can only grow to a certain level of accuracy before being hurt by overfitting.

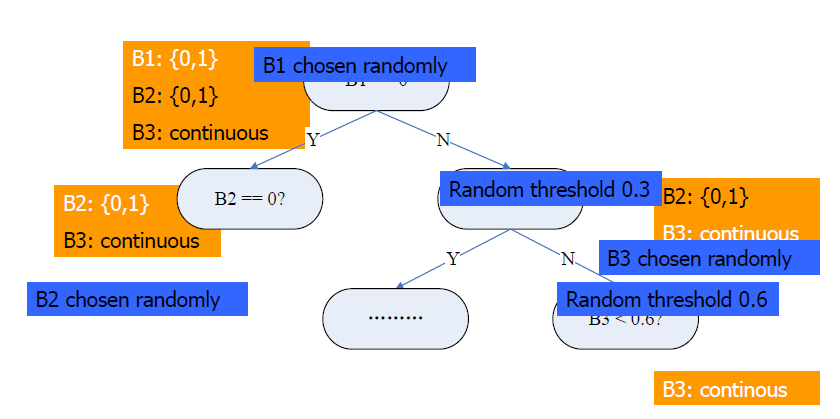
### Variable importance in Random Forests:

Random forests can be used to rank the importance of variables in a regression or classification problem in a natural way. The following technique was described in Breiman's original paper and is implemented in the [R](https://en.wikipedia.org/wiki/R_(programming_language)) package randomForest.

The first step in measuring the variable importance in a data set is to fit a random forest to the data. During the fitting process the [out-of-bag error](https://en.wikipedia.org/wiki/Out-of-bag_error) for each data point is recorded and averaged over the forest. To measure the importance of the feature j-th after training, the values of the j-th feature are permuted among the training data and the out-of-bag error is again computed on this perturbed data set. The importance score for the j-th feature is computed by averaging the difference in out-of-bag error before and after the permutation over all trees. The score is normalized by the standard deviation of these differences.

Features which produce large values for this score are ranked as more important than features which produce small values.

This method of determining variable importance has some drawbacks. For data including categorical variables with different number of levels, random forests are biased in favor of those attributes with more levels. Methods such as [partial permutations](https://en.wikipedia.org/wiki/Partial_permutation) and growing unbiased trees can be used to solve the problem. If the data contain groups of correlated features of similar relevance for the output, then smaller groups are favored over larger groups.

 **RANDOM FOREST FLOW CHART**

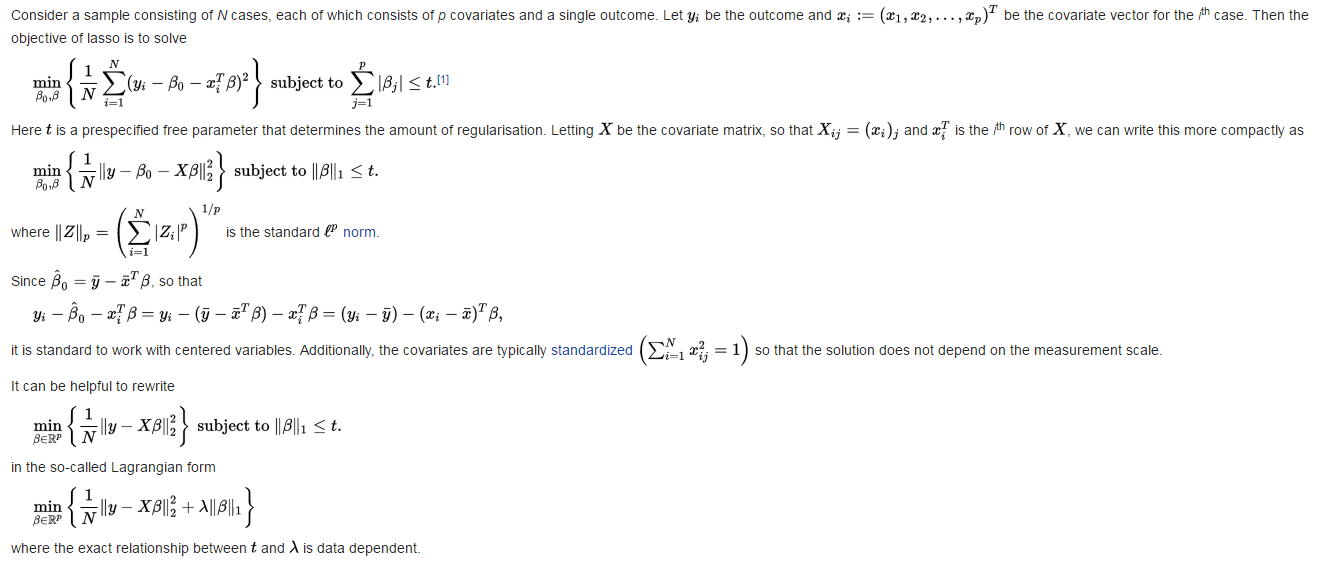
**Lasso Regression**: Another algorithm to extract and to perform regression on the top features from the data which are favoring the FDI inflows

In [statistics](https://en.wikipedia.org/wiki/Statistics) and [machine learning](https://en.wikipedia.org/wiki/Machine_learning), lasso (least absolute shrinkage and selection operator) (also Lasso or LASSO) is a [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis) method that performs both [variable selection](https://en.wikipedia.org/wiki/Variable_selection) and [regularization](https://en.wikipedia.org/wiki/Regularization_(mathematics)) in order to enhance the prediction accuracy and interpretability of the [statistical model](https://en.wikipedia.org/wiki/Statistical_model) it produces. It was introduced by [Robert Tibshirani](https://en.wikipedia.org/wiki/Robert_Tibshirani) in 1996 based on [Leo Breiman’s](https://en.wikipedia.org/wiki/Leo_Breiman) Nonnegative Garrote. Lasso was originally formulated for [least squares](https://en.wikipedia.org/wiki/Least_squares) models and this simple case reveals a substantial amount about the behavior of the estimator, including its relationship to [ridge regression](https://en.wikipedia.org/wiki/Ridge_regression) and best subset selection and the connections between lasso coefficient estimates and so-called soft thresholding. It also reveals that (like standard [linear regression](https://en.wikipedia.org/wiki/Linear_regression)) the coefficient estimates need not be unique if [covariates](https://en.wikipedia.org/wiki/Covariate) are [collinear](https://en.wikipedia.org/wiki/Collinear).

Prior to lasso, the most widely used method for choosing which covariates to include was [stepwise selection](https://en.wikipedia.org/wiki/Stepwise_regression), which only improves prediction accuracy in certain cases, such as when only a few covariates have a strong relationship with the outcome. However, in other cases, it can make prediction error worse. Also, at the time, ridge regression was the most popular technique for improving prediction accuracy. Ridge regression improves prediction error by [shrinking](https://en.wikipedia.org/wiki/Shrinkage_(statistics)) large [regression coefficients](https://en.wikipedia.org/wiki/Regression_coefficients) in order to reduce [overfitting](https://en.wikipedia.org/wiki/Overfitting" \o "Overfitting), but it does not perform covariate selection and therefore does not help to make the model more interpretable.

Lasso is able to achieve both of these goals by forcing the sum of the absolute value of the regression coefficients to be less than a fixed value, which forces certain coefficients to be set to zero, effectively choosing a simpler model that does not include those coefficients. This idea is similar to ridge regression, in which the sum of the squares of the coefficients is forced to be less than a fixed value, though in the case of ridge regression, this only shrinks the size of the coefficients, it does not set any of them to zero. However, data is expensive; models performing the complete selection of coefficients (setting then to zero) can ignore relevant information from datasets. Probabilistic approaches that penalize, and select regression coefficient based on the information from the data combined with prior knowledge (Multilevel Bayesian LASSO) may outperform the classical LASSO model, given that all information are used considering the uncertainty of the data, instead of simply setting them directly to zero.

Lasso was originally introduced in the context of least squares, and it can be instructive to consider this case first, since it illustrates many of lasso’s properties in a straightforward setting.



**Mulitple Linear Regression:** To predict the type of factors which are affecting the FDIs, Multiple linear regression analysis has been used

Statistics are used in medicine for data description and inference. Inferential statistics are used to answer questions about the data, to test hypotheses (formulating the alternative or null hypotheses), to generate a measure of effect, typically a ratio of rates or risks, to describe associations (correlations) or to model relationships (regression) within the data and, in many other functions. Usually point estimates are the measures of associations or of the magnitude of effects. Confounding, measurement errors, selection bias and random errors make unlikely the point estimates to equal the true ones. In the estimation process, the random error is not avoidable. One way to account for is to compute p-values for a range of possible parameter values (including the null). The range of values, for which the *p-value* exceeds a specified alpha level (typically 0.05) is called confidence interval. An interval estimation procedure will, in 95% of repetitions (identical studies in all respects except for random error), produce limits that contain the true parameters. It is argued that the question if the pair of limits produced from a study contains the true parameter could not be answered by the ordinary (frequentist) theory of confidence intervals[1](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/#R1). Frequentist approaches derive estimates by using probabilities of data (either p-values or likelihoods) as measures of compatibility between data and hypotheses, or as measures of the relative support that data provide hypotheses. Another approach, the Bayesian, uses data to improve existing (prior) estimates in light of new data. Proper use of any approach requires careful interpretation of statistics

The goal in any data analysis is to extract from raw information the accurate estimation. One of the most important and common question concerning if there is statistical relationship between a response variable (Y) and explanatory variables (Xi). An option to answer this question is to employ regression analysis in order to *model* its relationship. There are various types of regression analysis. The type of the regression model depends on the type of the distribution of Y; if it is continuous and approximately normal we use linear regression model; if dichotomous we use logistic regression; if Poisson or multinomial we use log-linear analysis; if time-to-event data in the presence of censored cases (survival-type) we use Cox regression as a method for modeling. By modeling we try to predict the outcome (Y) based on values of a set of predictor variables (Xi). These methods allow us to assess the impact of multiple variables (covariates and factors) in the same model

 Linear regression is the procedure that estimates the coefficients of the linear equation, involving one or more independent variables that best predict the value of the dependent variable which should be quantitative.

**Linear Regression Equation**

The purpose of regression is to predict Y on the basis of X or to describe how Y depends on X (regression line or curve)

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e001.jpg

The Xi (X1, X2, , Xk) is defined as "predictor", "explanatory" or "independent" variable, while Y is defined as "dependent", "response" or "outcome" variable.

Assuming a linear relation in population, mean of Y for given X equals α+βX i.e. the "population regression line".

If Y = a + bX is the estimated line, then the fitted

Ŷi = a + bXi is called the fitted (or predicted) value, and Yi Ŷi is called the residual.

The estimated regression line is determined in such way that (residuals)2 to be the minimal i.e. the standard deviation of the residuals to be minimized (residuals are on average zero). This is called the "least squares" method. In the equation

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e002.jpg

b is the slope (the average increase of outcome per unit increase of predictor)

a is the intercept (often has no direct practical meaning)

A more detailed (higher precision of the estimates a and b) regression equation line can also be written as

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e003.jpg

Further inference about regression line could be made by the estimation of confidence interval (95%CI for the slope b). The calculation is based on the standard error of b:

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e004.jpg

so, 95% CI for β is b ± t0.975\*se(b) [t-distr. with df = n-2]

and the test for H0: β=0, is t = b / se(b) [p-value derived from t-distr. with df = n-2].

If the p value lies above 0.05 then the null hypothesis is not rejected which means that a straight line model in X does not help predicting Y. There is the possibility that the straight line model holds (slope = 0) or there is a curved relation with zero linear component. On the other hand, if the null hypothesis is rejected either the straight line model holds or in a curved relationship the straight line model helps, but is not the best model. Of course there is the possibility for a type II or type I error in the first and second option, respectively. The standard deviation of residual (σres) is estimated by

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e005.jpg

The standard deviation of residual (σres) characterizes the variability around the regression line i.e. the smaller the σres, the better the fit. It has a number of degrees of freedom. This is the number to divide by in order to have an unbiased estimate of the variance. In this case df = n-2, because two parameters, α and β, are estimated[7](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3049417/#R7).

Multiple Linear Regression Equation

In the multiple linear regression model, Y has normal distribution with mean

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e006.jpg

The model parameters β0 + β1 + +βρ and σ must be estimated from data.

β0 = intercept

β1 βρ = regression coefficients

σ = σres = residual standard deviation

Interpretation of Regresiion coefficients

In the equation **Y = β0** **+ β11** **+ +βρXρ**

β1 equals the mean increase in Y per unit increase in Xi , while other Xi's are kept fixed. In other words βi is influence of Xi corrected (adjusted) for the other X's. The estimation method follows the least squares criterion.

If b0, b1, , bρ are the estimates of β0, β1, , βρ then the "fitted" value of Y is

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e007.jpg

The b0, b1, ... , b are computed such that An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-24-e008.jpg to be minimal. Since Y – Yfit is called the residual; one can also say that the sum of squared residuals is minimized.

Test on overall or reduced model:

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-25-e005.jpg

In our example Tpers = β0 + β1 time outdoors + β2 Thome +β3 wind speed + residual

The null hypothesis (H0) is that there is no regression overall i.e. β1= β2=+βρ = 0

The test is based on the proportion of the SS explained by the regression relative to the residual SS. The test statistic (F= MSreg / MSres) has F-distribution with df1 = p and df2 = n p 1 (F- distribution table). In our example F= 5.49 (P<0.01)

If now we want to test the hypothesis Ho: β1= β2= β5 = 0 (k = 3)

In general k of p regression coefficients are set to zero under H0. The model that is valid if H0=0 is true is called the "reduced model". The Idea is to compare the explained variability of the model at hand with that of the reduced model.

The test statistic (F):

An external file that holds a picture, illustration, etc.
Object name is hippokratia-14-25-e006.jpg

follows a F-distribution with df1 = k and df2 = n p 1.

If one or two variables are left out and we calculate SS reg (the statistical package does) and we find that the test statistic for F lies between 0.05 < P < 0.10, that means that there is some evidence, although not strong, that these variables together, independently of the others, contribute to the prediction of the outcome.

**Assumptions:**

If a linear model is used, the following assumptions should be met. For each value of the independent variable, the distribution of the dependent variable must be normal. The variance of the distribution of the dependent variable should be constant for all values of the independent variable. The relationship between the dependent variable and the independent variables should be linear, and all observations should be independent. So the assumptions are: independence; linearity; normality; homoscedasticity. In other words the residuals of a good model should be normally and randomly distributed i.e. the unknown does not depend on X ("homoscedasticity")

**Checking for violations of model assumptions:**

To check model assumptions we used residual analysis. There are several kinds of residuals most commonly used are the standardized residuals (ZRESID) and the studentized residuals (SRESID) [6]. If the model is correct, the residuals should have a normal distribution with mean zero and constant sd (i.e. not depending on X). In order to check this we can plot residuals against X. If the variation alters with increasing X, then there is violation of homoscedasticity. We can also use the Durbin-Watson test for serial correlation of the residuals and casewise diagnostics for the cases meeting the selection criterion (outliers above n standard deviations). The residuals are (zero mean) independent, normally distributed with constant standard deviation (homogeneity of variances)

To discover deviations form linearity and homogeneity of variables we can plot residuals against each predictor or against predicted values. Alternatively by using the PARTIAL plot we can assess linearity of a predictor variable. The partial plot for a predictor X1 is a plot of residuals of Y regressed on other Xs and against residuals of Xi regressed on other X's. The plot should be linear. To check the normality of residuals we can use an histogram (with normal curve) or a normal probability plot

The goodness-of-fit of the model is assessed by studying the behavior of the residuals, looking for "special observations / individuals" like outliers, observations with high "leverage" and influential points. Observations deserving extra attention are outliers i.e. observations with unusually large residual; high leverage points: unusual x - pattern, i.e. outliers in predictor space; influential points: individuals with high influence on estimate or standard error of one or more β's. An observation could be all three. It is recommended to inspect individuals with large residual, for outliers; to use distances for high leverage points i.e. measures to identify cases with unusual combinations of values for the independent variables and cases that may have a large impact on the regression model. For influential points use influence statistics i.e. the change in the regression coefficients (DfBeta(s)) and predicted values (DfFit) that results from the exclusion of a particular case. Overall measure for influence on all β's jointly is "Cook's distance" (COOK). Analogously for standard errors overall measure is COVRATIO

**Deviations from model assumptions:**

We can use some tips to correct some deviation from model assumptions. In case of curvilinearity in one or more plots we could add quadratic term(s). In case of non homogeneity of residual sd, we can try some transformation: log Y if Sres is proportional to predicted Y; square root of Y if Y distribution is Poisson-like; 1/Y if Sres2 is proportional to predicted Y; Y2 if Sres2 decreases with Y. If linearity and homogeneity hold then non-normality does not matter if the sample size is big enough (n≥50- 100). If linearity but not homogeneity hold then estimates of β's are correct, but not the standard errors. They can be corrected by computing the "robust" se's (sandwich, Huber's estimate)

**Stepwise Regression (StepAIC):**

In [statistics](https://en.wikipedia.org/wiki/Statistics), stepwise regression is a method of fitting [regression models](https://en.wikipedia.org/wiki/Regression_model) in which the choice of predictive variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of [explanatory variables](https://en.wikipedia.org/wiki/Explanatory_variable) based on some prespecified criterion. Usually, this takes the form of a sequence of [F-tests](https://en.wikipedia.org/wiki/F-test) or [t-tests](https://en.wikipedia.org/wiki/T-test), but other techniques are possible, such as [adjusted R2](https://en.wikipedia.org/wiki/Adjusted_R-squared), [Akaike information criterion](https://en.wikipedia.org/wiki/Akaike_information_criterion" \o "Akaike information criterion), [Bayesian information criterion](https://en.wikipedia.org/wiki/Bayesian_information_criterion), [Mallows's Cp](https://en.wikipedia.org/wiki/Mallows%27s_Cp" \o "Mallows's Cp), [PRESS](https://en.wikipedia.org/wiki/PRESS_statistic), or [false discovery rate](https://en.wikipedia.org/wiki/False_discovery_rate).

The frequent practice of fitting the final selected model followed by reporting estimates and confidence intervals without adjusting them to take the model building process into account has led to calls to stop using stepwise model building altogether or to at least make sure model uncertainty is correctly reflected.

The main approaches are:

* **Forward selection**, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit, and repeating this process until none improves the model to a statistically significant extent.
* **Backward elimination**, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.
* **Bidirectional elimination**, a combination of the above, testing at each step for variables to be included or excluded.

**Selection Criterion:**

 This is an automatic procedure for statistical model selection in cases where there is a large number of potential explanatory variables, and no underlying theory on which to base the model selection. The procedure is used primarily in [regression analysis](https://en.wikipedia.org/wiki/Regression_analysis), though the basic approach is applicable in many forms of model selection. This is a variation on forward selection. At each stage in the process, after a new variable is added, a test is made to check if some variables can be deleted without appreciably increasing the [residual sum of squares](https://en.wikipedia.org/wiki/Residual_sum_of_squares) (RSS). The procedure terminates when the measure is (locally) maximized, or when the available improvement falls below some critical value.

One of the main issues with stepwise regression is that it searches a large space of possible models. Hence it is prone to [overfitting](https://en.wikipedia.org/wiki/Overfitting" \o "Overfitting) the data. In other words, stepwise regression will often fit much better in sample than it does on new out-of-sample data. This problem can be mitigated if the criterion for adding (or deleting) a variable is stiff enough. The key line in the sand is at what can be thought of as the [Bonferroni](https://en.wikipedia.org/wiki/Bonferroni" \o "Bonferroni) point: namely how significant the best spurious variable should be based on chance alone. On a t-statistic scale, this occurs at about (sqrt 2log p), where p is the number of predictors. Unfortunately, this means that many variables which actually carry signal will not be included. This fence turns out to be the right trade-off between over-fitting and missing signal. If we look at the [risk](https://en.wikipedia.org/wiki/Risk_function) of different cutoffs, then using this bound will be within a 2logp factor of the best possible risk. Any other cutoff will end up having a larger such [risk inflation](https://en.wikipedia.org/w/index.php?title=Risk_inflation&action=edit&redlink=1)

**Model Accuracy:**

A way to test for errors in models created by step-wise regression, is to not rely on the model's F-statistic, significance, or multiple R, but instead assess the model against a set of data that was not used to create the model. This is often done by building a model based on a sample of the dataset available (e.g., 70%) and use the remaining 30% dataset to assess the accuracy of the model. Accuracy is then often measured as the actual standard error (SE), MAPE, or mean error between the predicted value and the actual value in the hold-out sample. This method is particularly valuable when data are collected in different settings (e.g., different times, social vs. solitary situations) or when models are assumed to be generalizable.

### Time Series Regression: Lagged Variables

Many econometric models are *dynamic*, using lagged variables to incorporate feedback over time. By contrast, *static* time series models represent systems that respond exclusively to current events.

Lagged variables come in several types:

$\bullet$ **Distributed Lag (DL**) variables are lagged values $x_{t-k}$ of observed exogenous predictor variables $x_t$.

$\bullet$ **Autoregressive (AR)** variables are lagged values $y_{t-k}$ of observed endogenous response variables $y_t$.

$\bullet$ **Moving Average (MA)** variables are lagged values $e_{t-k}$ of unobserved stochastic innovations processes .

Dynamic models are often constructed using linear combinations of different types of lagged variables, to create ARMA, ARDL, and other hybrids. The modeling goal, in each case, is to reflect important interactions among relevant economic factors, accurately and concisely.

Dynamic model specifications ask the question: Which lags are important? Some models, such as seasonal models, use lags at distinct periods in the data. Other models base their lag structure on theoretical considerations of how, and when, economic agents react to changing conditions. In general, lag structures identify the time delay of the response to known leading indicators.

However, lag structures must do more than just represent the available theory. Because dynamic specifications produce interactions among variables that can affect standard regression techniques, lag structures must also be designed with accurate model estimation in mind.

**Spectifications:**

Consider the multiple linear regression (MLR) model:

$$y_t = Z_t \beta + e_t,$$

where $y_t$ is an observed response, $Z_t$ includes columns for each potentially relevant predictor variable, including lagged variables, and $e_t$ is a stochastic innovations process. The accuracy of estimation of the coefficients in $\beta$ depends on the constituent columns of $Z_t$, as well as the joint distribution of $e_t$. Selecting predictors for $Z_t$ that are both statistically and economically significant usually involves cycles of estimation, residual analysis, and respecification.

Classical linear model (CLM) assumptions allow ordinary least squares (OLS) to produce estimates of $\beta$ with desirable properties: unbiased, consistent, and efficient relative to other estimators. Lagged predictors in $Z_t$, however, can introduce violations of CLM assumptions. Specific violations depend on the types of lagged variables in the model, but the presence of dynamic feedback mechanisms, in general, tends to exaggerate the problems associated with static specifications.

Model specification issues are usually discussed relative to a data-generating process (DGP) for the response variable $y_t$. Practically, however, the DGP is a theoretical construct, realized only in simulation. No model ever captures real-world dynamics entirely, and model coefficients in $\beta$ are always a subset of those in the true DGP. As a result, innovations in $e_t$ become a mix of the inherent stochasticity of the process and a potentially large number of omitted variables (OVs). Autocorrelations in $e_t$ are common in econometric models where OVs exhibit persistence over time. Rather than comparing a model to a theoretical DGP, it is more practical to evaluate whether, or to what degree, dynamics in the data have been distinguished from autocorrelations in the residuals.

Initially, lag structures may include observations of economic factors at multiple, proximate times. However, observations at time t are likely to be correlated with observations at times t - 1, t - 2, and so forth, through economic inertia. Thus, a lag structure may overspecify the dynamics of the response by including a sequence of lagged predictors with only marginal contributions to the DGP. The specification will overstate the effects of past history, and fail to impose relevant restrictions on the model. Extended lag structures also require extended presample data, reducing the sample size and decreasing the number of degrees of freedom in estimation procedures. Consequently, overspecified models may exhibit pronounced problems of collinearity and high estimator variance. The resulting estimates of $\beta$ have low precision, and it becomes difficult to separate individual effects.

To reduce predictor dependencies, lag structures can be restricted. If the restrictions are too severe, however, other problems of estimation arise. A restricted lag structure may underspecify the dynamics of the response by excluding predictors that are actually a significant part of the DGP. This leads to a model that underestimates the effects of past history, forcing significant predictors into the innovations process. If lagged predictors in $e_t$ are correlated with proximate lagged predictors in $Z_t$, the CLM assumption of strict exogeneity of the regressors is violated, and OLS estimates of $\beta$ become biased and inconsistent.

Specific issues are associated with different types of lagged predictors.

Lagged exogenous predictors $x_{t-k}$, by themselves, do not violate CLM assumptions. However, DL models are often described, at least initially, by a long sequence of potentially relevant lags, and so suffer from the problems of overspecification mentioned above. Common, if somewhat ad hoc, methods for imposing restrictions on the lag weights (that is, the coefficients in $\beta$) are discussed in the example on "Lag Order Selection." In principle, however, the analysis of a DL model parallels that of a static model. Estimation issues related to collinearity, influential observations, spurious regression, autocorrelated or heteroscedastic innovations, etc. must still be examined.

Lagged endogenous predictors $y_{t-k}$ are more problematic. AR models introduce violations of CLM assumptions that lead to biased OLS estimates of $\beta$. Absent any other CLM violations, the estimates are, nevertheless, consistent and relatively efficient. Consider a simple first-order autoregression of $y_t$ on $y_{t-1}$:

$$y_t = \beta y_{t-1} + e_t.$$

In this model, $y_t$ is determined by both $y_{t-1}$ and $e_t$. Shifting the equation backwards one step at a time, $y_{t-1}$ is determined by both $y_{t-2}$ and $e_{t-1}$, $y_{t-2}$ is determined by both $y_{t-3}$ and $e_{t-2}$, and so forth. Transitively, the predictor $y_{t-1}$ is correlated with the entire previous history of the innovations process. Just as with underspecification, the CLM assumption of strict exogeneity is violated, and OLS estimates of $\beta$ become biased. Because $\beta$ must absorb the effects of each $e_{t-k}$, the model residuals no longer represent true innovations [10].

The problem is compounded when the innovations in an AR model are autocorrelated. As discussed in the example on "Residual Diagnostics," autocorrelated innovations in the absence of other CLM violations produce unbiased, if potentially high variance, OLS estimates of model coefficients. The major complication, in that case, is that the usual estimator for the standard errors of the coefficients becomes biased. (The effects of heteroscedastic innovations are similar, though typically less pronounced.) If, however, autocorrelated innovations are combined with violations of strict exogeneity, like those produced by AR terms, estimates of $\beta$ become both biased and inconsistent.

If lagged innovations $e_{t-k}$ are used as predictors, the nature of the estimation process is fundamentally changed, since the innovations cannot be directly observed. Estimation requires that the MA terms be inverted to form infinite AR representations, and then restricted to produce a model that can be estimated in practice. Since restrictions must be imposed during estimation, numerical optimization techniques other than OLS, such as maximum likelihood estimation (MLE), are required. Models with MA terms are considered in the example on "Lag Order Selection."

### Vector Autoregressive models VAR(p) models VAR models (vector autoregressive models) are used for multivariate time series. The structure is that each variable is a linear function of past lags of itself and past lags of the other variables.

As an example suppose that we measure three different time series variables, denoted by xt,1xt,1, xt,2xt,2, and xt,3xt,3.

The vector autoregressive model of order 1, denoted as VAR(1), is as follows:

xt,1=α1+ϕ11xt−1,1+ϕ12xt−1,2+ϕ13xt−1,3+wt,1xt,1=α1+ϕ11xt−1,1+ϕ12xt−1,2+ϕ13xt−1,3+wt,1

xt,2=α2+ϕ21xt−1,1+ϕ22xt−1,2+ϕ23xt−1,3+wt,2xt,2=α2+ϕ21xt−1,1+ϕ22xt−1,2+ϕ23xt−1,3+wt,2

xt,3=α3+ϕ31xt−1,1+ϕ32xt−1,2+ϕ33xt−1,3+wt,3xt,3=α3+ϕ31xt−1,1+ϕ32xt−1,2+ϕ33xt−1,3+wt,3

Each variable is a linear function of the lag 1 values for all variables in the set.

In a VAR(2) model, the lag 2 values for all variables are added to the right sides of the equations, In the case of three x-variables (or time series) there would be six predictors on the right side of each equation, three lag 1 terms and three lag 2 terms.

In general, for a VAR(p) model, the first p lags of each variable in the system would be used as regression predictors for each variable.

VAR models are a specific case of more general VARMA models. VARMA models for multivariate time series include the VAR structure above along with moving average terms for each variable. More generally yet, these are special cases of ARMAX models that allow for the addition of other predictors that are outside the multivariate set of principal interest.

Here, as in Section 5.8 of the text, we’ll focus on VAR models.

On page 304, the authors fit the model of the form

xt=Γut+ϕxt−1+wtxt=Γut+ϕxt−1+wt

where ut=(1,t)′ut=(1,t)′ includes terms to simultaneously fit the constant and trend. It arose from macroeconomic data where large changes in the data permanently affect the level of the series.

# Method

## Possibilities

* Of the above algorithms, pick Random Forest as it has highest node purity levels with better performance than the other approaches for extracting the top features from the given data
* Lasso implementation on the given data is quite a challenging task to get the top features as its sparsity of the matrix and it shrinks the variables too much.
* Logistic regression ,Support Vector regression & Multiple linear regression models can be used to predict the significant variables
* Vector auto regression can also be used for time series analysis

In this study, I have considered Random Forest, Multiple linear regression for significant variables, stepAIC method for choosing refined variables and time series by lags

Multiple Linear regression terminologies in this given problem:

Response Variable or Y variable - "Foreign.direct.investment..net.inflows....of.GDP.",

Predictor variables or X variables – "Net.foreign.assets..current.LCU.",

"Net.primary.income..BoP..current.US..",

"Total.reserves..includes.gold..current.US..",

"Service.imports..BoP..current.US..",

"GDP.growth..annual...",

"Commercial.service.imports..current.US..",

"Real.effective.exchange.rate.index..2010...100.",

"Inflation..GDP.deflator..annual...",

"Permanent.cropland....of.land.area.",

"Reserves.and.related.items..BoP..current.US..",

"Energy.imports..net....of.energy.use.",

"Trade....of.GDP.",

"Primary.income.receipts..BoP..current.US..",

"Energy.use..kg.of.oil.equivalent.per.capita.",

"Net.income.from.abroad..current.US..",

"Primary.income.payments..BoP..current.US.." ,

"Domestic.credit.to.private.sector.by.banks....of.GDP.",

"Claims.on.central.government..etc.....GDP.",

"Goods.imports..BoP..current.US.." ,

"Adjusted.savings..natural.resources.depletion....of.GNI.",

"Railways..passengers.carried..million.passenger.km." ,

"Age.dependency.ratio..young....of.working.age.population." ,

"Food.exports....of.merchandise.exports." ,

"Imports.of.goods.and.services....of.GDP.",

"Population.growth..annual...",

"Final.consumption.expenditure..etc.....of.GDP.",

"Merchandise.trade....of.GDP.",

"Industry..value.added..annual...growth.",

"Import.value.index..2000...100.",

"Ores.and.metals.exports....of.merchandise.exports."

## Action points

* Process the data
* Load the required libraries
* Check the missing values if any
* Impute the missing values with imputeTS library
* If the columns does have less than 50% values, then delete those columns
* Identify the correlations b/w the variables if any
* Identify the response variable or Y variable & predictor variables or X variables
* Nullify the Year column by equating it to NULL value in the main data
* Apply the linear regression on the data
* Observe the Adjusted R2 values obtained from the model
* Apply step AIC method to on the extracted variables
* Apply regression model and find out significant variables
* Visualize the output and assumptions of regression model
* Apply time series by lagging the variables as many as lags we want
* Create a data frame which contains the maximum adjusted R2 values along with the other variables like Var, Lagvalue,& Multiple R2
* Finally, observe the maximum adjusted R2 values and its corresponding lags and decide the each variable takes how many years to give a positive impact on the FDIs inflows

## Assumptions

* Data taken from a reliable source ‘World bank’
* The Variable values collected are based on the surveys and the expert systems
* [Independence](https://en.wikipedia.org/wiki/Independence_(probability_theory)) of observations from each other (this assumption is an especially common error)
* Independence of observational error from potential [confounding](https://en.wikipedia.org/wiki/Confounding) effects.
* Assuming Foreign direct investment is an independent variable and the remaining variables as dependent in this study
* Same variables have mentioned in different units in the data, so considered all variables in a single unit by assuming that the values are interdependent
* Algorithms chosen in this study by referring different sources of literature

# Data

FDI data collected from World Bank by using the following procedure

Procedure CRAWLERTHREAD(frontier)  
 while not frontier.done() do  
 website🡨 frontier.nextSite()  
 url🡨 website.nextURL()  
 if website.permitsCrawl(url) then  
 text🡨 retrieveURL(url)  
 storeDocument(url,text)  
 for each url in parse(text) do  
 frontier .addURL(url)  
 end for  
 end if  
 Frontier.releaseSite(website)  
 end while   
 end procedure

## Data Processing

* From the data processing, it is identified that Same variables have mentioned in different units in the data
* So we can remove the data related to different units for a single variable by retaining the common unit of values for all variables to get faster and accurate results
* Data is having many NA values so, Imputed the missing values
* Deleted those columns which have less than 50% of the total no.of records

# Results

## Multiple Linear Regression values

Call:

lm(formula = Foreign.direct.investment..net.inflows....of.GDP. ~

., data = Top30)

Residuals:

Min 1Q Median 3Q Max

-1.3562 -0.5261 -0.1393 0.4719 1.6177

Coefficients:

Estimate Std. Error

(Intercept) 1.408e+01 1.647e+01

Adjusted.savings..natural.resources.depletion....of.GNI. 4.862e-02 1.124e-01

Age.dependency.ratio..young....of.working.age.population. 1.550e-02 1.118e-01

Claims.on.central.government..etc.....GDP. -8.112e-03 3.480e-02

Commercial.service.imports..current.US.. 3.600e-10 1.136e-10

Domestic.credit.to.private.sector.by.banks....of.GDP. -5.444e-03 1.419e-01

Energy.imports..net....of.energy.use. -4.096e-04 8.357e-04

Energy.use..kg.of.oil.equivalent.per.capita. -3.884e-04 1.010e-03

Final.consumption.expenditure..etc.....of.GDP. -9.501e-02 8.974e-02

Food.exports....of.merchandise.exports. 3.103e-01 1.591e+00

GDP.growth..annual... -1.648e-01 1.498e-01

Goods.imports..BoP..current.US.. -1.466e-10 1.253e-10

Import.value.index..2000...100. 5.232e-02 3.336e-02

Imports.of.goods.and.services....of.GDP. 6.464e-02 1.080e-01

Industry..value.added..annual...growth. 1.597e-02 6.978e-02

Inflation..GDP.deflator..annual... -7.975e-02 1.600e-02

Merchandise.trade....of.GDP. -8.697e-03 1.068e-01

Net.foreign.assets..current.LCU. -4.451e-12 8.869e-12

Net.income.from.abroad..current.US.. 7.247e-10 3.123e-10

Net.primary.income..BoP..current.US.. -1.769e-08 8.138e-09

Ores.and.metals.exports....of.merchandise.exports. -3.359e+00 2.958e+00

Permanent.cropland....of.land.area. -2.190e+01 3.881e+01

Population.growth..annual... 5.104e-01 7.594e-01

Primary.income.payments..BoP..current.US.. -1.663e-08 8.130e-09

Primary.income.receipts..BoP..current.US.. 1.756e-08 8.175e-09

Railways..passengers.carried..million.passenger.km. 4.844e-03 1.169e-02

Real.effective.exchange.rate.index..2010...100. -5.327e-02 2.678e-02

Reserves.and.related.items..BoP..current.US.. -5.218e-12 1.381e-11

Service.imports..BoP..current.US.. -1.707e-10 9.175e-11

Total.reserves..includes.gold..current.US.. -1.769e-11 2.307e-11

Trade....of.GDP. -8.433e-02 3.903e-02

t value Pr(>|t|)

(Intercept) 0.855 0.39992

Adjusted.savings..natural.resources.depletion....of.GNI. 0.433 0.66876

Age.dependency.ratio..young....of.working.age.population. 0.139 0.89079

Claims.on.central.government..etc.....GDP. -0.233 0.81745

Commercial.service.imports..current.US.. 3.169 0.00378 \*\*

Domestic.credit.to.private.sector.by.banks....of.GDP. -0.038 0.96967

Energy.imports..net....of.energy.use. -0.490 0.62803

Energy.use..kg.of.oil.equivalent.per.capita. -0.385 0.70358

Final.consumption.expenditure..etc.....of.GDP. -1.059 0.29910

Food.exports....of.merchandise.exports. 0.195 0.84682

GDP.growth..annual... -1.100 0.28097

Goods.imports..BoP..current.US.. -1.170 0.25223

Import.value.index..2000...100. 1.568 0.12845

Imports.of.goods.and.services....of.GDP. 0.598 0.55454

Industry..value.added..annual...growth. 0.229 0.82065

Inflation..GDP.deflator..annual... -4.984 3.18e-05 \*\*\*

Merchandise.trade....of.GDP. -0.081 0.93571

Net.foreign.assets..current.LCU. -0.502 0.61986

Net.income.from.abroad..current.US.. 2.320 0.02813 \*

Net.primary.income..BoP..current.US.. -2.174 0.03865 \*

Ores.and.metals.exports....of.merchandise.exports. -1.135 0.26620

Permanent.cropland....of.land.area. -0.564 0.57723

Population.growth..annual... 0.672 0.50727

Primary.income.payments..BoP..current.US.. -2.045 0.05068 .

Primary.income.receipts..BoP..current.US.. 2.148 0.04083 \*

Railways..passengers.carried..million.passenger.km. 0.415 0.68178

Real.effective.exchange.rate.index..2010...100. -1.989 0.05690 .

Reserves.and.related.items..BoP..current.US.. -0.378 0.70848

Service.imports..BoP..current.US.. -1.861 0.07374 .

Total.reserves..includes.gold..current.US.. -0.767 0.44988

Trade....of.GDP. -2.161 0.03974 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.01 on 27 degrees of freedom

Multiple R-squared: 0.9347, Adjusted R-squared: 0.8622

F-statistic: 12.89 on 30 and 27 DF, p-value: 1.181e-09

## stepAIC

Step: AIC=-1.37

Foreign.direct.investment..net.inflows....of.GDP. ~ Commercial.service.imports..current.US.. +

Energy.use..kg.of.oil.equivalent.per.capita. + Final.consumption.expenditure..etc.....of.GDP. +

GDP.growth..annual... + Goods.imports..BoP..current.US.. +

Import.value.index..2000...100. + Imports.of.goods.and.services....of.GDP. +

Inflation..GDP.deflator..annual... + Net.income.from.abroad..current.US.. +

Net.primary.income..BoP..current.US.. + Ores.and.metals.exports....of.merchandise.exports. +

Primary.income.payments..BoP..current.US.. + Primary.income.receipts..BoP..current.US.. +

Real.effective.exchange.rate.index..2010...100. + Service.imports..BoP..current.US.. +

Total.reserves..includes.gold..current.US.. + Trade....of.GDP.

Df Sum of Sq RSS

<none> 30.449

+ Reserves.and.related.items..BoP..current.US.. 1 0.925 29.525

+ Claims.on.central.government..etc.....GDP. 1 0.771 29.678

+ Energy.imports..net....of.energy.use. 1 0.698 29.751

+ Domestic.credit.to.private.sector.by.banks....of.GDP. 1 0.603 29.846

+ Population.growth..annual... 1 0.562 29.887

- Energy.use..kg.of.oil.equivalent.per.capita. 1 1.613 32.062

+ Food.exports....of.merchandise.exports. 1 0.510 29.939

+ Railways..passengers.carried..million.passenger.km. 1 0.264 30.185

+ Merchandise.trade....of.GDP. 1 0.132 30.317

+ Age.dependency.ratio..young....of.working.age.population. 1 0.127 30.322

+ Permanent.cropland....of.land.area. 1 0.121 30.328

- Imports.of.goods.and.services....of.GDP. 1 2.049 32.498

+ Net.foreign.assets..current.LCU. 1 0.064 30.386

+ Adjusted.savings..natural.resources.depletion....of.GNI. 1 0.062 30.387

+ Industry..value.added..annual...growth. 1 0.014 30.436

- Ores.and.metals.exports....of.merchandise.exports. 1 2.635 33.084

- Goods.imports..BoP..current.US.. 1 4.370 34.819

- Service.imports..BoP..current.US.. 1 4.862 35.311

- Import.value.index..2000...100. 1 5.001 35.450

- GDP.growth..annual... 1 6.400 36.849

- Primary.income.payments..BoP..current.US.. 1 6.433 36.882

- Final.consumption.expenditure..etc.....of.GDP. 1 6.885 37.334

- Primary.income.receipts..BoP..current.US.. 1 7.094 37.543

- Net.primary.income..BoP..current.US.. 1 7.272 37.721

- Trade....of.GDP. 1 8.120 38.569

- Total.reserves..includes.gold..current.US.. 1 10.583 41.032

- Commercial.service.imports..current.US.. 1 13.983 44.432

- Real.effective.exchange.rate.index..2010...100. 1 14.502 44.951

- Net.income.from.abroad..current.US.. 1 16.241 46.690

- Inflation..GDP.deflator..annual... 1 51.958 82.407 AIC

<none> -1.374

+ Reserves.and.related.items..BoP..current.US.. -1.163

+ Claims.on.central.government..etc.....GDP. -0.863

+ Energy.imports..net....of.energy.use. -0.719

+ Domestic.credit.to.private.sector.by.banks....of.GDP. -0.535

+ Population.growth..annual... -0.455

- Energy.use..kg.of.oil.equivalent.per.capita. -0.381

+ Food.exports....of.merchandise.exports. -0.354

+ Railways..passengers.carried..million.passenger.km. 0.120

+ Merchandise.trade....of.GDP. 0.373

+ Age.dependency.ratio..young....of.working.age.population. 0.383

+ Permanent.cropland....of.land.area. 0.395

- Imports.of.goods.and.services....of.GDP. 0.403

+ Net.foreign.assets..current.LCU. 0.504

+ Adjusted.savings..natural.resources.depletion....of.GNI. 0.507

+ Industry..value.added..annual...growth. 0.600

- Ores.and.metals.exports....of.merchandise.exports. 1.439

- Goods.imports..BoP..current.US.. 4.404

- Service.imports..BoP..current.US.. 5.218

- Import.value.index..2000...100. 5.446

- GDP.growth..annual... 7.690

- Primary.income.payments..BoP..current.US.. 7.742

- Final.consumption.expenditure..etc.....of.GDP. 8.449

- Primary.income.receipts..BoP..current.US.. 8.773

- Net.primary.income..BoP..current.US.. 9.046

- Trade....of.GDP. 10.337

- Total.reserves..includes.gold..current.US.. 13.927

- Commercial.service.imports..current.US.. 18.544

- Real.effective.exchange.rate.index..2010...100. 19.217

- Net.income.from.abroad..current.US.. 21.419

- Inflation..GDP.deflator..annual... 54.371

Call:

lm(formula = Foreign.direct.investment..net.inflows....of.GDP. ~

Commercial.service.imports..current.US.. + Energy.use..kg.of.oil.equivalent.per.capita. +

Final.consumption.expenditure..etc.....of.GDP. + GDP.growth..annual... +

Goods.imports..BoP..current.US.. + Import.value.index..2000...100. +

Imports.of.goods.and.services....of.GDP. + Inflation..GDP.deflator..annual... +

Net.income.from.abroad..current.US.. + Net.primary.income..BoP..current.US.. +

Ores.and.metals.exports....of.merchandise.exports. +

Primary.income.payments..BoP..current.US.. + Primary.income.receipts..BoP..current.US.. +

Real.effective.exchange.rate.index..2010...100. + Service.imports..BoP..current.US.. +

Total.reserves..includes.gold..current.US.. + Trade....of.GDP.,

data = Top30)

Coefficients:

(Intercept)

1.741e+01

Commercial.service.imports..current.US..

3.088e-10

Energy.use..kg.of.oil.equivalent.per.capita.

-7.875e-04

Final.consumption.expenditure..etc.....of.GDP.

-1.420e-01

GDP.growth..annual...

-1.065e-01

Goods.imports..BoP..current.US..

-1.185e-10

Import.value.index..2000...100.

3.367e-02

Imports.of.goods.and.services....of.GDP.

9.869e-02

Inflation..GDP.deflator..annual...

-8.079e-02

Net.income.from.abroad..current.US..

7.320e-10

Net.primary.income..BoP..current.US..

-1.764e-08

Ores.and.metals.exports....of.merchandise.exports.

-1.616e+00

Primary.income.payments..BoP..current.US..

-1.661e-08

Primary.income.receipts..BoP..current.US..

1.744e-08

Real.effective.exchange.rate.index..2010...100.

-3.462e-02

Service.imports..BoP..current.US..

-1.286e-10

Total.reserves..includes.gold..current.US..

-2.575e-11

Trade....of.GDP.

-7.753e-02

Call:

lm(formula = Foreign.direct.investment..net.inflows....of.GDP. ~

Commercial.service.imports..current.US.. + Imports.of.goods.and.services....of.GDP. +

Net.income.from.abroad..current.US.. + Primary.income.receipts..BoP..current.US.. +

Import.value.index..2000...100., data = Main\_data)

Residuals:

Min 1Q Median 3Q Max

-5.9770 -1.4256 0.0018 0.8878 4.9353

Coefficients:

Estimate Std. Error t value

(Intercept) -3.768e+00 2.062e+00 -1.827

Commercial.service.imports..current.US.. 1.591e-10 6.956e-11 2.288

Imports.of.goods.and.services....of.GDP. 1.565e-01 7.456e-02 2.099

Net.income.from.abroad..current.US.. -2.306e-10 1.651e-10 -1.397

Primary.income.receipts..BoP..current.US.. 3.539e-11 1.937e-10 0.183

Import.value.index..2000...100. -1.039e-02 5.931e-03 -1.751

Pr(>|t|)

(Intercept) 0.0734 .

Commercial.service.imports..current.US.. 0.0262 \*

Imports.of.goods.and.services....of.GDP. 0.0407 \*

Net.income.from.abroad..current.US.. 0.1683

Primary.income.receipts..BoP..current.US.. 0.8557

Import.value.index..2000...100. 0.0858 .

---

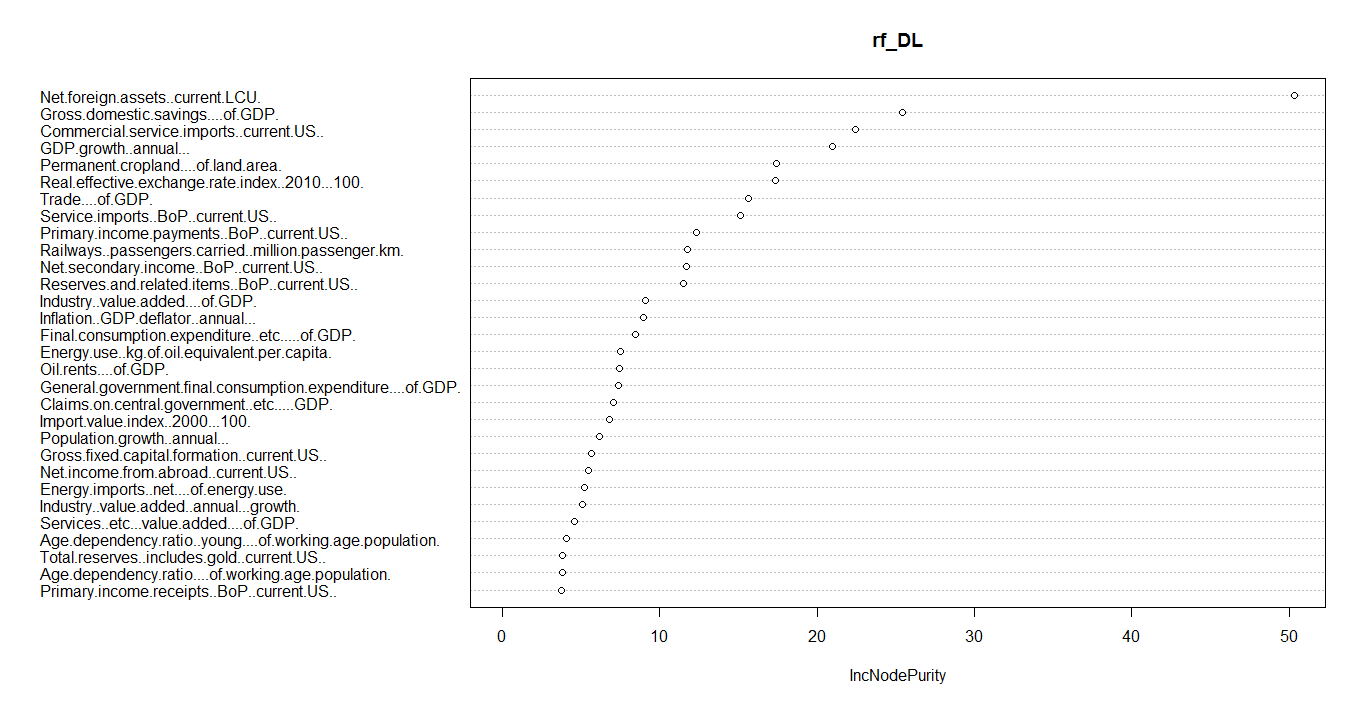
Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.168 on 52 degrees of freedom

Multiple R-squared: 0.4202, Adjusted R-squared: 0.3644

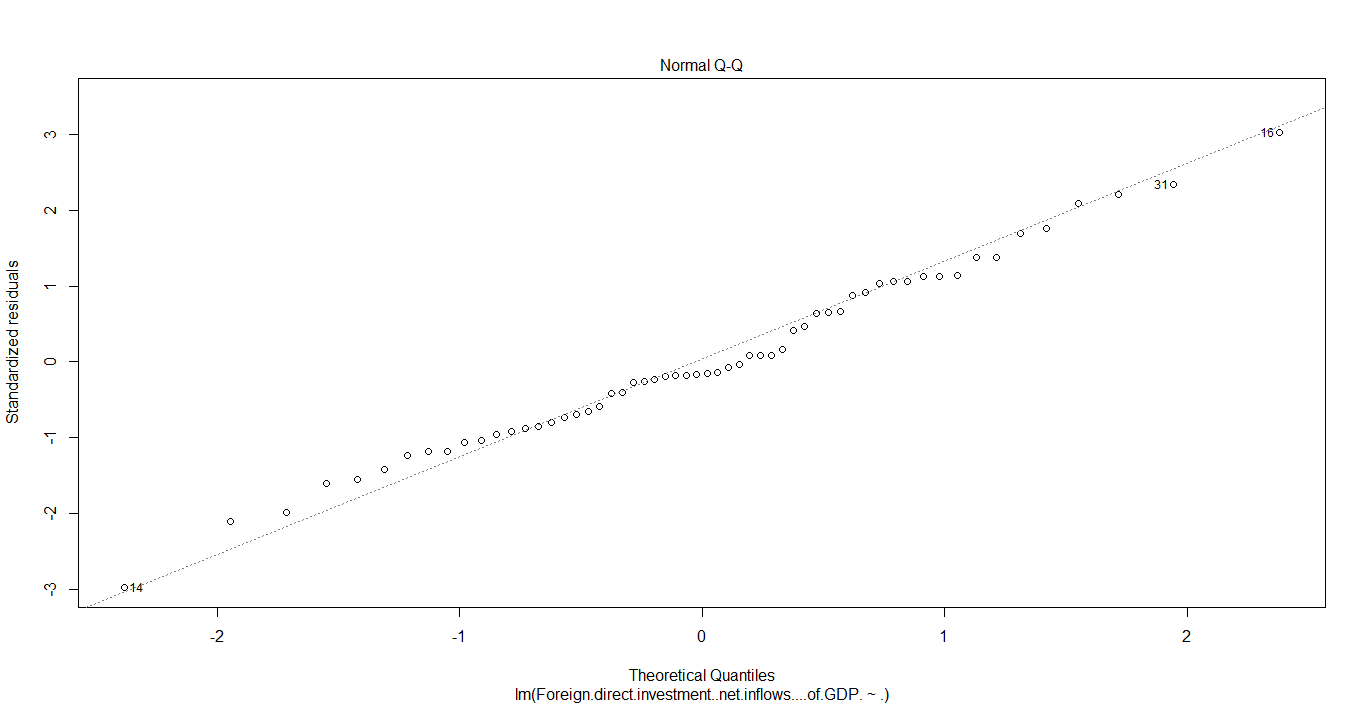
F-statistic: 7.536 on 5 and 52 DF, p-value: 2.203e-05

## Below plot showing the top 30 variables obtained from RF

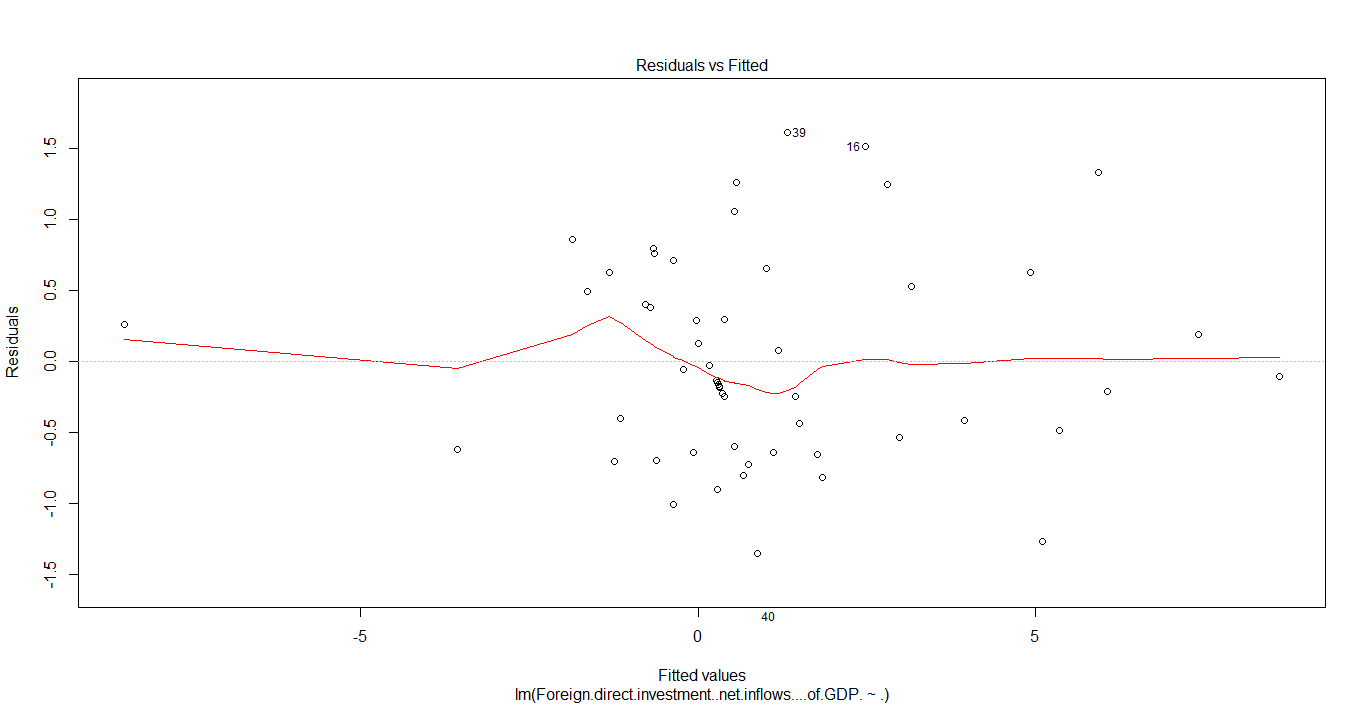


## Linear Regression assumption plots

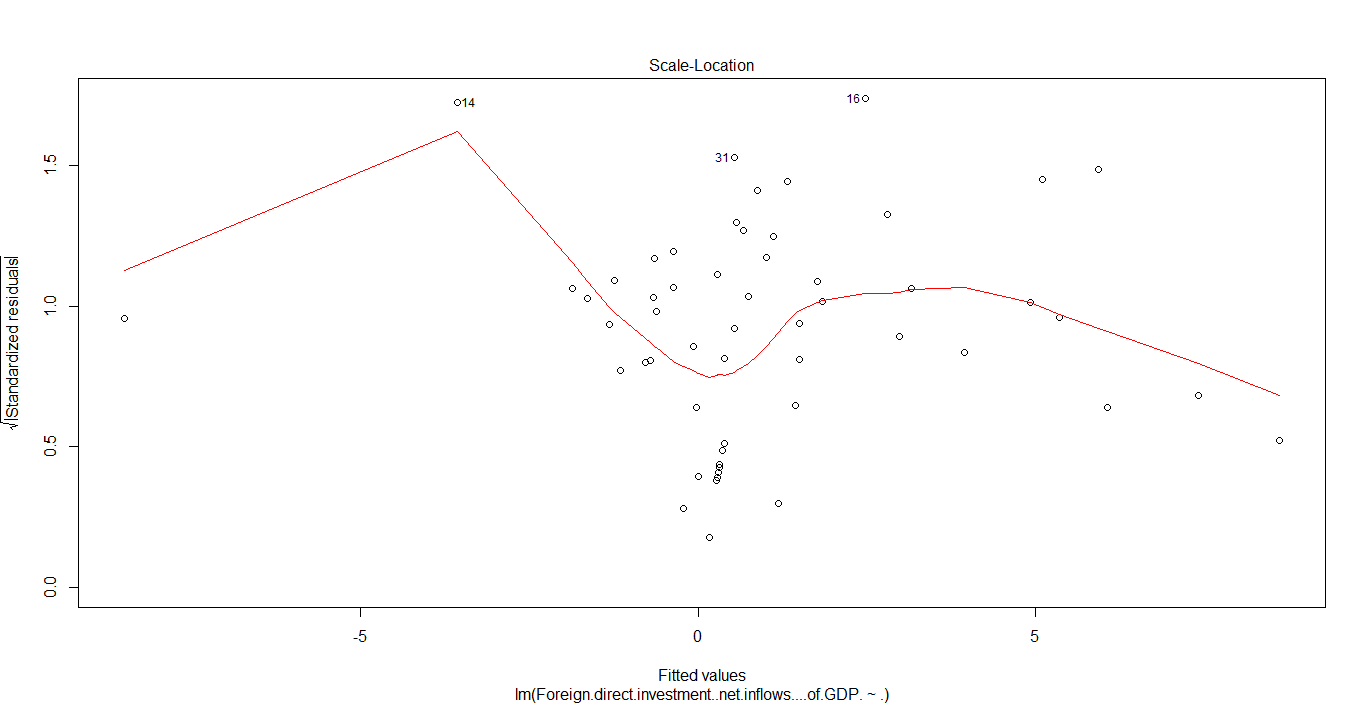
## 5. 4.1 Below plot showing multivariate Normality



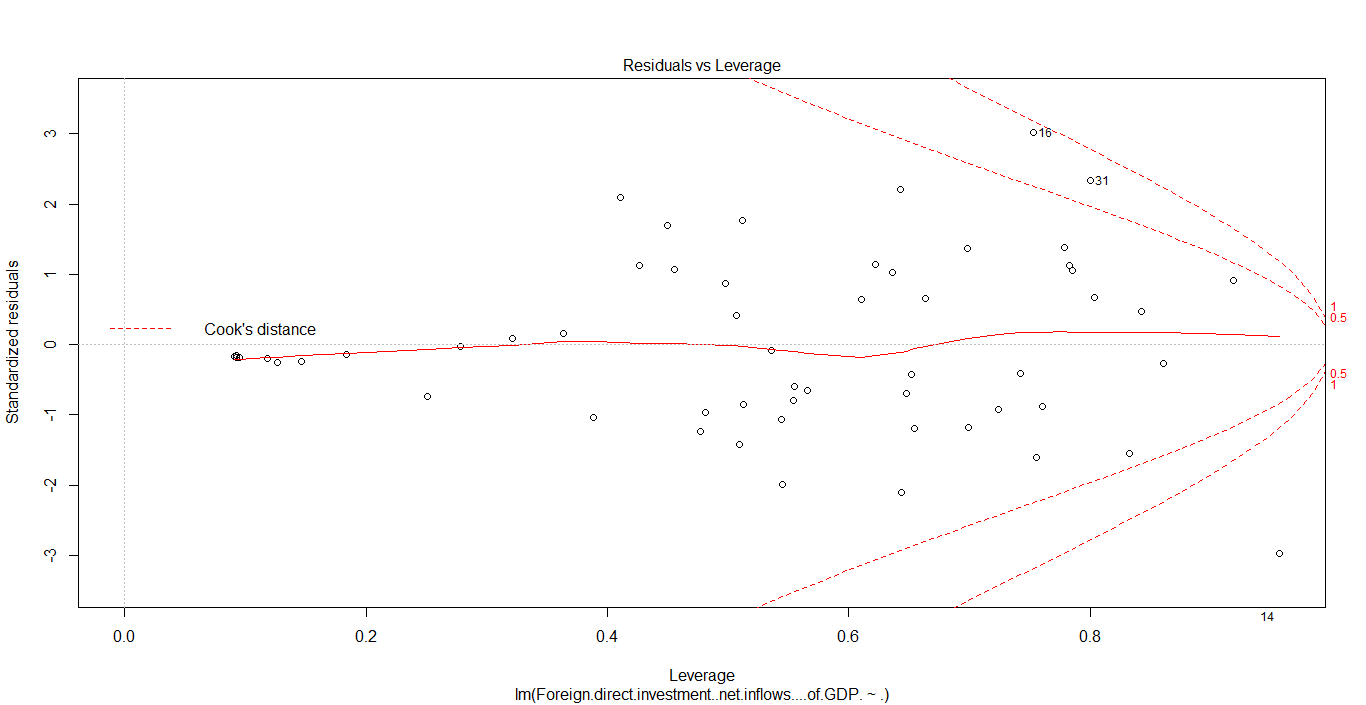
## 5.4.2 Below plot showing Residuals vs Fitted values



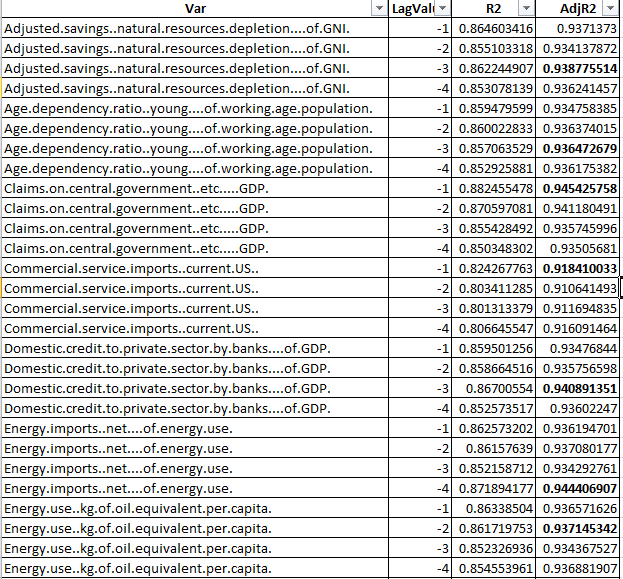
## 5.4.3 Below plot showing sqrt of Standardized Residuals vs Fitted values

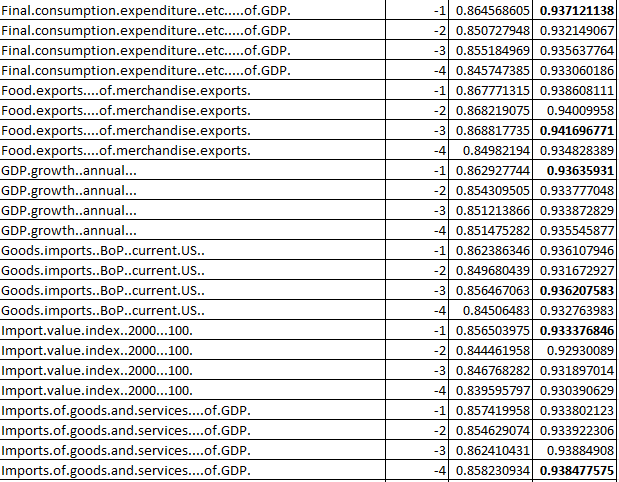


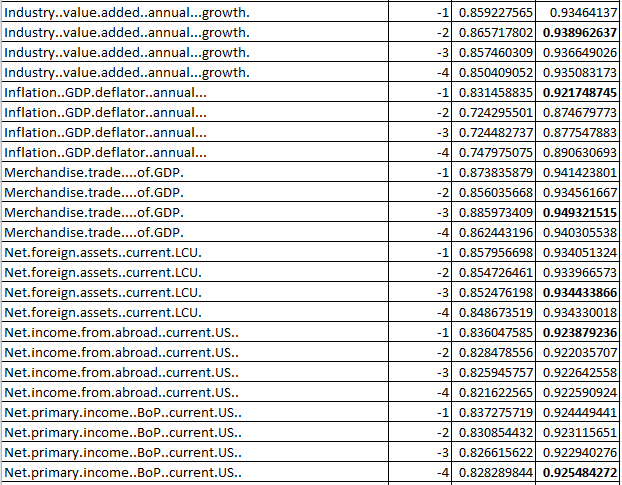
## 5.4.4 Below plot showing Homoscedacity

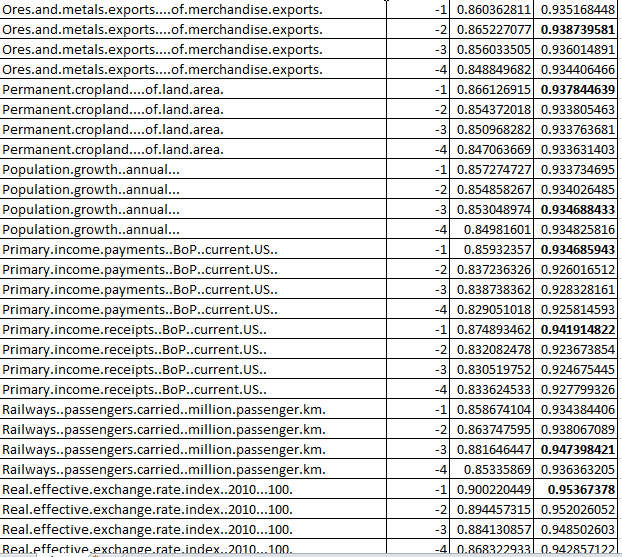


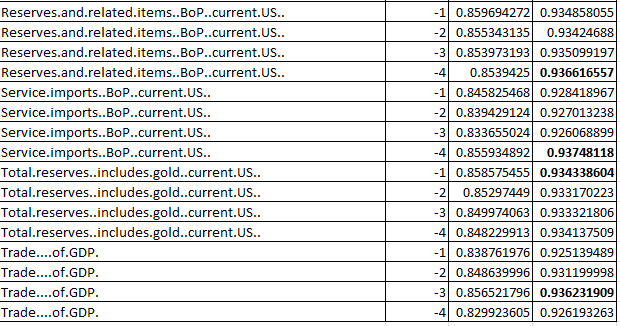
## Time series lags results











# Analysis

## Decision point

Factors determining FDI inflows into Kingdom of Saudi Arabia (KSA) obtained from the Regression model are listed below

* Commercial.service.imports..current.US..
* Inflation..GDP.deflator..annual
* Net.income.from.abroad..current.US
* Net.primary.income..BoP..current.US..
* Primary.income.payments..BoP..current.US..
* Primary.income.receipts..BoP..current.US.
* Real.effective.exchange.rate.index..2010...100
* Service.imports..BoP..current.US
* Trade....of.GDP

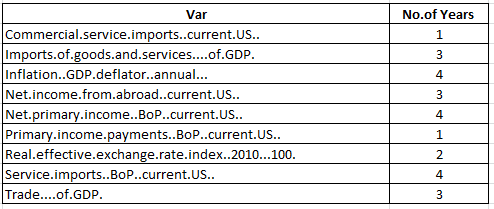
Top factors which are impacting more on attracting FDIs into KSA obtained from the stepAIC model are listed below

* Commercial.service.imports..current.US
* Imports.of.goods.and.services....of.GDP.

Saudi\_Arabian Government can attract the FDIs by concentrating on the above factors and can make policies in favor to the above determinants.

After implementing the policies and necessary steps, each factor would take the no.of years to

attract FDIs inflows is mentioned below.



## Scope of enhancements

* In the current model:
* By knowing the 1% increase in the values of each attributes , we could predict how much percentage increment will be there in the FDI inflows
* Lasso regression can be tried, but Linear regression is suitable due to its simplicity and we can explain the model by looking at the Adj.R2 values
* Hybrid algorithms can be used to find out the significant values

# Appendices

rm(list=ls(all=TRUE))

## Loading the data###

setwd("G:\\Project")

Saudiarabia\_data <- read.csv("Saudi Arabia.csv",header = T, sep = ",")

##Checking for NAs..

sum(is.na(Saudiarabia\_data))

summary(Saudiarabia\_data)

##There are are 1532 NAs, hence imputing the values using Time series Imputation..

library("imputeTS")

Main\_data = na.locf(Saudiarabia\_data, option = "nocb", na.remaining = "rev")

sum(is.na(Main\_data))

summary(Main\_data)

##Removing Year from Main\_data and train

Main\_data$Year=NULL

#Scaling the data

##library(vegan)

##Main\_data\_std <- data.frame(decostand(Main\_data,method = "standardize"),"Foreign.direct.investment..net.inflows....of.GDP." = Main\_data$Foreign.direct.investment..net.inflows....of.GDP.)

## Performing Random Forest to get the top variables based on the NodePurity levels

require(randomForest)

rf\_DL <- randomForest(Foreign.direct.investment..net.inflows....of.GDP. ~ ., data=Main\_data, keep.forest=TRUE, ntree=30)

rf\_DL

# importance of attributes

(rf\_DL$importance)

round(importance(rf\_DL), 2)

importanceValues = data.frame(attribute=rownames(round(importance(rf\_DL), 2)),

MeanDecreaseGini = round(importance(rf\_DL), 2))

importanceValues = importanceValues[order(-importanceValues$IncNodePurity),]

str(importanceValues)

# Top 30 Important attributes

Top30ImpAttrs = as.character(importanceValues$attribute[1:30])

Top30ImpAttrs

varImpPlot(rf\_DL)

###Considering the top 30 variables to run linear regression##

name<-c("Foreign.direct.investment..net.inflows....of.GDP.",

"Net.foreign.assets..current.LCU.",

"Net.primary.income..BoP..current.US..",

"Total.reserves..includes.gold..current.US..",

"Service.imports..BoP..current.US..",

"GDP.growth..annual...",

"Commercial.service.imports..current.US..",

"Real.effective.exchange.rate.index..2010...100.",

"Inflation..GDP.deflator..annual...",

"Permanent.cropland....of.land.area.",

"Reserves.and.related.items..BoP..current.US..",

"Energy.imports..net....of.energy.use.",

"Trade....of.GDP.",

"Primary.income.receipts..BoP..current.US..",

"Energy.use..kg.of.oil.equivalent.per.capita.",

"Net.income.from.abroad..current.US..",

"Primary.income.payments..BoP..current.US.." ,

"Domestic.credit.to.private.sector.by.banks....of.GDP.",

"Claims.on.central.government..etc.....GDP.",

"Goods.imports..BoP..current.US.." ,

"Adjusted.savings..natural.resources.depletion....of.GNI.",

"Railways..passengers.carried..million.passenger.km." ,

"Age.dependency.ratio..young....of.working.age.population." ,

"Food.exports....of.merchandise.exports." ,

"Imports.of.goods.and.services....of.GDP.",

"Population.growth..annual...",

"Final.consumption.expenditure..etc.....of.GDP.",

"Merchandise.trade....of.GDP.",

"Industry..value.added..annual...growth.",

"Import.value.index..2000...100.",

"Ores.and.metals.exports....of.merchandise.exports.")

Top30<-Main\_data[names(Main\_data)%in% name]

##Running the model on top30 attributes data###

LinReg1<-lm(Foreign.direct.investment..net.inflows....of.GDP. ~ ., data=Top30)

summary(LinReg1)

plot(LinReg1)

## Running stepAIC model to find out the significant variables

library(MASS)

stepAIC(LinReg1, direction = "both")

LinReg\_AIC1<-lm(Foreign.direct.investment..net.inflows....of.GDP.~ Commercial.service.imports..current.US..+Imports.of.goods.and.services....of.GDP.+

Net.income.from.abroad..current.US..+ Primary.income.receipts..BoP..current.US..+Import.value.index..2000...100.,data=Main\_data)

summary(LinReg\_AIC1)

#################Using Lasso Regression#############

####Defining target variable####

#Target Varaible

##Main\_data1 <- as.matrix(data.frame(Main\_data))

##str(Main\_data1)

##dim(Main\_data1)

##y=Main\_data$Foreign.direct.investment..net.inflows....of.GDP.[trainRows]

##ytest = Main\_data$Foreign.direct.investment..net.inflows....of.GDP.[-trainRows]

##head(ytest)

# Lasso Regression using glmnet - L1 norm

##library(glmnet)

##summary(train)

# fit model

##fit1 <- glmnet(train,y,alpha = 0,family = "gaussian")

#plot(fit1,xvar="lambda",label=TRUE)

#plot(fit1,xvar="dev",label=TRUE)

###TimeSeries###

library(DataCombine)

r=data.frame()

r

j=2

summary\_matrix = matrix(nrow=120, ncol = 4, dimnames = list(NULL, c('Var1','LagValue','R2','AdjR2')))

counter = 1

for(j in 2:31)

{

for (i in 1:4)

{

v = colnames(Top30[j])

x = slide(Top30, Var= v , NewVar = v, slideBy = -i)

model2 =lm(Foreign.direct.investment..net.inflows....of.GDP. ~.,data = x)

r[i,(j-1)]=summary(model2)$r.squared

summary\_matrix[counter,] = c(v,-i,summary(model2)$adj.r.squared,summary(model2)$r.squared)

print(summary\_matrix)

print(summary(model2)$adj.r.squared)

counter = counter + 1

}

}

Final\_result= as.data.frame(summary\_matrix)

Final\_result

write.csv(Final\_result,"data.csv", row.names=F)