Pension Expenditure Modeling: A Multifactor Risk Analysis

University of the Aegean
School of Sciences
Department of Mathematics
Track in Statistics and Actuarial - Financial Mathematics
MSc in Actuarial and Financial Mathematics

Kimon Ntotsis

2018

Contents

Acknowledgements

After an intensive period of almost a year, today is the day: writing this note of thanks is the finishing touch on my Master Thesis. It has been a period of intense learning for me, not only in the scientific arena, but also on a personal level. Writing this Thesis has had a big impact on me. I would like to reflect on the people who have supported and helped me so much throughout this period.

At first I would like to express my very great appreciation to my supervisor Professor Alexandros Karagrigoriou of the University of the Aegean for the patient guidance, encouragement and advice he has provided, both at scientific and at personal level, throughout my time as his student that will accompany me forever. I have been extremely lucky to be able to work with someone who has such passion about his work and unintentionally inspires everyone around him to want to do better.

I would also like to offer my special thanks to the staff of the Laboratory of Statistics and Data Analysis of the University of the Aegean and especially to my dear friends and colleagues Emmanouil Kalligeris, Dr. Christina Parpoula and Paschalini Trentou, with whom I have shared moments of deep anxiety but also of big excitement. Their presence was very important in a process that is often felt as a jenga tower ready to collapse. I greatly look forward to having all of you as colleagues in the years ahead.

In addition I would like to thank Assistant Professor Petros Hatzopoulos of the University of the Aegean and Mrs Marianna Papamichail Vice Chair-

person at National Actuarial Authority, who trusted me with this project and guided me throughout its duration.

Furthermore, I would like to thank Associate Professor Stylianos Xanthopoulos of the University of the Aegean and Assistant Professor Andreas Artemiou of Cardiff University for their contributions to this Thesis.

Some special words of gratitude go to my friends who have always been a major source of support when things get a bit discouraging. You all know who you are and what you have done for me the past year. Thanks guys for always being there for me.

Last but definetely not least I would also like to express my gratitude to my parents, Ioannis and Anna, and my sister Angeliki for their wise counsel and sympathetic ear. You are always there for me and I would not even imagine being here and doing what I love if you had not sacrificed so much.

Abstract

The aim of this master thesis is to locate, collect and analyze the factors which either on short-term or on long-term may have an impact on the shaping of the pension system especially the *Pension Expenditures*, for various European countries including Greece, as well as to use the previous analysis and attempt to create a platform for a future life assurance product to partly hedge pending state pension deficit. This product can be sold by insurance companies to people who wish to secure a stable future pension and wish to avoid possible future reductions.

Social security pensions, especially the minimum or the Pay As You Go part are normally guaranteed by the State. Much more benefits have been promised to be paid to citizens in the past than state budgeting, ageing or investment yields permit now days.

Insurance industry might provide lifetime smoothing first pillar pension benefits by valuating most of the factors affecting state *Pension Deficit*. In the European Union (EU) there are certain indicators that refer to the robustness of a state budget: S0 for the short run, S1 for the middle run and S2 for the long run. There are also indicators worldwide concerning the ageing of the population as well as fertility or migration. Labour markets, public or private investment and even corruption play a role on social security reductions.

Pension Expenditures along with *Pension Contributions* form the *Pension Deficit* which is the basis on which we rely on the creation of the product mentioned above.

Περίληψη

Σκοπός αυτής της μεταπτυχιακής διατριβής είναι να εντοπίσει, να συλλέξει και να αναλύσει τους παράγοντες οι οποίοι είτε βραχυπρόθεσμα είτε μακροπρόθεσμα μπορεί να έχουν αντίκτυπο στη διαμόρφωση του συνταξιοδοτικού συστήματος, και πιο συγκεκριμένα στις *Συνταξιοδοτικές Δαπάνες*, για διάφορες χώρες της Ευρώπης συμπεριλαμβανομένης και της Ελλάδας, καθώς και να αξιοποιήσει την προαναφερθείσα ανάλυση ώστε να προσπαθήσει να δημιουργήσει μια πλατφόρμα για ένα μελλοντικό προϊόν ασφάλισης ζωής το οποίο έχει ως στόχο να αντισταθμίσει εν μέρει το συνταξιοδοτικό έλλειμμα. Αυτό το προϊόν μπορεί να προσφερθεί από ασφαλιστικές εταιρείες σε άτομα που επιθυμούν αφενός να εξασφαλίσουν μια σταθερή μελλοντική σύνταξη και αφετέρου να αποφύγουν ενδεχόμενες μελλοντικές μειώσεις.

Οι συντάξεις κοινωνικής ασφάλισης, και πιο συγκεκριμένα το Pay As You Go, είναι συνήθως εγγυημένο από το κράτος. Πολλά περισσότερα οφέλη έχουν υποσχεθεί να καταβληθούν στους πολίτες στο παρελθόν από τον κρατικό προϋπολογισμο αλλά οι αποδόσεις των επενδύσεων δεν το επιτρέπουν στις μέρες μας.

Ο ασφαλιστικός κλάδος θα μπορούσε να παράσχει διαχρονικές συνταξιοδοτικές παροχές έχοντας ως πυλώνα την εκτίμηση των περισσότερων από τους παράγοντες που επηρεάζουν το *Συνταξιοδοτικό Έλλειμμα*. Στην Ευρωπαϊκή Ένωση (EU) υπάρχουν ορισμένοι δείκτες που αναφέρονται στην ευρωστία ενός κρατικού προϋπολογισμού: S0, S1 και S2. Υπάρχουν επίσης δείκτες σε παγκόσμιο επίπεδο σχετικά με τη γήρανση του πληθυσμού καθώς και τη γονιμότητα ή τη μετανάστευση. Οι αγορές εργασίας, οι δημόσιες ή ιδιωτικές

επενδύσεις και ακόμη και η διαφθορά παίζουν ρόλο στις μειώσεις της κοινωνικής ασφάλισης.

Οι *Συνταξιοδοτικές Δαπάνες* σε συνδυασμό με τις *Συνταξιοδοτικές Εισφορές* συνθέτουν το *Συνταξιοδοτικό Έλλειμμα*, το οποίο είναι η βάση πάνω στην οποία στηρίζουμε τη δημιουργία του προϊόντος που αναφέρθηκε παραπάνω.

Chapter 1

Dimension Reduction Techniques

For many years scientists, especially in the field of statistics, are involved with data science but in recent years it seems that a tendancy around *Big Data Analytics* exists. As IBM states "*Big data analytics is the use of advanced analytic techniques against very large, diverse data sets that include structured, semi-structured and unstructured data in different sizes.*

Big data is a term applied to data sets whose size or type is beyond the ability of traditional relational databases to capture, manage, and process the data with low-latency. And it has one or more of the following characteristics – high volume, high velocity, or high variety"[**26**].

The tremendous increase development of technology as well as the creation of new databases on a variety of topics day after day makes *Big Data Analytics* more efficient to work with. However, more is not always better. Large amounts of data might sometimes produce worse performances in data analytics applications. To deal with this, among other issues, special techniques, called *Dimension Reduction Techniques* were created.

In the field of statistics as well as others, *dimension reduction*, also called dimensionality reduction, is the process of reducing the number of random variables under consideration and creating a new smaller set of variables based on the original one. Through this process it is easier to interpret different statistical tests without losing the accuracy of the original variables. This master thesis is based on the *Principal Component Analysis*, dimension reduction technique as well as in its predecessor, *Beale et al.* dimension reduction technique.

1.1 Beale et al.

This technique is a very simple three steps procedure proposed by *E.M.L. Beale, M.G. Kendall* and *D.W.Mann* [**32**] for discarding variables in multivariate analysis. Their technique can be summed up as follows:

- Locate the minimum eigenvalue and the corresponding eigenvector.
- Locate the highest absolute value of the eigenvector. This value corresponds to a variable which will be removed from the model.
- Repeat the above steps until *p-k* variables have been removed.

Note that, p is the number of all variables and k is the number of eigenvalues which are larger that one.

1.2 Principal Component Analysis

PCA technique was proposed independently by *Pearson* [**45**] and *Hotelling* [**46**, **47**]. The thought behind the *Principal Component Analysis* is the conversion of a data set with interdependent variables into a new one with uncorrelated variables *(principal components)*, which will be arranged in such a way so that the first ones maintain the greater part of the variance that exists in all the original variables. With this procedure the reduction of the dimension of the original data set is achieved while leaving unchanged as much as possible, the variation [**4**].

The *Principal Component Analysis* relies on the the covariance or correlation matrix of the original dataset in order to obtain the eigenvalues and the eigenvectors which are essential in this procedure. The analysis consists of the following:

1.2.1 Covariance Matrix

The variance of a variable defined by the expectation of the squared deviation of a random variable from its mean [**2**] and can be expressed as follow:

$$\mathrm{Var}(X) = \mathrm{Cov}(X, X) = \mathrm{E}\left[(X - \mu)^2\right] = \mathrm{E}\left[X^2\right] - \mathrm{E}[X]^2 \qquad (1.1)$$

The covariance between 2 variables measures the variance between two variables, i.e. how one variable varies with another. If X and Y are the two variables, then the above can be expressed as follow:

$$\operatorname{cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))] = \mathrm{E}[XY] - \mathrm{E}[X]\,\mathrm{E}[Y] \tag{1.2}$$

while for n variables, $X_1, X_2, ..., X_n$, the covariance matrix is typically presented like this

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mathrm{E}(X_1))(X_1 - \mathrm{E}(X_1))] & \cdots & \mathrm{E}[(X_1 - \mathrm{E}(X_1))(X_n - \mathrm{E}(X_n))] \\ \mathrm{E}[(X_2 - \mathrm{E}(X_2))(X_1 - \mathrm{E}(X_1))] & \cdots & \mathrm{E}[(X_2 - \mathrm{E}(X_2))(X_n - \mathrm{E}(X_n))] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mathrm{E}(X_n))(X_1 - \mathrm{E}(X_1))] & \cdots & \mathrm{E}[(X_n - \mathrm{E}(X_n))(X_n - \mathrm{E}(X_n))] \end{bmatrix}. \tag{1.3}$$

1.2.2 Correlation Matrix

The correlation matrix is a statistical tool from which one distinguishes the strength of the relationship, if it exists, between any two variables involved in the matrix. The correlation between two variables is called correlation coefficient and ranges between -1 and 1 [**3**].

The population correlation coefficient $\rho_{X,Y}$ between two random variables X and Y with expected values $\mathrm{E}(X)$ and $\mathrm{E}(Y)$ and standard deviations σ_X and σ_Y is defined by:

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mathrm{E}(X))(Y - \mathrm{E}(Y))]}{\sigma_X \sigma_Y} \tag{1.4}$$

1.2.3 Eigenvalues - Eigenvectors

$$Av = \lambda v \tag{1.5}$$

As Cheever, E. states "*In this equation A is an n-by-n matrix, v is a non-zero n-by-1 vector and λ is a scalar (which may be either real or complex). Any value of λ for which this equation has a solution is known as an eigenvalue of the matrix A. It is sometimes also called the characteristic value. The vector v, which corresponds to this value is called an eigenvector*" [**28**].

Eigenvalues and eigenvectors are often focused on matrices and can be presented as follows:

$$\begin{bmatrix} A_{11} & A_{12} & \dots & A_{1n} \\ A_{21} & A_{22} & \dots & A_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{n1} & A_{n2} & \dots & A_{nn} \end{bmatrix} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \lambda \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \tag{1.6}$$

1.2.4 Extracted Eigenvalues - Eigenvectors

In order to determine the number of components in *PCA*, several techniques have been created. All those method are using the eigenvalues in order to achieve that. Most popular ones are:

- ***Kaiser's* rule**
  This technique keeps those components with eigenvalues greater than one.

- ***Scree Plot***
  Catell and Jasper's technique [**48**] is known also as scree test. According to this technique a plot with the eigenvalues sorted from maximum to minimum is created. At some point the curve connecting the values will create an angle, like an elbow, after which the curve will remain become a straight line parallele to the x-axis. The number of components selected is defined by the number of values that appear to that elbow.

- ***Proportion of variance explained***
  It is sometimes thought that a good factor analysis should explain two-thirds of the variance [**50**].

1.2.5 Components Matrix

The components are a set of uncorrelated vectors which have been created by the following methodology:

Let us denote by C_i, the i^{th} component, λ_i the corresponding eigenvalue and v_i the corresponding eigenvector.

Then,

$$C_i = -\sqrt{\lambda_i}\ v_i = -\sqrt{\lambda_i}\ \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} \tag{1.7}$$

1.2.6 New Variables

Each new variable (NV) is a linear combination between the components and the original data set.

Let us denote by

C : the components matrix
X : the original dataset matrix and
NV : the new variables matrix.

Then,

$$NV_{ij} = \sum_{r=1}^{n} X_{ir} C_{rj} \qquad \forall\ i, j = 1, ..., n \tag{1.8}$$

1.3 Application

To understand better the meaning of all the above, a very well known example used by *N.R. Draper* and *H. Smith* discussed below. Note that the data were first given in *Woods et al.* [**50**].

Note: *The interpretation of the results will be held in Chapter 3.*

X_1	X_2	X_3	X_4	Y
7,00	26,00	6,00	60,00	78,50
1,00	29,00	15,00	52,00	74,30
11,00	56,00	8,00	20,00	104,30
11,00	31,00	8,00	47,00	87,60
7,00	52,00	6,00	33,00	95,90
11,00	55,00	9,00	22,00	109,20
3,00	71,00	17,00	6,00	102,70
1,00	31,00	22,00	44,00	72,50
2,00	54,00	18,00	22,00	93,10
21,00	47,00	4,00	26,00	115,90
1,00	40,00	23,00	34,00	83,80
11,00	66,00	9,00	12,00	113,30
10,00	68,00	8,00	12,00	109,40

The variables shown in the above table are:

- X_1 = amount of tricalcium aluminate
- X_2 = amount of tricalcium silicate
- X_3 = amount of tetracalcium alumino ferrite
- X_4 = amount of dicalcium silicate
- Y = heat evolved in calories per gram of cement

X_1, X_2, X_3, X_4 are measured as percent of the weight of the clinkers from which the cement was made.
This example will be used in order to perfom *PCA* with the use of the correlation matrix. The following results are calculated with the use of *IBM SPSS* and *Microsoft Excel*.

Pearson Correlation Matrix				
	X1	X2	X3	X4
X1	1	,229	-,824	-,245
X2	,229	1	-,139	-,973
X3	-,824	-,139	1	,030
X4	-,245	-,973	,030	1

Note: *As it can be seen from the above correlation matrix, the use of PCA is essential due to the high correlation between* X_1 *with* X_3 *and* X_2 *with* X_4.

Total Variance Explained			
	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	2,236	55,893	55,893
2	1,576	39,402	95,294
3	,187	4,665	99,959
4	,002	,041	100,000

Total Variance Explained			
	Extraction Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %
1	2,236	55,893	55,893
2	1,576	39,402	95,294

Component Matrix		
	Component	
	1	2
1	,712	-,639
2	,843	,520
3	-,589	,759
4	-,819	-,566

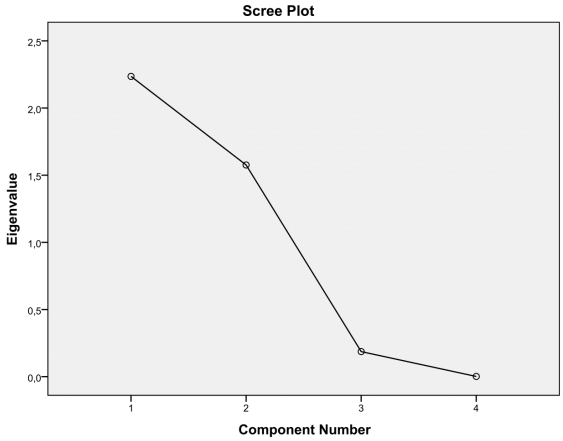


Figure 1.1: Scree Plot

Based on Beale et al. method as well as the total variance explained by each variable we come to the conclusion that the first two components must be kept to the model which contain more than 95% of the total information/-variability.

New Variables		
Variables	NV1	NV2
1	-25,7902	-20,359
2	-26,2819	-3,606
3	33,94062	16,843
4	-9,25498	-11,439
5	18,24839	8,443
6	30,86984	15,95
7	47,05659	44,51
8	-22,1662	7,275
9	18,31586	28,012
10	30,9148	-0,659
11	-6,97561	18,374
12	48,33565	27,33
13	49,89887	28,25

Chapter 2

Linear Regression Analysis

2.1 Regression Analysis

The term *regression* in statistics was first introduced and used by *Galton* [**53**], when in the course of an experiment he introduced the term regression to mediocrity. Nowadays the term regression analysis is well defined by D.N.Gujarati [**37**], who states that *"Regression analysis is concerned with the study of the dependence of one variable, the dependent variable, on one or more other variables, the explanatory variables, with a view to estimating and/or predicting the (population) mean or average value of the former in terms of the known or fixed (in repeated sampling) values of the latter"* [**36**].

2.1.1 Linear Regression

This type of regression is used when a linear relationship between the response *(dependent)* variable and the explanatory variable(s) *(independent)* occurs.

Simple linear regression

This type of regression is taking place when only one independent variable exists. The model formation is:

$$Y_i = \alpha + \beta X_i + \varepsilon_i \tag{2.1}$$

where, α and β are called coefficients of regression. More specific α is the

value of the dependent variable Y corresponding to the value X=0 of the dependent variable and β is the variation of the dependent variable corresponding to a unit change of X. Finally, ε_i is the error term that represents the deviation of the observed value from the true value of the quantity of interest.

Multiple linear regression

This type of regression is taking place when more than one independent variable exist. The model formation is:

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ...\beta_p X_{ip} + \varepsilon_i \tag{2.2}$$

Multivariate linear regression

A *multivariate*, also called factor, *analysis* is applied in models with one response variable and three or more multiple correlated dependent variables and is generally considered to be the extension of the two-way analysis and that is why the same patterns are followed. This method is used to identify which factors of a model are significant, if any interaction between them exist, and also the percentage of influence of the independent variables on the dependent one [**44**]. The formation of the model is the same as the *multiple regression analysis (2.2)*.

A typical linear regression analysis consists of the following:

- Estimates of coefficients
- Standard Error (SE)
- T- statistics
- P-value of t-statistics
- Number of observations
- Error degrees of freedom
- Root Mean Squared Error

- R-square
- Adjusted R-square
- F-statistics
- P-value of F-statistics

2.1.2 Interpretation of Linear Regression Elements

The Interpretation of the elements which constitute the *Linear Regression Analysis*, with X_1,...,X_p the number of the existing independent variables [**43**], it follows.

- **Estimate**
  This column will display the values for theregression coefficients for predicting the dependent variable from the independent variable, i.e. the values of α, β_1, ..., β_p.

- **SE**
  SE displays the standard error of each coefficient.

- **T-statistics**
  This column displays the t-statistic values, namely the values of the Student's t-test. A t-test is commonly used to determine whether a regression coefficient is significant, i.e. it differs or not from zero.
  In other words, the null hypothesis of this test is used to decide whether each variable is statistical significant. The null and alternatives are of the form:

$$H_0 : \beta_i = 0 \tag{2.3}$$

$$H_\alpha : \beta_i \neq 0 \tag{2.4}$$

  where *i=1,2,...,p*

- **P-value**
  This column displays the 2-tailed *p-values* associated with the t-test and are used to determine whether a given coefficient is significantly different from zero.

- **Number of observations**
  The number of observations is the size of the sample.

- **Error degrees of freedom**
  As it mentioned before in *ANOVA* interpretation the degrees of freedom of an estimate is the number of independent pieces of information that enter into the estimate calculation.

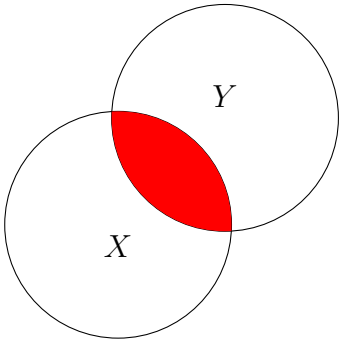
- **Root Mean Squared Error**
  Root Mean Squared Error, denoted by *(RMSE)*, is defined by the standard deviation of the variance. The *RMSE* of an estimator $\hat{\theta}$ is defined by:
  $$\mathrm{RMSE}(\hat{\theta}) = \sqrt{\mathrm{MSE}(\hat{\theta})} \tag{2.5}$$

- **R-square**
  R-square, denoted by R^2, represents the percentage of the total variability of the dependent variable interpreted on the basis of the regression model
  $$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{SST} \tag{2.6}$$
  where,
  - TSS is the Total Sum of Squares
  - ESS is the Explained Sum of Squares
  - RSS is the Residual Sum of Squares

  Note: *In some cases the symbols might differ. Explained Sum of Squares, ESS, can be seen as regression sum of squares, which symbolized by RSS and residuals sum of squares, RSS, can be seen as error sum of squares, which sumbolized as ESS. It is essential for the reader to focus on the*

meaning behind the symbolisation, so to be able to understand the procedure. For more details see Appendix A.

R^2 can be also seen via a *Venn diagram*. The one below displays a simple linear regression model with one dependent and one independent variable. R^2 is the two-circle intersection, which shows the extent to which the variation of Y is interpreted by the variation of X



- **Adjusted R-Square**
  Adjusted R-Square, denoted by R^2_{adj}, is used to decide about the usefulness of the independent variables in the model. The addition of a useless variable to the model, will cause decrease to the adjusted R-square, while the addition of a useful variable, will cause increase, but will never exceeds the R^2.

- **F-statistics**
  F-test and the associated p-value have been thoroughly analyzed in ANOVA interpretation.

2.2 Regression Assumptions

Analysis of variance as well as regression analysis answers some "questions" about the given data set that being analyzed. It should be noted though that those answers are not always trustworthy, or as in the field of statistics referred as statistically significant. To come to the conclusion that the model is significant some assuption must be fulfilled. Those assumptions are called *linear regression assumptions* and are based on the residuals ε_i of

the model. Depending on the formation *(i.e. simple, multiple, multivariate)* those assumptions might differ from case to case, but the following four are the most important and must be satisfied in every model formations [**41**].

- Normality
- Independence
- Homoscedasticity
- Linearity between Y and X_i

2.2.1 Normality

The errors must follow a normal distribution with zero mean and σ^2 variance, symbolized by:

$$X \sim \mathcal{N}(0, \sigma^2) \tag{2.7}$$

Most common tests that check this assumption are the *Lilliefors* test for normality, which is an improvement of the *Kolmogorov-Smirnov* test as well as the *Shapiro-Wilk* normality test.

$$H_0 : \textit{The residuals come from a standard normal distribution.} \tag{2.8}$$

$$H_\alpha : \textit{The residuals do not come from such a distribution.} \tag{2.9}$$

When it comes to real data sets, usually this assumption is not satisfied. For that reason some transformations have been proposed, which correct this problem. These transformations are mostly based on functions of the dependent variable with the *Logarithm, Root and Box & Cox* transformations being the most popular ones.

Note: *The following theorem provides the condition for the independence of residuals.*

Theorem 1 *If the residuals are normally distributed and uncorrelated, then they are independent. The inverse of the theorem is not true [**37**].*

2.2.2 Independence

The residuals must be independent. There are two ways that usually used in order to make a decision about the independence. The first one is with the implementation of *Runs* test for independency, which states the following hypotheses:

$$H_0 : \textit{The values of the residuals come in random order.} \tag{2.10}$$

$$H_\alpha : \textit{The values of the residuals do not come in random order.} \tag{2.11}$$

The second way to check independence is based on the previous theorem. After the normality has been checked, one can check the assumption of correlation via a *Durbin-Watson* test with the following hypotheses.

$$H_0 : \textit{The residuals are not autocorrelated.} \tag{2.12}$$

$$H_\alpha : \textit{The residuals are autocorrelated.} \tag{2.13}$$

2.2.3 Homoscedasticity

The errors must have the same finite variance. Levene's test for homoscedasticity is the most common test that checks this assumption, which states the following hypotheses:

$$H_0 : \textit{All population variances are equal.} \tag{2.14}$$

$$H_\alpha : \textit{At least one population variance differs from the others.} \tag{2.15}$$

2.2.4 Linearity

Linearity between Y and X_i can be seen through a scatter diagram.

Note: *In multiple or multivariate regression, the assumption of multicollinearity or simply collinearity, must also be checked. The collinearity is defined by the correlation among the predictor variables. The variance inflation factor, denoted by VIF, is the way to measure collinearity. The VIF expresses the rate at which the varience of the estimator increases when collinearity exist. A empirical rule states that if VIF > 10 then multicollinearity is considered high. Another way to locate the collinearity is with the use of the condition index (CI), which derived by the variance decomposition table. The following empirical rule is commonly applied in CI interpretation:*

- *If Condition Index <10, multicollinearity is small*
- *If 10 ≤ Condition Index ≤ 30, multicollinearity is medium*
- *If Condition Index >30, multicollinearity is high*

2.3 Graphical Assumptions Interpretation

In recent years more and more scientists in the field of statistics tend to use graphical representations of their data in order to come to conclusions about various tests. When it comes to *ANOVA* assumptions, this is not an exception. This view becomes more powerful day after day as new articles that support this theory are written, with one of the most influential to be an article by *Kozak et al.* [**40**].

The graphs that can check the assumptions above are:

- **Normality**
  The symmetry plot of residuals can be used for the interpretation of normality. A symmetrical distribution of the residuals around their median suggests the existence of normal distribution.

- **Independence**
  The residuals versus lagged residuals plot can be used for the interpretation of correlation. A trend among the residuals indicates a possible

correlation between them. If the residuals plots confirm the assumptions of correlations and normality, then the residuals are independent.

- **Homoscedasticity**
  The residuals versus the fitted values plot can be used for the interpretation of homoscedasticity. The increase in the variance as the fitted values increase suggests possible heteroscedasticity.

- **Linearity**
  Residuals versus every single one independent variable of the model.

2.4 Application

To understand the above theory, the same data set from the *Application 1.3* will be used for illustrative purposes.

Notes:

- The following results are calculated with the use of *IBM SPSS* and *Matlab*

- The interpretation of the results will be provided in Chapter 3.

- It must be noted that all four independent variables of the example will be used for the regression and not just the two that were kept in the model after the *Principal Component Analysis*. The reason is that *Linear Regression* and *PCA* techniques are not dependent and can be used separately.

Descriptive Statistics					
	X1	X2	X3	X4	Y
Mean	7,4615	48,1538	11,7692	30,0000	95,4231
Median	7,0000	52,0000	9,0000	26,0000	95,9000
Std. Deviation	5,88239	15,56088	6,40513	16,73818	15,04372
Variance	34,603	242,141	41,026	280,167	226,314
Minimum	1,00	26,00	4,00	6,00	72,50
Maximum	21,00	71,00	23,00	60,00	115,90
Range	20,00	45,00	19,00	54,00	43,40

Model Summary					
Model	R	R Square	Adjusted R Square	Std. Error of the Estimates	Durbin-Watson
1	,991	,982	,974	2,44601	2,053

ANOVA					
Model	Sum of Squared	df	Mean Square	F	Sig.
Regression	2667,899	4	666,975	111,479	,000
Residuals	47,864	8	5,983		
Total	2715,763	12			

Collinearity Diagnostics							
Model	Eigenvalue	Condition Index	Variance Proportions				
			(Constant)	X1	X2	X3	X4
1	4,120	1,000	,00	,00	,00	,00	,00
2	,554	2,727	,00	,01	,00	,00	,00
3	,289	3,778	,00	,00	,00	,00	,00
4	,038	10,462	,00	,06	,00	,05	,00
5	6,614E-5	249,578	1,00	,93	1,00	,95	1,00

Coefficients							
Model	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B		Collinearity Statistics
	B	Std. Error			Lower Bound	Upper Bound	VIF
(Constant)	62,405	70,071	,891	,399	-99,179	223,989	38,496
X1	1,551	,745	2,083	,071	-,166	3,269	38,496
X2	,510	,724	,705	,501	-1,159	2,179	254,423
X3	,102	,755	,135	,896	-1,638	1,842	46,868
X4	-,144	,709	-,203	,844	-1,779	1,491	282,513

```
Kolmogorov-Smirnov Normality test

 Kolm_Smirnov_pvalue
      0.1296


Runs test for independence

 Runs_Test_pvalue
      1.0000
```

Chapter 3

LabSTADA Application

In the context of operation of the *Laboratory of Statistics and Data Analysis* of the *University of the Aegean*, it was created an application within this master thesis that implements all the above theory that was presented in the 1st and 2nd *Chapter*.

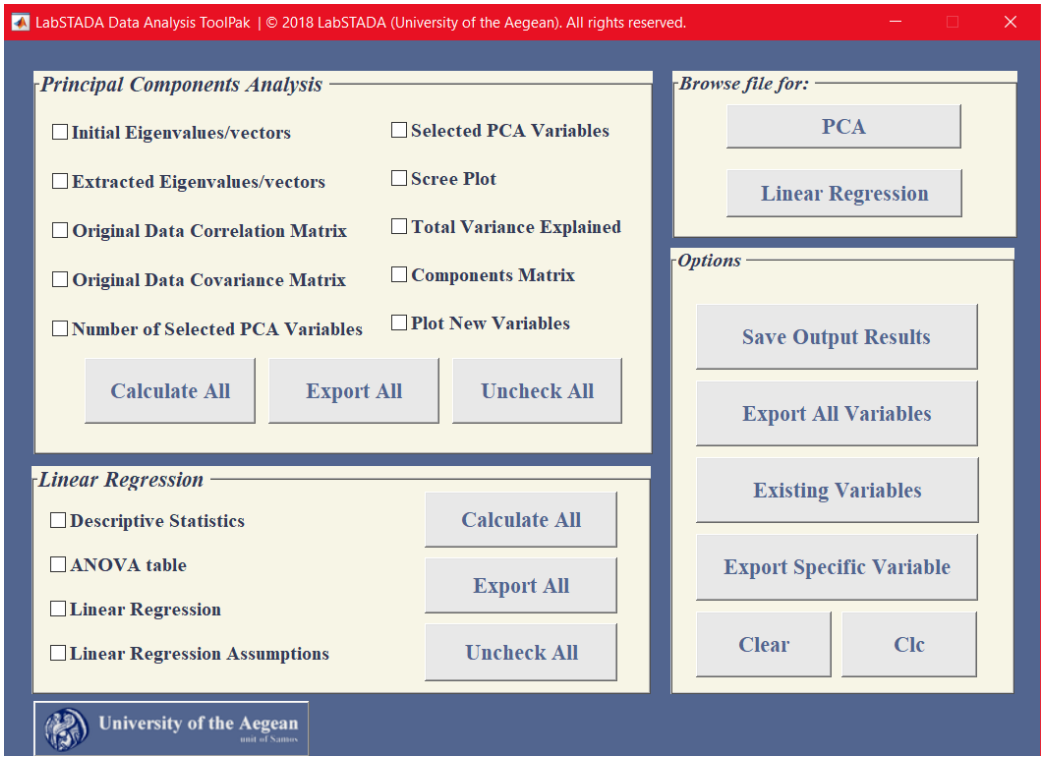
This application was created with *Matlab* code via *Matlab gui*. As MathWorks [**24**] states "*GUIs (also known as graphical user interfaces or UIs) provide point-and-click control of software applications, eliminating the need to learn a language or type commands in order to run the application.*

Matlab® apps are self-contained MATLAB programs with GUI front ends that automate a task or calculation. The GUI typically contains controls such as menus, toolbars, buttons, and sliders. Many MATLAB products, such as Curve Fitting Toolbox^{TM}, Signal Processing Toolbox^{TM}, and Control System Toolbox^{TM} include apps with custom user interfaces. You can also create your own custom apps, including their corresponding UIs, for others to use.

GUIDE (GUI development environment) provides tools to design user interfaces for custom apps. Using the GUIDE Layout Editor, you can graphically design your UI. GUIDE then automatically generates the MATLAB code for constructing the UI, which you can modify to program the behavior of your app" [**25**].

The goal of this task was to give someone the ability to perform *Principal Component Analysis* and *Linear Regression* as simply as possible. The only think that the user should beware of is the formation of the import data set. This application in addition to the results, it provides their interpretation which we consider very useful not only regarding the outputs but also because they can be used by someone not very familiar with these procedures.

While running *LabSTADA*, a windows command shell will be shown which will display the output results as well as the following window:



3.1 Options Description

For the application to work, one must verify that the *MATLAB Runtime version 9.0.1 (R2016a)* is installed.

3.1.1 Browse file for

Both files selection must follow some rules

1. The files must be in ***.xlsx*** form.
2. The files must contain ***only*** one sheet.
3. The files must be in the ***same*** folder as the *LabSTADA* application.

4. ***No*** letters must exists in the file, ***only*** numeric characters.

5. The data must begin from the cell ***A1***

PCA

This pushbutton gives the ability to the user to choose the file that wishes to use to perform the *Principal Component Analysis*.

Linear Regression

This pushbutton gives the ability to the user to choose the file that wishes to use to perform the *Linear Regression* **Caution:** The file must contain the dependent variable in the first column and the independent one(s) in the others.

3.1.2 Principal Component Analysis

Initial Eigenvalues/vectors

This check box displays the initial eigenvalues of the correlation matrix of the chosen data set arranged from maximum to minimum value along with the corresponding eigenvectors.

Extracted Eigenvalues/vectors

This check box displays the eigenvalues that their value is more than 0,99. Those eigenvalues usually are the ones that hold more than 80% of the information of the data and used for the creation of the principal components and then the new variables. Their corresponding eigenvectors displayed also.

Caution: This rule is known as *Kaiser's* technique but it is up to the user to choose the eigenvalue threshold. The LabSTADA app works by applying the above 0,99 rule.

Original Data Correlation Matrix

This check box will display the correlation matrix of the original data set values.

Original Data Covariance Matrix

This check box will display the covariance matrix of the original data set values.

Number of Selected PCA Variables

This check box will display the number of the new *PCA* variables that have been created.

Selected PCA Variables

This check box will display the new variables that have been created via *PCA*.

Scree Plot

This check box will display a figure of the components that hold at least 95% of the information of the original data along with the Pareto chart.

Total Variance Explained

This check box will display the total variance explained table.

Components

This check box will display the components that have been chosen from the correlation matrix of the initial data set.

Plot New Variables

This check box will display a figure with the values and linear relationship between two variables of the choice of the user.

Calculate All

This pushbutton gives the ability to the user to calculate all the above options at once.
Caution: This option calculates all the check boxes except the *Plot New Variables*

Export All

This pushbutton gives the option to the user to save all the above calculations in one *xlsx.* file with the name of his choice. Every sheet has the name of the check box that it contains.
Caution: For this button to work, **All** boxes must be calculated either by the *Calculate All* button or by checking each and every one box, otherwise error will occur.

Uncheck All

This pushbutton will uncheck all the checked boxes of *Principal Component Analysis* panel.

3.1.3 Linear Regression

Descriptive Statistics

This check box will display a *Descriptive statistics* table which will contain the mean, median,variance, standard deviation, minimum and maximum values of each variable.
Caution: The first variable is the dependent one and the rest are the independent ones.

ANOVA table

This check box will display the *ANOVA* table as well as the *Coefficient Confidence Intervals* and the *Variance Decomposition* tables of the data set.
Caution: The user has to interfere and choose the *Confidence level* of his choice, which the app will use to create the *Coefficient Confidence Intervals*. The value must be written using full stop (i.e. 0.95)

Linear Regression

This check box will display the *Linear Regression* table.

Linear Regression Assumptions

This check box will check the four assumptions of linear regression that have been mentioned in the previous chapter, as well as the multicollinearity

of the data set. This option will check the assumption both using statistical tests as well as graphs. The user has to interfere and choose the *significance level* α, which the app will use to interpret the assumptions.
Caution:

- ***statistical tests***
  The user has to interfere and choose the *significance level* α, which the app will use to interpret the assumptions.

- ***graphs***
  A figure with graphs of the three assumptions will be displayed with each one having as title the assumption which will be investigated. A second figure with the linearity assumption will also be presented. A *Graphs Explanation* text will be displayed as well to help with the interpretation.

Calculate All

This pushbutton gives the ability to the user to calculate all the above options at once.

Export All

This pushbutton gives the option to the user to save all above calculations in one *xlsx* file with the name of his choice. Every sheet has the name of the check box that it contains.
Caution: For this button to work **All** the boxes must be calculated either by the *Calculate All* button or by checking each and every one box, otherwise error will occurr.

Uncheck All

This pushbutton will uncheck all the checked boxes of *Principal Component Analysis* panel.

3.1.4 Options

Save Output Results

This pushbutton will save everything that the command window will display in a txt file named *Output_Results* in the same folder that the LabSTADA exists.

Export All Variables

This pushbutton gives the ability to the user to save all the *Options* (results) **except** the plot *Options* in the same *.xlsx* file.
Caution: This push button works **ONLY** when all options have been calculated, otherwise error will occur.

Existing Variables

This pushbutton displays all the existing variables in this app.

Export Specific Variable

This pushbutton gives the ability to the user to export in a *.xlsx* file one variable of his choice. The name of the variable must be displayed exactly as it appears in the *Exisitng Variables* pushbutton.

Clear

This pushbutton is useful when the program runs as an app inside *Matlab* environment and clears the workspace. When the app runs as a standalone, this button has no usefulness.

Clc

This pushbutton clears all input and output from the *Command Window* display.

Link

At the bottom left of the panel, there is a pushbutton that links the user to the official site of the *Laboratory of Statistics and Data Analysis* of the

University of the Aegean.

Note: *While runing the LabSTADA if the input data set has big values then, while inserting the file the values will be displayed as zeros multiplied with a value i.e.*

```
Chosen data set for Principal Components Analysis
   1.0e+11 *



    0.0000    0.0000    0.0000    1.5906    1.4895    0.0000    0.0000
    0.0000    0.0000    0.0000    2.4014    2.2937    0.0000    0.0000
    0.0000    0.0000    0.0001    0.5400    0.5435   -0.0000    0.0000
    0.0000    0.0000    0.0000    0.9814    0.8602    0.0000    0.0000
    0.0000    0.0000    0.0000    0.6537    0.5286    0.0000    0.0000
    0.0000    0.0000    0.0003    4.8951    4.7220    0.0000    0.0000
```

This is a transformation that Matlab does in order to make the dataset easy for display. The data set is the same as the user has chosen.

3.2 Application

In the previous *Chapters* the same data set for both *Principal Component Analysis* and *Linear Regression* was used. The results were implemented with the use of *IBM SPSS*, *Matlab* and *Microsoft Excel Office*. Now the same dataset will be used in *LabSTADA* to create the output of the previous theory. One can easily see that the results are the same as in *Chapters 1 & 2*.

Note: *The results are implemented with significance level* $\alpha = 5\%$

The following results provide the output of the *PCA* part of the *LabSTADA* app.

```
Chosen data set for Principal Components Analysis

     7    26     6    60
     1    29    15    52
    11    56     8    20
    11    31     8    47
```

```
 7    52     6    33
11    55     9    22
 3    71    17     6
 1    31    22    44
 2    54    18    22
21    47     4    26
 1    40    23    34
11    66     9    12
10    68     8    12

Eigenvalues Matrix

  2.2357         0         0         0
       0    1.5761         0         0
       0         0    0.1866         0
       0         0         0    0.0016

Eigenvectors Matrix

   0.4760    0.5090    0.6755    0.2411
   0.5639   -0.4139   -0.3144    0.6418
  -0.3941   -0.6050    0.6377    0.2685
  -0.5479    0.4512   -0.1954    0.6767

Extracted Eigenvalues Matrix (Eigenvalues >= 1)

   2.2357         0
        0    1.5761

Extracted Eigenvectors Matrix

   0.4760    0.5090
   0.5639   -0.4139
  -0.3941   -0.6050
  -0.5479    0.4512

The Components Matrix becomes as follows
```

```
   -0.7117   -0.6390
   -0.8431    0.5197
    0.5892    0.7595
    0.8193   -0.5665

Number of Selected PCA Variables

     2

New Variables via PCA

   25.7896  -20.3941
   26.2790   -3.6340
  -33.9433   16.8181
    9.2552  -11.4685
  -18.2520    8.4121
  -30.8724   15.9249
  -47.0637   44.4910
   22.1630    7.2536
  -18.3213   27.9915
  -30.9130   -0.6855
    6.9714   18.3549
  -48.3395   27.3060
  -49.9032   28.2248

Original Data Correlation Matrix

    1.0000    0.2286   -0.8241   -0.2454
    0.2286    1.0000   -0.1392   -0.9730
   -0.8241   -0.1392    1.0000    0.0295
   -0.2454   -0.9730    0.0295    1.0000

TOTAL VARIANCE EXPLAINED

Initial Eigenvalues
 Total_Initial_   Percentage_    Cumulative_
  Eigenvalues      Variance       Variance
---------------   ------------    -----------
```

```
   2.2357          0.55893         0.55893
   1.5761          0.39402         0.95294
   0.18661         0.046652        0.99959
   0.0016237       0.0004059              1

Extraction Sums of Squared Loadings

  New_Total_Initial_    New_Percentage_    New_Cumulative_
   Eigenvalues           Variance           Variance
  -----------------    -----------------   ---------------

    2.2357               0.55893             0.55893
    1.5761               0.39402             0.95294

Original Data Covariance Matrix

   34.6026   20.9231  -31.0513  -24.1667
   20.9231  242.1410  -13.8782 -253.4167
  -31.0513  -13.8782   41.0256    3.1667
  -24.1667 -253.4167    3.1667  280.1667
```

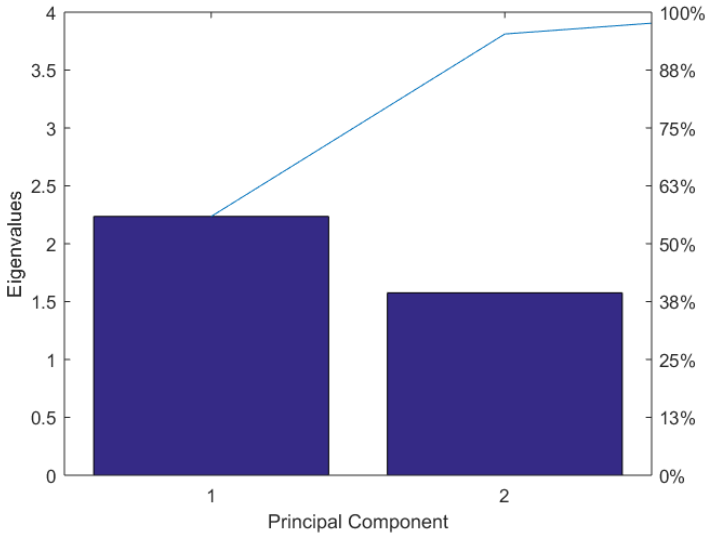


Figure 3.1: Scree Plot

Note: *Observe that the procedure chooses only 2 components that describe 95,294% of the total variability. Note also that the values of the new 2 PCA variables are been provided for the convenience of the user.*

The following results provide the output of the *Linear Regression* part of the *LabSTADA* app.

```
Chosen data set for Linear Regression

   78.5000    7.0000   26.0000    6.0000   60.0000
   74.3000    1.0000   29.0000   15.0000   52.0000
  104.3000   11.0000   56.0000    8.0000   20.0000
   87.6000   11.0000   31.0000    8.0000   47.0000
   95.9000    7.0000   52.0000    6.0000   33.0000
  109.2000   11.0000   55.0000    9.0000   22.0000
  102.7000    3.0000   71.0000   17.0000    6.0000
   72.5000    1.0000   31.0000   22.0000   44.0000
   93.1000    2.0000   54.0000   18.0000   22.0000
  115.9000   21.0000   47.0000    4.0000   26.0000
   83.8000    1.0000   40.0000   23.0000   34.0000
```

```
  113.3000   11.0000   66.0000    9.0000   12.0000
  109.4000   10.0000   68.0000    8.0000   12.0000


Descriptive_Statistics

 Maximum_     Minimum_     Average_     Median_     Standard_     Variance
  Value        Value        Value        Value       Deviation
---------    ---------    ---------    --------    ----------    --------

  115.9        72.5         95.423       95.9        15.044        226.31
   21           1            7.4615       7           5.8824        34.603
   71          26           48.154       52          15.561        242.14
   23           4           11.769        9           6.4051        41.026
   60           6           30           26          16.738        280.17

First row corresponds to the dependent variable followed by
the independent ones, in order.

Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4

Estimated Coefficients:

                   Estimate      SE        tStat       pValue
                   --------    -------    --------    --------

    (Intercept)      62.405     70.071      0.8906     0.39913
    x1               1.5511    0.74477      2.0827    0.070822
    x2              0.51017    0.72379     0.70486      0.5009
    x3              0.10191    0.75471     0.13503     0.89592
    x4             -0.14406    0.70905    -0.20317     0.84407


Number of observations: 13, Error degrees of freedom: 8
Root Mean Squared Error: 2.45
```

```
R-squared: 0.982,  Adjusted R-Squared 0.974
F-statistic vs. constant model: 111, p-value = 4.76e-07


ANOVA
                SumSq     DF    MeanSq      F         pValue
                ------    --    ------    ------    ----------

    Total       2715.8    12    226.31
    Model       2667.9     4    666.97    111.48    4.7562e-07
    Residual    47.864     8     5.983

Coefficient_Confidence_Intervals

  -99.1786  223.9893
   -0.1663    3.2685
   -1.1589    2.1792
   -1.6385    1.8423
   -1.7791    1.4910

Kolmogorov-Smirnov Normality test

Test Hypothesis:
Ho: The Residuals come from a standard normal distribution
Ha: The Residuals do not come from such a distribution
Kolm_Smirnov_pvalue
    0.1296

The test fail to reject the null hypothesis at the chosen significance level


Durbin Watson autocorrelation test

Test Hypothesis:
Ho: The residuals are not autocorrelated
Ha: The residuals are autocorrelated

Durbin_Watson_pvalue =
```

```
    0.6285

The test fail to reject the null hypothesis at the Assumptions
Significance Level


Runs test for independence

Test Hypothesis:
Ho: The values of the residuals come in random order
Ha: The values of the residuals do not come in random order
Runs_Test_pvalue
    1.0000

The test fail to reject the null hypothesis at the chosen significance level


Collinearity test

VIF:

   38.4962  254.4232   46.8684  282.5129

The VIF expresses the rate at which the estimator's variance increases when
collinearity exist.
If VIF > 10 then multicollinearity is high

Variance Decomposition

 sValue   condIdx   var1    var2    var3    var4    var5
-----------------------------------------------------------
 2.0262     1      0.0000  0.0004  0.0001  0.0013  0.0006
 0.7491    2.7049  0.0000  0.0083  0.0000  0.0182  0.0026
 0.5441    3.7236  0.0000  0.0012  0.0009  0.0088  0.0295
 0.1928   10.5083  0.0005  0.0601  0.0092  0.2748  0.0234
 0.0173  117.0267  0.9995  0.9301  0.9898  0.6969  0.9439
```

```
Condition_Indices
    1.0000
    2.7049
    3.7236
   10.5083
  117.0267


If Condition Index < 10, multicollinearity is small
If 10 <= Condition Index <= 30, multicollinearity is medium
If Condition Index > 30, multicollinearity is high

Graphs Explanation

For Normality:
An equal distribution of the residuals around their median suggests
the existence of normal distribution

For Homoscedasticity:
The increase in the variance as the fitted values increase suggests
possible heteroscedasticity
On the contrary, the decrease in the variance as the fitted values increase
suggests possible homoscedasticity

For Autocorrelation:
A trend among the residuals indicates a possible correlation between them

For Independence:
If the data are normally distributed and uncorrelated, then they are
independent

For Linearity:
If the data do not exhibit a pattern, then linearity could be accepted
```

Notes:

- For illustrative purposes we use the original data although one may wish to use the new data obtained after the implementation of the *PCA* procedure.

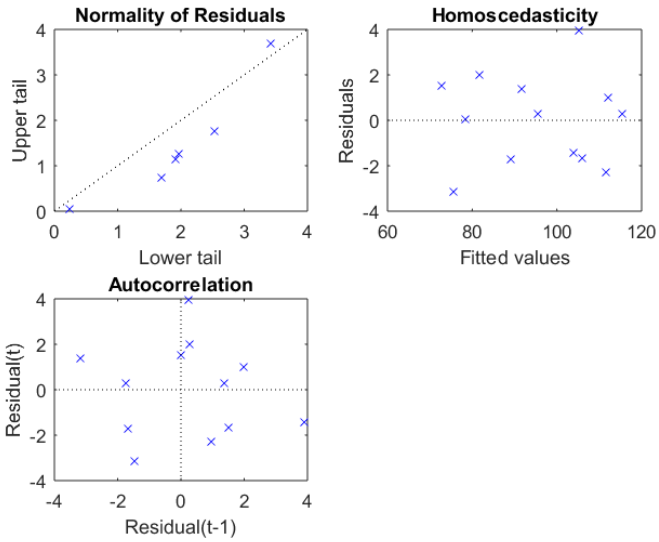


Figure 3.2: Linear Regression Assumptions Graphs

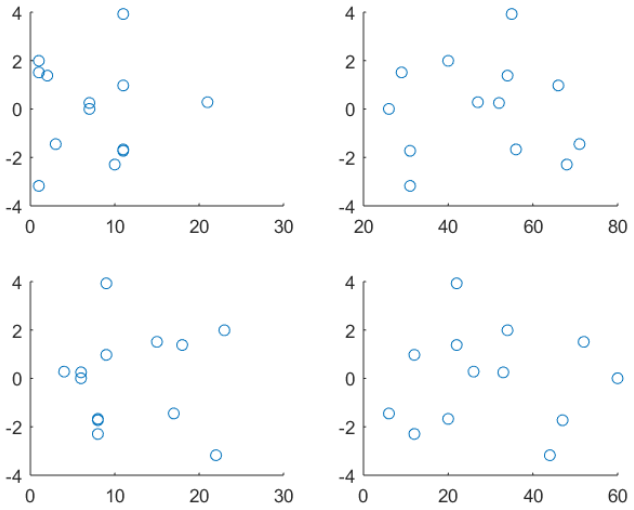


Figure 3.3: Linear Regression Linearity Assumption

- The *ANOVA* table together with the coefficient estimates are provided.
- the app is not designed to perform model selection but one may choose to repeat the app with the variables that appear to be significant according to the above output in conjunction with the assumptions investigated below.
- The app provides also the results for the regression assumptions including collinearity.

Chapter 4

Actuarial Project

4.1 Variables Selection

4.1.1 Pension Expenditure

The purpose of this project is to model the *Pension Expenditures* of European countries including Greece. As *OECD* states, *Pension expenditure, also named pension spendings, is defined by all cash expenditures (including lump-sum payments) on old-age and survivors pensions. Old-age cash benefits provide an income for persons retired from the labour market or guarantee incomes when a person has reached a 'standard' pensionable age or fulfilled the necessary contributory requirements. This category also includes early retirement pensions: pensions paid before the beneficiary has reached the 'standard' pensionable age relevant to the programme. It excludes programmes concerning early retirement for labour market reasons. Old-age pensions includes supplements for dependants paid to old-age pensioners with dependants under old-age cash benefits. Old age also includes social expenditure on services for the elderly people, services such as day care and rehabilitation services, home-help services and other benefits in kind. It also includes expenditure on the provision of residential care in an institution. This indicator is measured in percentage of GDP broken down by public and private sector. Private pension spending includes payments made to private pension plan members, or dependants after retirement and covers persons working in both the public and private sectors.*

The modeling of *pension expenditures*, according to the relevant theory, is based on 20 explanatory variables which most likely are related and pos-

sibly affect pension expenditures. For this project 20 European countries were selected based on the completeness of the available data mostly derived through *Knoema, OECD and Eurostat*. The data which are annual cover the period 2001 to 2015. Note that at the time of this work the data for 2016 and 2017 were not fully available. Based on the available data, three datasets were created corresponding to the periods 2001-2005, 2006-2010 and 2011-2015. The value of each variable for each time period is taken to be equal to the average of all values of the specific variable for the specific time period. For convenience and future reference the datasets are denoted by A, B and C respectively.

The selected countries, in alphabetical order, are:

- Austria
- Belgium
- Czech Republic
- Denmark
- Finland
- France
- Germany
- Greece
- Iceland
- Italy
- Latvia
- Netherlands
- Poland
- Portugal
- Slovak Republic
- Republic of Slovenia

- Spain
- Sweden
- Switzerland
- United Kingdom

4.1.2 Independent Variables

Based on the relevant theory on *Pension Expenditure*, the selected variables which are likely to affect it are the following:

Note: *The descriptions of the variables are presented as displayed in the original sites, from which the data were derived* [**18-23**].

- **Gross Domestic Product**

  Gross domestic product, denoted by GDP, at purchaser's prices is the sum of gross value added by all resident producers in the economy plus any product taxes and minus any subsidies not included in the value of the products. It is calculated without making deductions for depreciation of fabricated assets or for depletion and degradation of natural resources. Data are in current U.S. dollars.

- **Unemployment Rate**

  Unemployment rate can be defined by either the national definition, the ILO harmonized definition, or the OECD harmonized definition. The OECD harmonized unemployment rate gives the number of unemployed persons as a percentage of the labor force (the total number of people employed plus unemployed). [OECD Main Economic Indicators, OECD, monthly] As defined by the International Labour Organization, "unemployed workers" are those who are currently not working but are willing and able to work for pay, currently available to work, and have actively searched for work.(ILO, http://www.ilo.org/public/english/bureau/stat/res/index.htm)

- **Total Labor Force**

  Labor force participation rate is the proportion of the population ages 15-64 that is economically active: all people who supply labor for the production of goods and services during a specified period.

- **Exports of goods and services**

  Exports of goods and services comprise all transactions between residents of a country and the rest of the world involving a change of ownership from residents to nonresidents of general merchandise, net exports of goods under merchanting, nonmonetary gold, and services. Data are in current U.S. dollars.

- **Imports of goods and services**

  Imports of goods and services comprise all transactions between residents of a country and the rest of the world involving a change of ownership from nonresidents to residents of general merchandise, nonmonetary gold, and services. Data are in current U.S. dollars.

- **Current Account Balance**

  Current account, denoted by CAB, refers to all transactions other than those in financial and capital items. The major classifications are goods and services, income and current transfers. The focus of the BOP is on transactions (between an economy and the rest of the world) in goods, services, and income.

- **Investments**

Capital investment refers to funds invested in a firm or enterprise for the purpose of furthering its business objectives. Capital investment may also refer to a firm's acquisition of capital assets or fixed assets such as manufacturing plants and machinery that is expected to be productive over many years. Sources of capital investment are manifold and can include equity investors, banks, financial institutions, venture capital and angel investors.

- **Consumer price index**

  Consumer price index reflects changes in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used.

- **Median age of the population**

  Median age is the age that divides a population into two numerically equal groups - that is, half the people are younger than this age and half are older.

- **Number of births**

  Number of births is the total number of live births per thousands in a population in a year or period.

- **Net number of migrants**

  The difference between the immigrants and emigrants of a country.

- **Demographic Dependency**

population ages 65+ divided by the population ages 0-64.

- **Inflation**

  Inflation as measured by the consumer price index reflects the annual percentage change in the cost to the average consumer of acquiring a basket of goods and services that may be fixed or changed at specified intervals, such as yearly. The Laspeyres formula is generally used. Data for inflation are averages for the year, not end-of-period data.

- **Long-term interest rates**

  Long-term interest rates refer to government bonds maturing in ten years. Rates are mainly determined by the price charged by the lender, the risk from the borrower and the fall in the capital value. Long-term interest rates are generally averages of daily rates, measured as a percentage. These interest rates are implied by the prices at which the government bonds are traded on financial markets, not the interest rates at which the loans were issued. In all cases, they refer to bonds whose capital repayment is guaranteed by governments.

- **Short-term interest rates**

  Short-term interest rates are the rates at which short-term borrowings are effected between financial institutions or the rate at which short-term government paper is issued or traded in the market. Short-term interest rates are generally averages of daily rates, measured as a percentage. Short-term interest rates are based on three-month money market rates where available. Typical standardised names are "money market rate" and "treasury bill rate".

- **Total saving rate**

Saving is the difference between disposable income plus the change in net equity of households in pension funds and final consumption expenditure. Saving therefore reflects the residual income used to acquire financial and non-financial assets. It is important to note that disposable income does not include any capital gains or indeed losses, and so neither does saving. Net saving is equal to saving less depreciation.

- **Total household spendings**

Household spending is the amount of final consumption expenditure made by resident households to meet their everyday needs, such as: food, clothing, housing (rent), energy, transport, durable goods (notably, cars), health costs, leisure, and miscellaneous services.

- **Total household saving**

Net household saving is defined by the subtraction of household consumption expenditure from household disposable income, plus the change in net equity of households in pension funds. Household saving is the main domestic source of funds to finance capital investment, a major impetus for long-term economic growth. This indicator is measured as a percentage of household disposable income.

- **Private sector debt**

The private sector debt is the stock of liabilities held by the sectors Non-Financial corporations and Households and Non-Profit institutions serving households. The instruments that are taken into account to compile private sector debt are Debt securities and Loans

- **Compensation of employees**

Compensation of employees is defined by the total remuneration, in cash or in kind, payable by an employer to an employee in return for

work done by the latter during the accounting period. Compensation of employees consists of wages and salaries, and of employers' social contributions.

4.2 Discarding Variables Technique

In order to discard some variables with limited usefull information the *Beale et al. (1967)* technique was used. In all three time periods this technique was not completed, instead it was chosen to be stopped when, based on theory, a very important variable was proposed for rejection. Then, in *Section 4.3*, the *Principal Component Analysis* will be implemented for the 2^{nd} step of the dimension reduction procedure.

4.2.1 2001-2005 Dataset

For this time period out of the original *p=20* variables only *k=3* were found with eigenvalues greater than one. Thus, according to the procedure 17 variables must be discarded.

Below the rejected variables are displayed in the order they have been rejected. The associated p-value based on which each variable is rejected, is given in parenthesis.

1. Total Household Saving (0.53)
2. Exports of Goods and Services (0.64)
3. Total Household Spendings (0.70)
4. Total Saving Rate (0.49)
5. GDP (0.65)
6. Imports of Goods and Services (0.49)
7. Number of Births (0.75)
8. Total Labor Force (0.64)
9. Short-term Interest Rates (0.57)
10. Inflation Rate (0.51)
11. Median Age of Population (0.57)
12. Net Number of Migrants (0.63)

13. Long-term Interest Rates (0.67)
14. Investments (0.56)
15. Demographic Dependency (0.71)
16. Compensation of Employees (0.68)
17. CPI (0.73)

It was decided to interrupt the procedure after the 4^{th} iteration/extraction since GDP, which was chosen to be removed in the 5^{th} iteration, is considered as one of the most important and significant variables. The variability described by the 16 variables remained after the 4^{th} iteration is 79% of the total variability of the data.

The top three most significant variables according to the *Baile et al.* technique are the following.

1. Unemployment Rate
2. CAB
3. Private Sector Debt

4.2.2 2006-2010 Dataset

For this time period out of the original *p=20* variables only *k=4* were found with eigenvalues greater than one. Thus, according to the procedure 16 variables must be discarded.

Below the rejected variables are displayed in the order they have been rejected. The associated p-value based on which each variable is rejected, is given in parenthesis.

1. Exports of Goods and Services (0.59)
2. Short-term Interest Rates (0.49)
3. GDP (0.66)
4. Number of Births (0.66)
5. Total Household Spendings (0.64)

6. Total Saving Rate (0.52)
7. Imports of Goods and Services (0.57)
8. Long-term Interest Rates (0.65)
9. Private Sector Debt (0.52)
10. Median Age of Population (0.63)
11. Demographic Dependency (0.68)
12. Inflation Rate (0.68)
13. Total Household Saving (0.71)
14. Net Number of Migrants (0.60)
15. Compensation of Employees (0.71)
16. Investments (0.73)

It was decided to interrupt the procedure after the 2^{nd} iteration/extraction since GDP, which was chosen to be removed in the 3^{rd} iteration, is considered as one of the most important and significant variables. The variability described by the 18 variables remained after the 2^{nd} iteration is 90% of the total variability of the data.

The top four most significant variables according to the *Baile et al.* technique are the following.

1. Unemployment Rate
2. Total Labor Force
3. CAB
4. CPI

4.2.3 2011-2015 Dataset

For this time period out of the original *p=20* variables only *k=4* were found with eigenvalues greater than one. Thus, according to the procedure 16 variables must be discarded.

Below the rejected variables are displayed in the order they have been rejected. The associated p-value based on which each variable is rejected, is given in parenthesis.

1. Exports of Goods and Services (0.69)
2. Total Household Spendings (0.74)
3. GDP (0.65)
4. Total Labor Force (0.56)
5. Short-term Interest Rates (0.68)
6. Private Sector Debt (0.63)
7. Imports of Goods and Services (0.49)
8. Total Saving Rate (0.67)
9. Inflation Rate (0.52)
10. Net Number of Migrants (0.43)
11. Median Age of Population (0.60)
12. Total Household Saving (0.54)
13. Investments (0.67)
14. Unemployment Rate (0.69)
15. Demographic Dependency (0.65)
16. Compensation of Employees (0.71)

It was decided to interrupt the procedure after the 2^{nd} iteration/extraction since GDP, which was chosen to be removed in the 3^{rd} iteration, is considered as one of the most important and significant variables. The variability described by the 18 variables remained after the 2^{nd} iteration is 89.9% of the total variability of the data.

The top four most significant variables according to the *Baile et al.* technique are the following

1. CAB

2. CPI

3. Number of Births

4. Long-term interest rates

4.2.4 Conclusion

Table 4.2.4 presents the results of *Beale et al.* method for all three datasets examined.

Beale et al. Discarding Variables Technique Synopsis			
Dataset	**2001-2005**	**2006-2010**	**2011-2015**
1^{st} **extraction**	Total Household Saving	Exports of Goods and Services	Exports of Goods and Services
2^{nd} **extraction**	Exports of Goods and Services	Short-term Interest Rates	Total Household Spendings
3^{rd} **extraction**	Total Household Spendings	GDP	GDP
4^{th} **extraction**	Total Saving Rate	Number of Births	Total Labor Force
5^{th} **extraction**	GDP	Total Household Spendings	Short-term Interest Rates
6^{th} **extraction**	Imports of Goods and Services	Total Saving Rate	Private Sector Debt
7^{th} **extraction**	Number of Births	Imports of Goods and Services	Imports of Goods and Services
8^{th} **extraction**	Total Labor Force	Long-term Interest Rates	Total Saving Rate
9^{th} **extraction**	Short-term Interest Rates	Private Sector Debt	Inflation Rate
$10th$ **extraction**	Inflation Rate	Median Age of Population	Net Number of Migrants
11^{th} **extraction**	Median Age of Population	Demographic Dependency	Median Age of Population
12^{th} **extraction**	Net Number of Migrants	Inflation Rate	Total Household Saving
13^{th} **extraction**	Long-term Interest Rates	Total Household Saving	Investments
14^{th} **extraction**	Investments	Net Number of Migrants	Unemployment Rate
15^{th} **extraction**	Demographic Dependency	Compensation of Employees	Demographic Dependency
16^{th} **extraction**	Compensation of Employees	Investments	Compensation of Employees
17^{th} **extraction**	CPI		

Based on the above discussion and taking into consideration the significance of the *Gross Domestic Product (GDP)* we arrived at the conclusion that, irrespectively of the time period, only those variables extracted from the original data before *GDP* must be excluded from further analysis, while all other variables should remain and considered for the next step of the reduction process.

As a result the variables *Exports of goods and services*, *Total Household Spendings*, *Short-term interest rates*, *Total Household Saving*, *Total Saving Rate* are being extracted from all three datasets.

The reduction process continues in the next section, with the implementation of the *Principal Component Analysis (PCA)* using the following 15 variables for each time period.

Notes:

- *PCA* appears to be necessary due to high correaltion between at least some of the variables involved *(the correlation matrix appears in Appendix **Actuarial Project Results**)*.
- The remaining 15 independent variables which will be used are the following:
  - Gross Domestic Product
  - Unemployment Rate
  - Total Labor Force
  - Imports of goods and services
  - Current Account Balance
  - Investments
  - Consumer price index
  - Median age of the population
  - Number of births
  - Net number of migrants
  - Demographic Dependency
  - Inflation
  - Long-term interest rates

- Private sector debt
- Compensation of employees

4.3 Principal Component Analysis

In this section we apply the *Principal Component Analysis* and obtain the full 15 *principal components* for each time period with the corresponding eigenvalues ranging from almost eight to nearly zero. Based on the overall results *(see Appendix **Actuarial Project Results**)* and the fact that we did not want to lose important information of the model, we came to the conclusion that the first seven components should be kept regardless of the eignvalues because they contain a considerable amount of the total information/variability. The described variability played a key role in our decision since our intention was to keep that many components so that a considerable proportion of the original variability will be decribed by the components chosen. Note that the seven first components have variability around 75% of the original data for each of the three time periods.

Note: *To determine which variables are significant in every component, the procedure that was used is the following. For the first two out of seven components we keep as significants variables, those which absolute value is* $\geqslant$ *0.8. For the rest five components we keep as significant only one variable,the one that has the highest absolute value among all.(The reason to do this is that the last five components have not as large information as the first two)*

4.3.1 2001-2005 Dataset

The result of *PCA* are the following:

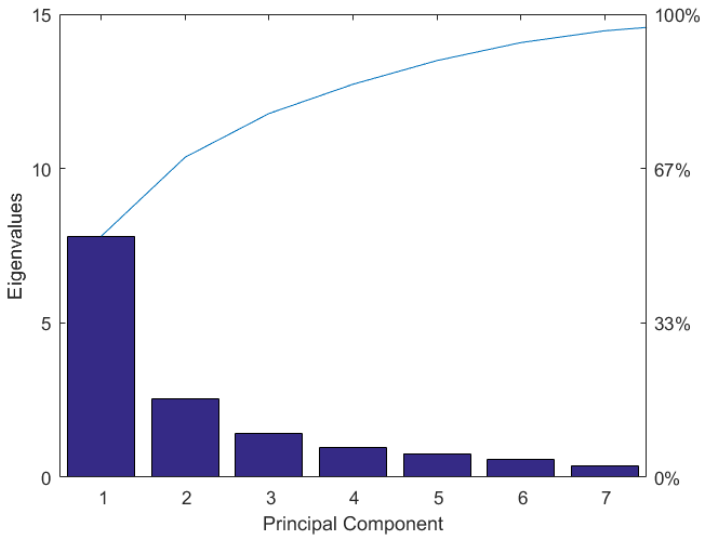


Figure 4.1: Scree Plot

Interpretation of Components

- 1^{st} **component**
  The first component holds the 52% of the information/variability. The variables that are statistical significant in this component are
  - GDP
  - Total Labor Force
  - Imports of goods and services
  - Investments
  - Number of Births
  - Inflation
  - Private Sector Debt
- 2^{nd} **component**
  The second component holds the 17% of the information. The variables that are statistical significant in this component are
  - Median age of the population

  - Long-term interest rates

- 3^{rd} **component**
  The third component holds the 9.4% of the information. The variable that is statistical significant in this component is the *CAB*

- 4^{th} **component**
  The forth component holds the 6.2% of the information. The variable that is statistical significant in this component is the *Unemployment Rate*

- 5^{th} **component**
  The fifth component holds the 5% of the information. The variable that is statistical significant in this component is the *Compensation of Employees*

- 6^{th} **component**
  The sixth component holds the 3.8% of the information. The variable that is statistical significant in this component is the *Investments*

- 7^{th} **component**
  The seventh component holds the 2.5% of the information. The variable that is statistical significant in this component is the *Compensation of Employees*

4.3.2 2006-2010 Dataset

The result of *PCA* are the following:

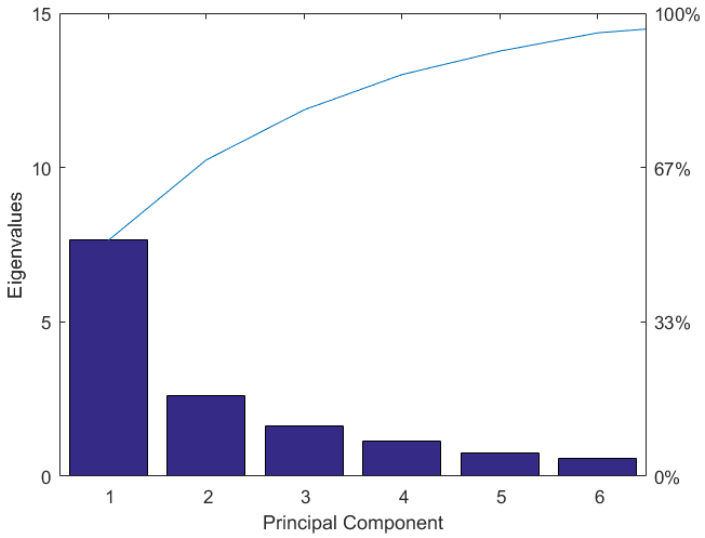


Figure 4.2: Scree Plot

Interpretation of Components

- 1^{st} **component**
  The first component holds the 50.9% of the information/variability. The variables that are statistical significant in this component are
  - GDP
  - Total Labor Force
  - Imports of goods and services
  - Investments
  - Number of Births
  - Inflation
  - Private Sector Debt

- 2^{nd} **component**
  The second component holds the 17.3% of the information. The variable that is statistical significant in this component is the *Median age of population.*

- 3^{rd} **component**
  The third component holds the 10.8% of the information. The variable that is statistical significant in this component is the *CAB.*

- 4^{th} **component**
  The forth component holds the 7.6% of the information. The variable that is statistical significant in this component is the *Unemployment Rate.*

- 5^{th} **component**
  The fifth component holds the 5% of the information. The variable that is statistical significant in this component is the *Demographic Dependency.*

- 6^{th} **component**
  The sixth component holds the 3.9% of the information. The variable that is statistical significant in this component is the *Compensation of Employees.*

- 7^{th} **component**
  The seventh component holds the 1.6% of the information. The variable that is statistical significant in this component is the *Investments.*

4.3.3 2011-2015 Dataset

The result of *PCA* are the following:

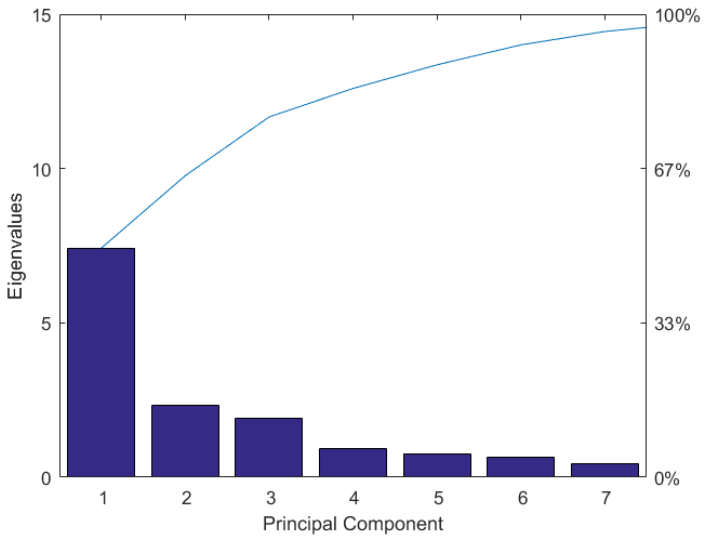


Figure 4.3: Scree Plot

Interpretation of Components

- 1^{st} **component**
  The first component holds the 49.4% of the information/variability. The variables that are statistical significant in this component are
  - GDP
  - Total Labor Force
  - Imports of goods and services
  - Investments
  - Number of Births
  - Net Number of Migrants
  - Inflation
  - Private Sector Debt
- 2^{nd} **component**
  The second component holds the 15.6% of the information. The variable that is statistical significant in this component is the *Unemployment Rate*.

- 3^{rd} **component**
  The third component holds the 12.6% of the information. The variable that is statistical significant in this component is the *CAB*.

- 4^{th} **component**
  The forth component holds the 6.1% of the information. The variable that is statistical significant in this component is the *CPI*.

- 5^{th} **component**
  The fifth component holds the 5% of the information. The variable that is statistical significant in this component is the *Compensation of Employees*.

- 6^{th} **component**
  The sixth component holds the 4.2% of the information. The variable that is statistical significant in this component is the *Dempographic Dependency*.

- 7^{th} **component**
  The seventh component holds the 2.8% of the information. The variable that is statistical significant in this component is the *Net Number of Migrants*.

4.3.4 Conclusion

Table 4.3.4 presents the results of *Principal Component Analysis* method for all three datasets examined.

Principal Component Analysis Synopsis			
Dataset	**2001-2005**	**2006-2010**	**2011-2015**
1^{st} **Component**	GDP Total Labor Force Imports of Goods and Services Investments Number of Births Inflation Private Sector Debt	GDP Total Labor Force Imports of Goods and Services Investments Number of Births Inflation Private Sector Debt	GDP Total Labor Force Imports of Goods and Services Investments Number of Births Inflation Private Sector Debt Net Number of Migrants
2^{nd} **Component**	Median Age of Population Long-term Interest Rates	Median Age of Population	Unemployment Rate
3^{rd} **Component**	CAB	CAB	CAB
4^{th} **Component**	Unemployment Rate	Unemployment Rate	CPI
5^{th} **Component**	Compensation of Employees	Demographic Dependency	Compensation of Employees
6^{th} **Component**	Investments	Compensation of Employees	Demographic Dependency
7^{th} **Component**	Compensation of Employees	Investments	Net Number of Migrants

Based on the above analysis, we observed that in all three datasets the variables that were significant in every component were almost always the same. It should be pointed out though that there is one important exception. Indeed, in the *C* period the *Number of Migrants* was significant both in the first Component as well as in the last one. This variable might have an impact in the modelling process that was not possibly not as important in the past as it is in this particular time period. This can be due to two very important events that have begun to emerge in Europe since 2010, the *European Debt Crisis* and *European Migrant Crisis*.

4.4 Linear Regression Analysis

In this section we proceed with the *multiple regression analysis* using the seven components of *PCA* from the previous section as independent variables and the logarithm of *Pension Expenditure* as the dependent variable. Our intention is to identify the significance of each component *(independent covariate)* and obtain an ideal model for *(the logarithm of)* the *Pension Expenditure* for descriptive as well as predictive purposes. The log transformation was decided to be used in order to achieve the linearity between the dependent variable and each independent one. Note that in cases linearity fails to exist, the *Kernel PCA* should be used instead of *PCA*.

Note: *For period A we have decided to remove 2 out of the original 7 independent variables while in periods B and C all 7 variables have been used. The details appear in the following sections.*

4.4.1 2001-2005 Dataset

The result of *Linear Regression* are the following:

```
Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:
                    Estimate         SE          tStat       pValue
                   ----------    ---------    ---------    ---------

    (Intercept)       0.87432      0.88044      0.99305      0.34029
    x1              0.0060387    0.0036301       1.6635      0.12208
    x2              -0.016255    0.0074334      -2.1867       0.0493
    x3              -0.014525     0.013117      -1.1074      0.28984
    x4              -0.019487    0.0087773      -2.2202     0.046423
    x5              -0.005766     0.013921      -0.4142      0.68603
    x6                0.03228     0.010483       3.0791    0.0095537
    x7             -0.00025197    0.0045291    -0.055633      0.95655
```

```
Number of observations: 20, Error degrees of freedom: 12
Root Mean Squared Error: 0.35
R-squared: 0.866,  Adjusted R-Squared 0.788
F-statistic vs. constant model: 11.1, p-value = 0.000215

ANOVA
                SumSq     DF    MeanSq       F         pValue
                ------    --    -------    ------    ----------

    Total       10.956    19    0.57665
    Model       9.4859     7     1.3551    11.059    0.00021478
    Residual    1.4704    12    0.12254
```

After consideration we decide to extract the $X5$ and $X7$ variables.

Then, the *Linear Regression* becomes as follow:

```
Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ---------    ---------    -------    ----------

    (Intercept)       1.1459      0.63003     1.8188      0.090389
    x1             0.0070111    0.0015125     4.6353    0.00038568
    x2             -0.017578    0.0038019    -4.6234    0.00039453
    x3             -0.018936    0.0041295    -4.5856    0.00042396
    x4             -0.021714    0.0047161    -4.6042    0.00040919
    x5              0.032552    0.0073719     4.4158    0.00058674


Number of observations: 20, Error degrees of freedom: 14
Root Mean Squared Error: 0.295
```

```
R-squared: 0.872,  Adjusted R-Squared 0.826
F-statistic vs. constant model: 19, p-value = 8.56e-06

ANOVA
               SumSq     DF    MeanSq       F         pValue
               ------    --    -------    ------    ----------

    Total      9.5141    19    0.50074
    Model      8.2926     5     1.6585    19.009    8.5633e-06
    Residual   1.2215    14    0.08725
```

4.4.2 2006-2010 Dataset

The result of *Linear Regression* are the following:

```
Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                   ----------    ---------    --------    ---------

    (Intercept)       -1.2595       1.3848    -0.90953      0.38098
    x1                0.010589    0.0030526      3.4689    0.0046407
    x2               -0.029168    0.0082644     -3.5294    0.0041511
    x3               -0.022032    0.0061539     -3.5802    0.0037806
    x4              -0.0086152    0.0025609     -3.3641     0.005631
    x5               -0.028823     0.007279     -3.9598    0.0018941
    x6                0.012096    0.0044256      2.7333     0.018156
    x7               -0.038647     0.010643     -3.6311    0.0034435


Number of observations: 20, Error degrees of freedom: 12
Root Mean Squared Error: 0.333
R-squared: 0.865,  Adjusted R-Squared 0.787
F-statistic vs. constant model: 11, p-value = 0.000219

ANOVA
                SumSq     DF    MeanSq       F          pValue
                ------    --    -------    ------    ----------

    Total        9.873    19    0.51963
    Model       8.5431     7     1.2204    11.013    0.00021921
    Residual    1.3298    12    0.11082
```

4.4.3 2011-2015 Dataset

The result of *Linear Regression* are the following:

```
Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:
                   Estimate          SE          tStat       pValue
                  ___________    __________    _________    ________

    (Intercept)         4.0111        1.3751       2.9169    0.012913
    x1             0.00014486    0.00060448      0.23965     0.81465
    x2             0.00024837     0.0073778     0.033665      0.9737
    x3            -0.00015032     0.0032925    -0.045656     0.96434
    x4             0.00074675     0.0034409      0.21702     0.83184
    x5             -0.0001515      0.012717    -0.011913     0.99069
    x6             0.00092311      0.010518     0.087764     0.93151
    x7            -0.00071601      0.016952    -0.042238       0.967


Number of observations: 20, Error degrees of freedom: 12
Root Mean Squared Error: 0.484
R-squared: 0.705,  Adjusted R-Squared 0.533
F-statistic vs. constant model: 4.1, p-value = 0.0159

ANOVA
                SumSq     DF    MeanSq       F         pValue
                ______    __    _______    ______    ________

    Total       9.5141    19    0.50074
    Model       6.7086     7    0.95837    4.0993    0.015891
    Residual    2.8055    12    0.23379
```

4.4.4 Overall Model

This regression results are between the *(logarithm of) Pension Expenditures* and the average of all observations by country and by variable of all three datasets.

```
Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                    Estimate          SE         tStat        pValue
                   -----------    ----------    --------    ----------

    (Intercept)         3.8466       0.17297      22.239    2.5401e-12
    x1              -1.8586e-07    3.3977e-07    -0.54701       0.59298
    x2              -2.5713e-07     5.127e-07    -0.50152        0.6238
    x3                1.119e-06    2.0533e-06     0.54496       0.59436
    x4              -1.0628e-06    1.9306e-06    -0.55053       0.59063
    x5               6.3235e-08    1.4799e-07     0.42729       0.67567


Number of observations: 20, Error degrees of freedom: 14
Root Mean Squared Error: 0.475
R-squared: 0.687,  Adjusted R-Squared 0.575
F-statistic vs. constant model: 6.13, p-value = 0.00328
ANOVA
                SumSq     DF    MeanSq       F       pValue
                ------    --    -------    -----    --------

    Total       10.086    19    0.53087
    Model       6.9249     5      1.385    6.133    0.003275
    Residual    3.1616    14    0.22583
```

4.4.5 Final Conclusions and Future Work

Table 4.4.5 presents the overall results for all regression models examined.

Linear Regression Synopsis								
	2001-2005		**2006-2010**		**2011-2015**		**Overall Model**	
	Test	**Graph**	**Test**	**Graph**	**Test**	**Graph**	**Test**	**Graph**
R^2	87%		86%		70%		68%	
Adj R^2	82%		78%		53%		57%	
F-statistic	Reject H_0		Reject H_0		Reject H_0		Reject H_0	
T-statistic Significance	All		All		None		None	
Normality Assumpion	✓*	✗	✓*	✗	✓**	✗	✓**	✗
Homoscedasticity Assumption		✓		✓				✓
Autocorrelation Assumption	✓	✓	✓	✓	✓	✓	✓	✓
Independence Assumption	✓		✓		✓		✓	
Linearity Assumption		✓		✓		✓		✓
Multicollinearity Assumption	High		High		High		High	

Notes:

- All the results were interpreted with $\alpha = 5\%$.
- ✓ means that the null hypothesis is failed to be rejected.
- ✗ means that the null hypothesis is being rejected.
- The existence of * means that we reject this null hypothesis in $\alpha = 5\%$, but the value is near 0.01 and if we choose $\alpha = 1\%$, then the null hypothesis can be failed to be rejected.
- The existence of ** means that we reject this null hypothesis in $\alpha = 5\%$ but in case of $\alpha = 1\%$, the null hypothesis would failed to be rejected.

We can observe that with the passing of time the same variables model is less and less good, that means that new factors with no big impact in the previous years, nowadays play an important role in the formation of the *Pension Expenditure*. As it was mentioned above, we believe that this two phenomena (*European Debt Crisis* and *European Migrant Crisis*) must be analyzed and being included as variables in the model of *Pension Expenditures*.

It can be seen that in the C model as well as in the *Overall* model the *F-test* comes to the conclusion that at least one variable is statistically significant while at the same time the *T-test* suggests that none of the variables is. Although that phenomenon seems quite paradoxical, it has thoroughly examined and explained by *Geary et al.* [**51**]. As he states there are two reason why this may occur.
The first one is the existance of multicollinearity in the model *(which in both cases is high)*, in which the existence of a relationship can be establislied but not the individual influence of each factor. The second reason why stems from the number of residuals degrees of freedom. If the residuals DE are $\geqslant$ 3, then, the significance point of F (k, n-k-1) is lower than the significance point of F (1, n-k-1) which corresponds to the significance point of t. Hence when all t_i are equal or approximately so, they may all be non-significant while F is significant. The explanation is that a significant F - ratio does not indicate the significance of any given regression coefficient but merely the existence of at least one linear combination which is significantly different from zero [**52**].

In conclusion, in this thesis, we suggest 3 models as well as an *Overall* model based on *Principal Component Analysis*. This work helps in achieving the following two tasks. Firstly by reducing the dimensionality of the original dataset we obtain a more "easy to use and handle" dataset and can apply various statistical methods and techniques without losing the accuracy of the original variables. Secondly, we are able to limit or eliminate the existing multicollinearity, and therefore we achieve a more accurate *Linear Regression* interpretation of the *Pension Expenditures* model.

Future work concerns deeper analysis of the particular model like comparison of our European model with other worldwide models, new proposals based on different alternative and more delicate and advanced dimension reduction techniques like *Support Vector Machines(SVM)*, also known as support vector network. As *Cortes et al.* [**57**] states, the *SVM* implements the following idea: *it maps the input vectors into some high dimensional feature space* $\mathbb{Z}$ *through some non-linear mapping chosen a priori. In this space a*

linear decision surface is constructed with special properties that ensure high generalization ability of the network. Another task that is also left for future work is the developement of a *LabSTADA* app into a statistical toolpak by adding more tests and techniques as options.

Appendices

Appendix A

Linear Regression

A.1 Descriptive statistics

Descriptive statistics provide simple summations, either quantitative or graphic, about the sample and the observations of a data set. Summary statistics is the most common quantitative summary and gives some measures i.e. measure of location, statistical dispersion, and shape, which are helpful to read large amount of information as simply as possible. Histogram, Box-plot and Stemplots are usually used as graphic representation.

In statistics and especially in research proccesses, descriptives are very important and usefull, as they present a first overall picture about the existing data set. The most common measures, that used to describe a data set, are measures of central tendency and variability [**1**].

A.1.1 Central tendency

Since 1920's, the term central tendency, also may referred to as *averages*, is used to describe the typical or the central value for a probability distribution. Most popular averages are the mean and the median.

Mean

The mean value of a group of observations, also knows as average or arithmetic mean, is defined by the summation of each and every observation divided by the total number of observations. Mean is typically denoted by $\bar{X}$ *(pronounced "X bar")* and defined by follows:

Let us suppose N, the number of observations and $x_1, x_2, ..., x_N$ the observed values.

Then,

$$\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N} \tag{A.1}$$

where,

$$\sum_{i=1}^{N} X_i = x_1 + x_2 + ... + x_N \tag{A.2}$$

Example 1

The coach of a basketball team, which consist of 12 players, wants to find the average weight of the players in order to conduct an experiment. The players weights(in kilograms) are the following: 65,60,73,79,67,81,85,74,69,90,86,73.

The weights summation is

$$\sum_{i=1}^{N} X_i = 65+60+73+79+67+81+85+74+69+90+86+73 = 902 \tag{A.3}$$

and the average weight of the team is

$$\bar{X} = \frac{902}{12} = 75.16kg \tag{A.4}$$

Median

Another important central tendency measure is the median, which is most commonly defined by the "middle" value separating the greater and lesser halves of a data set. Median is usually denotes as $\tilde{x}$, but in reality there is no widely accepted standard notation and so differs from author to author and always is defined while being introduced. First step to find the median in a given data set is to list the values in numerical order from smallest to largest

or vice versa. In a finite list of numbers the way to calculate the median depends on the number of values.

If the number of observations is odd, then after the arrangment, the middle value is chosen as the median. On the other hand, if the number of observations is even, then the median value is usually defined to be the mean of the two middle values.

Another way to find the median is through a stem-and-leaf display. Stem-and-leaf display also known as stem-and-leaf plot and Stemplot ,they first appeared in early 1900's by *Sir Arthur Lyon Bowley* and later evolved by *John Tukey* [**54**]. Stemplot as defined by business dictionary [**27**] *"is a table in which data values are divided into either a "leaf" or a "stem." In a stem and leaf plot, the stem values appear on the vertical axis and the leaf values are listed on the horizontal axis. Typically used to track the scores of a number of groups, such as those produced by sports teams or in a classroom setting, a stem and leaf plot clearly lists the scores of each group or individual in descending order."*

Example 2

Based on the data set of the previous example, the coach wants to calculate the median of the same team. The number of observations is N=12 , so even number. The arrangement of all the numbers from smallest to greatest is the following: 60, 65, 67, 69, 73, 73, 74, 79, 81, 85, 86, 90.

As it can be seen 73 and 74 are the two consecutive values, which averages will give the median value.

$$\tilde{x} = \frac{73 + 74}{2} = 73.5 \tag{A.5}$$

In the existing data set is easy to find the middle values and find the median, in a bigger data set it would be easier to locate the median via a stem and leaf plot. In the current example the steam and leaf plot will be presented as follows and someone can easily locate that 73 and 74 are the two middle consecutive values which will give the median.

```
Players_Weight Stem-and-Leaf Plot

 Frequency    Stem &  Leaf

     4,00        6 .  0579
     4,00        7 .  3349
     3,00        8 .  156
     1,00        9 .  0

 Stem width:     10,00
 Each leaf:        1 case(s)
```

A.1.2 Variability

Variability, also called dispersion, measures are used to determine how widely spread or how closely clustered the values of a data set are, and describes how much data sets differ from each other. Common measures of statistical variability include standard deviation, variance and range. The main feature of such a measure is that it is a nonnegative real number that is zero if all the data have the same value and increases as the data become more diverse.

Variance

Variance of a data set,let us suppose X, is defined by the expectation of the squared deviation of a random variable from its mean and usually is denoted by *var(X)* or σ^2 and can be expressed as follow:

$$Var(X) = E(X - E(X))^2 = E(X^2) - E(X)^2 \tag{A.6}$$

A small variance shows that the data set is tightly clustered together and a large number means the values are more spread apart. Variance is one of the most important measures in the field of statistics and it can be applied in many tasks such as descriptive statistics, statistical inference, hypothesis testing, goodness of fit, and Monte Carlo sampling.

Standard Deviation

Standard deviation, denoted by *SD* or σ, is the square root of the variance and is used to quantify the amount of variation of the values in the given data set. A small number for the *SD* means that the observations tend to the average price and a high number for the *SD* indicates that the observations of the data set are spread out over a wider range of values.

Range

In statistics, range is defined by the difference between the maximum value and the minimum value of a data set. Sample maximum and minimum value are the largest and the smallest observation respectively. Most common are denoted by *Max* and *Min*. Someone can very easily identify those values by sorting the given data set. Box plot is an altarnative graphic way to find the minimum and maximum values. Since the mathematician John W. Tukey introduced this type of visual data display, known as Box and Whisker Plot, in 1969. From this point to this day, the number of scientists using visual represantations to display their work is growing exponentially. A Box Plot is defined by the visual representation of the statistical five number summary of a given data set. This summary includes the minimum and maximum observatios, the first and third quartile and the median which is basically the second quartile.

Example 3
Given the data set from the example 1, the variance of the players height is

$$Var(X) = \sigma^2 = 84.69 \tag{A.7}$$

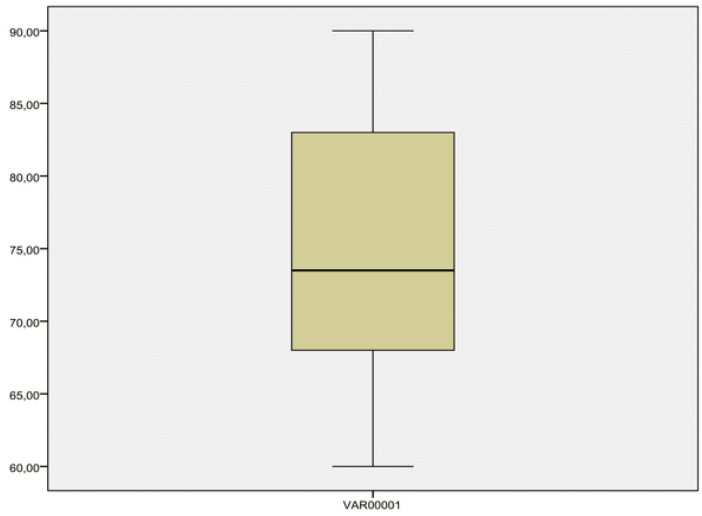
and the standard deviation of the players height is

$$\sigma = 9.20 \tag{A.8}$$

As for the range, it's very easy to see that Range is

$$Range = Max \text{ - } Min = 90 \text{ - } 60 = 30 \tag{A.9}$$

and visually via a Box plot



A.2 Analysis of Variance

The analysis of variance, most commonly known as ANOVA, is a statistical technique developed by *R.A. Fisher* [**55**, **56**], where he introduced the terms *variance* and *analysis of variance*. Is used for the comparison of two or more populations, while different kinds of effects operating at the same time. Basically is a procedure that determine if those effects are important, which their estimation is, and if there are differences between the means of populations [**39**].

An ANOVA table is formed as follows:

Source	Sum of Squares	DF	Mean Squared	F	Sig.
Explained	ESS	k-1	MSE = ESS / k-1	F = MSE / MSR	
Residuals	RSS	n-k	MSR = RSS / n-k		
Total	TSS	n-1			

Note: *The symbolism may vary from case to case and it must always be clarified because in some cases the same symbol may be misinterpreted by the reader.*

A.2.1 ANOVA interpretation

The goal of this procedure is to find if there are differences between the means as it mentioned above. The first step to achieve this is to define a hypothesis test. Hypothesis testing in statistics is a way to test the results of a survey or experiment and determine if the results are valid and meaningful [**42-43**]. The elements of a hypothesis test are the following:

1. **Null and alternative hypotheses**.
   The null hypothesis, denoted by H_0, is always the accepted fact while the alternative, denoted by H_α, is the one that is questionable and must be examined.

2. **Predetermined level of significance**
   The significance level is defined by the "tolerance" that the conductor of the experiment is giving in the existence of error type I.

   Note: *When conducting a statistical experiment, the risk of incorrect decision making may occur. There are two types of errors that may happen.*

   - Type I error: reject a true H_0
   - Type II error: failing to reject a false H_0

   Let us define,

$$\alpha = P(type\ I\ error) \tag{A.10}$$

   and

$$\beta = P(type\ II\ error) \tag{A.11}$$

   Then, one of the following options about the experiment decision occurred

Options	Fail to reject null hypothesis	Reject null hypothesis
H_0 **is true**	- with probability = 1-α	Type I error with probability = α
H_α **is false**	Type II error with probability = β	- with probability = 1-β

3. **Test statistic and critical zone of the test**
   Is used when it's time for ruling about the rejection or not of null hypothesis. It is a random quantity with known distribution when H_0 is true. It's formed based on the H_α, the distribution of the test statistic and the significance level α. The formation of the critical zone is defined by the H_α

4. **Value of the test statistic**
   The value of the test is computed based on the given values of the sample and if it exists in the critical zone, the null hypothesis is rejected otherwise is failed to be rejected.

5. **The decision to reject or not the null hypothesis**
   The conductor of the experiment is rulling about the rejection or not of the null hypothesis

Test Hypothesis

After said the above, the AVONA table null hypothesis is used to decide if there are any differences between the means. The null and alternatives are form as:

$$H_0 : \mu_1 = \mu_2 .... = \mu_n \tag{A.12}$$

$$H_\alpha : \textit{At least one } \mu_i \textit{ differs from the others} \tag{A.13}$$

where *i=1,2,...,n.*

ESS

Explained Sum of Squares, symbolised as *ESS*, expresses the variability between the sampling means, which measured as the sum of the squares of

the distances of each medium with the total mean.

$$ESS = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 \tag{A.14}$$

RSS

Residual Sum of Squares, symbolized by *RSS*, expresses the variability between the sampling means, which measured as the sum of the squares of the distances of each medium with the total mean.

$$RSS = \sum_{i=1}^{n} \hat{\varepsilon}_i^2 \tag{A.15}$$

TSS

Total Sum of Squares, symbolized by *TSS*, expresses the overall variability of the observations

$$\begin{aligned}
&\overline{\sum \text{height}}_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (Y_i - \overline{Y} + \hat{Y}_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} ((\hat{Y}_i - \bar{Y}) + \underbrace{(Y_i - \hat{Y}_i)}_{\hat{\varepsilon}_i})^2 \\
&= \sum_{i=1}^{n} ((\hat{Y}_i - \bar{Y})^2 + 2\hat{\varepsilon}_i(\hat{Y}_i - \bar{Y}) + \hat{\varepsilon}_i^2) \\
&= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n} \hat{\varepsilon}_i(\hat{Y}_i - \bar{Y}) \\
&= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2\sum_{i=1}^{n} \hat{\varepsilon}_i(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} - \overline{Y}) \\
&= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 + 2(\hat{\beta}_0 - \overline{Y}) \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i}_{0} + \\
&+ 2\hat{\beta}_1 \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i x_{i1}}_{0} + \cdots + 2\hat{\beta}_p \underbrace{\sum_{i=1}^{n} \hat{\varepsilon}_i x_{ip}}_{0} \\
&= \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n} \hat{\varepsilon}_i^2 = \text{ESS} + \text{RSS}
\end{aligned}$$

DF

Degrees of freedom, denoted by *DF*, of an estimate is the number of independent pieces of information that went into calculating the estimate and in particular:

- **k-1** are the *DF* of the divergence from the H_0
- **n-k** are the *DF* of residuals
- **n-1** are the total *DF*

MSE

Explained mean square, denoted by *MSR*, is defined by the error between the sample.

$$\text{MSE} = \frac{1}{k-1}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{A.16}$$

MSR

Mean square of residuals, denoted by *MSE*, is defined by the error within the sample.

$$\text{MSR} = \frac{1}{n-k}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{A.17}$$

F-test

As it mentioned above, F-test is the ratio of the variation between the sample and the variation within the sample. This particular test is the one that ANOVA use to check the test hypothesis.

If the test fails to reject the H_0, then,

$$F^* \sim F_{(k-1),(n-k)} \tag{A.18}$$

Otherwise,

$$F^* > F_{(k-1),(n-k),\alpha} \tag{A.19}$$

Sig.

Sig. in ANOVA table stands as the significant value of F-test. Is most commonly known as the *p-value*, which is defined by the probability to observe a random price of test statistics, same or even more extreme from the observed one in terms of H_0, given the fact that H_0 is failed to be rejected. Basically it is the lowest significance level which the H_0 is being rejected.

Decision cases about the hypothesis test

- If *p-value* $< \alpha$, then the H_0 is rejected
- If *p-value* $> \alpha$, then the H_0 is not rejected
- If *p-value* $= \alpha$, then no decision about the rejection or not of the H_0 can be made

Appendix B

Actuarial Project Results

B.1 Principal Components Analysis

B.1.1 2001-2005

Eigenvalues Matrix

Columns 1 through 7

7.8102	0	0	0	0	0	0
0	2.5581	0	0	0	0	0
0	0	1.4153	0	0	0	0
0	0	0	0.9465	0	0	0
0	0	0	0	0.7665	0	0
0	0	0	0	0	0.5826	0
0	0	0	0	0	0	0.3828
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Columns 8 through 15

0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

```
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
    0.2183         0         0         0         0         0         0         0
         0    0.1487         0         0         0         0         0         0
         0         0    0.0777         0         0         0         0         0
         0         0         0    0.0587         0         0         0         0
         0         0         0         0    0.0159         0         0         0
         0         0         0         0         0    0.0096         0         0
         0         0         0         0         0         0    0.0056         0
         0         0         0         0         0         0         0    0.0035
```

Eigenvectors Matrix

Columns 1 through 7

```
   -0.3407   -0.1371   -0.1320   -0.0665   -0.1000   -0.0309   -0.0006
    0.0404   -0.3153    0.0981    0.8403    0.0676   -0.1302    0.2578
   -0.3237   -0.2268   -0.1176    0.0934   -0.0047   -0.0285    0.0917
   -0.3364   -0.0780   -0.2010   -0.0615   -0.0822   -0.0575   -0.0791
   -0.0608    0.2778   -0.6319    0.2529   -0.1555    0.4007   -0.1288
   -0.2863   -0.1083    0.1110    0.1713    0.2096    0.5686   -0.3536
   -0.1862    0.3503   -0.1533    0.1807    0.5110   -0.3929   -0.0745
   -0.2105    0.4397    0.1433    0.1175   -0.0989    0.0898    0.4122
   -0.3213   -0.2395   -0.0773    0.0663    0.0152   -0.0530   -0.0431
   -0.2310   -0.1542    0.4306   -0.1634    0.4291    0.2593    0.0373
   -0.2440    0.2898    0.2617    0.0226   -0.2820    0.2831    0.3908
   -0.3168   -0.0748   -0.1179   -0.1860   -0.0336   -0.2415    0.3643
    0.1779   -0.4670   -0.1150   -0.0687   -0.2678    0.1615    0.2351
   -0.3314   -0.1442   -0.1312   -0.1725   -0.0580   -0.1506   -0.0501
   -0.2046    0.0986    0.3985    0.1911   -0.5473   -0.2729   -0.5053
```

Columns 8 through 15

```
    0.0056   -0.1387   -0.0318   -0.0324   -0.2140   -0.5711   -0.0942    0.6558
   -0.1831   -0.1394    0.1116    0.1003   -0.1396   -0.0114   -0.0064   -0.0096
    0.1307    0.0042   -0.2553   -0.0224    0.7418   -0.3166   -0.0099   -0.2805
   -0.1102   -0.0792   -0.1384    0.6190   -0.0200    0.4030   -0.4857    0.0296
   -0.3147    0.2014    0.1297    0.0498    0.0804    0.0043    0.2950    0.0463
    0.1850    0.0273    0.0334   -0.2996   -0.2339   -0.0315   -0.3660   -0.2385
    0.3852    0.4026    0.0877    0.1554   -0.1065   -0.1058    0.0217   -0.0067
   -0.0942    0.0658   -0.6744   -0.1550   -0.2103    0.0515   -0.0056   -0.0373
    0.3060   -0.0733   -0.0728   -0.3112    0.0707    0.6009    0.3781    0.3376
```

```
   -0.4130    0.2726    0.0099    0.2805    0.0899   -0.0028    0.3245    0.1446
    0.3844   -0.1169    0.4910    0.2449    0.0938   -0.0008    0.0617    0.0063
   -0.3593    0.2850    0.3944   -0.4301   -0.0375    0.1043   -0.2670   -0.1375
    0.3053    0.6143   -0.1395    0.1633   -0.2370   -0.0450    0.0439   -0.0498
   -0.0190   -0.3017    0.0009    0.1324   -0.4322   -0.1460    0.4514   -0.5226
   -0.1099    0.3231    0.0189   -0.0127    0.0328   -0.0304    0.0669   -0.0194

Extracted Eigenvalues Matrix

    7.8102         0         0         0         0         0         0
         0    2.5581         0         0         0         0         0
         0         0    1.4153         0         0         0         0
         0         0         0    0.9465         0         0         0
         0         0         0         0    0.7665         0         0
         0         0         0         0         0    0.5826         0
         0         0         0         0         0         0    0.3828

Extracted Eigenvectors Matrix

   -0.3407   -0.1371   -0.1320   -0.0665   -0.1000   -0.0309   -0.0006
    0.0404   -0.3153    0.0981    0.8403    0.0676   -0.1302    0.2578
   -0.3237   -0.2268   -0.1176    0.0934   -0.0047   -0.0285    0.0917
   -0.3364   -0.0780   -0.2010   -0.0615   -0.0822   -0.0575   -0.0791
   -0.0608    0.2778   -0.6319    0.2529   -0.1555    0.4007   -0.1288
   -0.2863   -0.1083    0.1110    0.1713    0.2096    0.5686   -0.3536
   -0.1862    0.3503   -0.1533    0.1807    0.5110   -0.3929   -0.0745
   -0.2105    0.4397    0.1433    0.1175   -0.0989    0.0898    0.4122
   -0.3213   -0.2395   -0.0773    0.0663    0.0152   -0.0530   -0.0431
   -0.2310   -0.1542    0.4306   -0.1634    0.4291    0.2593    0.0373
   -0.2440    0.2898    0.2617    0.0226   -0.2820    0.2831    0.3908
   -0.3168   -0.0748   -0.1179   -0.1860   -0.0336   -0.2415    0.3643
    0.1779   -0.4670   -0.1150   -0.0687   -0.2678    0.1615    0.2351
   -0.3314   -0.1442   -0.1312   -0.1725   -0.0580   -0.1506   -0.0501
   -0.2046    0.0986    0.3985    0.1911   -0.5473   -0.2729   -0.5053

The Components Matrix becomes as follows
    0.9521    0.2193    0.1571    0.0647    0.0875    0.0236    0.0004
   -0.1130    0.5043   -0.1167   -0.8175   -0.0592    0.0994   -0.1595
    0.9045    0.3627    0.1399   -0.0909    0.0041    0.0218   -0.0567
    0.9401    0.1247    0.2392    0.0598    0.0720    0.0439    0.0489
    0.1699   -0.4442    0.7517   -0.2461    0.1361   -0.3058    0.0797
    0.8001    0.1732   -0.1320   -0.1667   -0.1835   -0.4340    0.2188
    0.5204   -0.5603    0.1824   -0.1758   -0.4474    0.2998    0.0461
    0.5881   -0.7033   -0.1705   -0.1143    0.0866   -0.0685   -0.2550
    0.8978    0.3831    0.0920   -0.0645   -0.0133    0.0405    0.0267
```

```
    0.6457    0.2467   -0.5123    0.1589   -0.3757   -0.1979   -0.0231
    0.6819   -0.4635   -0.3114   -0.0220    0.2469   -0.2161   -0.2418
    0.8853    0.1196    0.1403    0.1810    0.0294    0.1843   -0.2254
   -0.4971    0.7469    0.1369    0.0668    0.2345   -0.1233   -0.1454
    0.9261    0.2307    0.1561    0.1679    0.0508    0.1149    0.0310
    0.5717   -0.1577   -0.4741   -0.1859    0.4791    0.2083    0.3126

Number of Selected PCA Variables
     7

New Variables via PCA
   1.0e+11 *

    1.4003    0.1858    0.3562    0.0891    0.1072    0.0654    0.0729
    2.1562    0.2861    0.5486    0.1373    0.1650    0.1007    0.1123
    0.5110    0.0678    0.1300    0.0325    0.0391    0.0239    0.0266
    0.8087    0.1073    0.2057    0.0515    0.0619    0.0378    0.0421
    0.4970    0.0659    0.1264    0.0316    0.0380    0.0232    0.0259
    4.4392    0.5890    1.1293    0.2826    0.3398    0.2074    0.2311
    6.9473    0.9218    1.7674    0.4422    0.5318    0.3246    0.3617
    0.4884    0.0648    0.1243    0.0311    0.0374    0.0228    0.0254
    0.0427    0.0057    0.0109    0.0027    0.0033    0.0020    0.0022
    3.4441    0.4570    0.8762    0.2192    0.2636    0.1609    0.1793
    0.0547    0.0073    0.0139    0.0035    0.0042    0.0026    0.0028
    2.8474    0.3778    0.7244    0.1812    0.2179    0.1330    0.1482
    0.7572    0.1005    0.1926    0.0482    0.0580    0.0354    0.0394
    0.5393    0.0716    0.1372    0.0343    0.0413    0.0252    0.0281
    0.2632    0.0349    0.0670    0.0168    0.0201    0.0123    0.0137
    0.1533    0.0203    0.0390    0.0098    0.0117    0.0072    0.0080
    2.4369    0.3234    0.6199    0.1551    0.1865    0.1138    0.1269
    1.0778    0.1430    0.2742    0.0686    0.0825    0.0503    0.0561
    1.4213    0.1886    0.3616    0.0905    0.1088    0.0664    0.0740
    5.1715    0.6862    1.3156    0.3292    0.3958    0.2416    0.2692

Original Data Correlation Matrix

  Columns 1 through 7

    1.0000   -0.0684    0.9569    0.9700    0.1660    0.7427    0.3504
   -0.0684    1.0000    0.1404   -0.1193   -0.1714    0.0725   -0.1919
    0.9569    0.1404    1.0000    0.9193    0.0989    0.7632    0.3194
    0.9700   -0.1193    0.9193    1.0000    0.2745    0.6955    0.4259
    0.1660   -0.1714    0.0989    0.2745    1.0000    0.1126    0.3556
    0.7427    0.0725    0.7632    0.6955    0.1126    1.0000    0.3017
    0.3504   -0.1919    0.3194    0.4259    0.3556    0.3017    1.0000
```

```
    0.3783   -0.2830    0.2851    0.4117    0.3256    0.3466    0.6084
    0.9485    0.1209    0.9782    0.8963    0.0312    0.7898    0.3173
    0.5543   -0.0001    0.5716    0.5038   -0.3784    0.7287    0.1682
    0.5145   -0.2614    0.4191    0.4994    0.1495    0.5172    0.4060
    0.9026   -0.1421    0.8463    0.8773    0.1234    0.5417    0.4047
   -0.2775    0.3347   -0.1502   -0.3405   -0.2755   -0.3066   -0.7492
    0.9816   -0.1315    0.9237    0.9640    0.0979    0.6767    0.3473
    0.4634    0.0034    0.3965    0.4501   -0.0995    0.4107    0.2047

  Columns 8 through 15

    0.3783    0.9485    0.5543    0.5145    0.9026   -0.2775    0.9816    0.4634
   -0.2830    0.1209   -0.0001   -0.2614   -0.1421    0.3347   -0.1315    0.0034
    0.2851    0.9782    0.5716    0.4191    0.8463   -0.1502    0.9237    0.3965
    0.4117    0.8963    0.5038    0.4994    0.8773   -0.3405    0.9640    0.4501
    0.3256    0.0312   -0.3784    0.1495    0.1234   -0.2755    0.0979   -0.0995
    0.3466    0.7898    0.7287    0.5172    0.5417   -0.3066    0.6767    0.4107
    0.6084    0.3173    0.1682    0.4060    0.4047   -0.7492    0.3473    0.2047
    1.0000    0.2392    0.2701    0.8431    0.4330   -0.7765    0.3229    0.5012
    0.2392    1.0000    0.5784    0.4089    0.8213   -0.1529    0.9267    0.4209
    0.2701    0.5784    1.0000    0.4028    0.5523   -0.2569    0.5506    0.3381
    0.8431    0.4089    0.4028    1.0000    0.4962   -0.5979    0.4576    0.5981
    0.4330    0.8213    0.5523    0.4962    1.0000   -0.3089    0.9005    0.3923
   -0.7765   -0.1529   -0.2569   -0.5979   -0.3089    1.0000   -0.2880   -0.4164
    0.3229    0.9267    0.5506    0.4576    0.9005   -0.2880    1.0000    0.4317
    0.5012    0.4209    0.3381    0.5981    0.3923   -0.4164    0.4317    1.0000


TOTAL VARIANCE EXPLAINED
Initial Eigenvalues
    Total_Initial_Eigenvalues    Percentage_Variance    Cumulative_Variance
    -------------------------    -------------------    -------------------

       7.8102                       0.52068             0.52068
       2.5581                       0.17054             0.69122
       1.4153                      0.094356             0.78558
      0.94654                      0.063103             0.84868
      0.76651                        0.0511             0.89978
      0.58256                      0.038837             0.93862
       0.3828                       0.02552             0.96414
       0.2183                      0.014553             0.97869
      0.14867                     0.0099111              0.9886
      0.07771                     0.0051807             0.99378
     0.058736                     0.0039157              0.9977
     0.015896                     0.0010597             0.99876
```

```
     0.009557                  0.00063713             0.99939
    0.0055601                  0.00037067             0.99976
    0.0035341                  0.00023561                   1

Extraction Sums of Squared Loadings
    New_Total_Initial_Eigenvalues    New_Percentage_Variance    New_Cumulative_Variance
    -----------------------------    -----------------------    -----------------------

     7.8102                            0.52068                   0.52068
     2.5581                            0.17054                   0.69122
     1.4153                           0.094356                   0.78558
    0.94654                           0.063103                   0.84868
    0.76651                             0.0511                   0.89978
    0.58256                           0.038837                   0.93862
     0.3828                            0.02552                   0.96414

Original Data Covariance Matrix

   1.0e+22 *

  Columns 1 through 7

    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
   -0.0000    0.0000    0.0000   -0.0000   -0.0000    0.0000   -0.0000
    0.0000    0.0000    0.0000    0.0002    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0002    4.1636    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
   -0.0000    0.0000   -0.0000   -0.0000   -0.0000   -0.0000   -0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000

  Columns 8 through 15

    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
   -0.0000    0.0000   -0.0000   -0.0000   -0.0000    0.0000   -0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000   -0.0000    0.0000    0.0000   -0.0000    0.0000   -0.0000
```

```
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
   -0.0000   -0.0000   -0.0000   -0.0000   -0.0000    0.0000   -0.0000   -0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
```

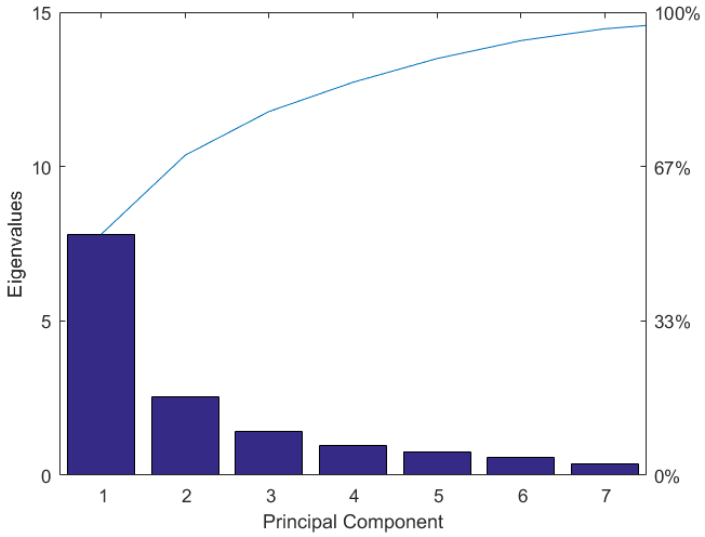


Figure B.1: Scree Plot

B.1.2 2006-2010

```
Eigenvalues Matrix
  Columns 1 through 7

    7.6403         0         0         0         0         0         0
         0    2.6049         0         0         0         0         0
         0         0    1.6241         0         0         0         0
         0         0         0    1.1416         0         0         0
         0         0         0         0    0.7605         0         0
         0         0         0         0         0    0.5915         0
         0         0         0         0         0         0    0.2483
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0

  Columns 8 through 15

         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
    0.1797         0         0         0         0         0         0         0
         0    0.0966         0         0         0         0         0         0
         0         0    0.0572         0         0         0         0         0
         0         0         0    0.0217         0         0         0         0
         0         0         0         0    0.0162         0         0         0
         0         0         0         0         0    0.0109         0         0
         0         0         0         0         0         0    0.0043         0
         0         0         0         0         0         0         0    0.0022



Eigenvectors Matrix
  Columns 1 through 7

   -0.3505   -0.0855   -0.1013   -0.0599   -0.0488    0.1423   -0.0203
   -0.0327   -0.2370    0.4132   -0.5123    0.4906   -0.2970   -0.3564
```

```
   -0.3389   -0.1455   -0.0746   -0.1099    0.0767    0.1084   -0.0458
   -0.3397    0.0004   -0.1799   -0.0619    0.0262    0.2429   -0.1422
   -0.0480    0.3675   -0.5189   -0.3053    0.2015    0.1049   -0.2194
   -0.3002   -0.0573    0.0193   -0.2858    0.1177   -0.1613    0.8156
   -0.1713    0.4242    0.2845    0.1840    0.3979   -0.0360   -0.0047
   -0.1895    0.4420    0.2072   -0.1968   -0.3033   -0.0610   -0.0306
   -0.3295   -0.2103   -0.0101   -0.0084    0.0559    0.1194    0.0690
   -0.2099   -0.2702    0.3595    0.3682   -0.1908   -0.0830   -0.1241
   -0.2102    0.3117    0.1814   -0.2440   -0.5227   -0.3167   -0.1105
   -0.3413   -0.0631   -0.0589    0.1059   -0.0280    0.0335   -0.2974
    0.1741   -0.4003   -0.1535   -0.4258   -0.3475   -0.0607   -0.0847
   -0.3450   -0.1372   -0.0182    0.0701   -0.0650    0.1484   -0.0634
   -0.1597   -0.0648   -0.4449    0.2792    0.0843   -0.7940   -0.0627

  Columns 8 through 15

    0.0650    0.0096    0.1154   -0.1117   -0.3256   -0.4040    0.7161    0.1440
   -0.0653   -0.0085    0.0626   -0.1309   -0.1594   -0.0002   -0.0479    0.0222
    0.2286   -0.3841   -0.2654   -0.1642    0.4689   -0.1893   -0.0074   -0.5231
   -0.0931   -0.1902    0.4190   -0.0468   -0.0413    0.7276    0.0957   -0.0754
   -0.4426   -0.0640   -0.0620    0.0567    0.2064   -0.2386   -0.1288    0.2779
   -0.2947    0.1347   -0.0680    0.0261    0.0175    0.0859   -0.0080   -0.0365
    0.1760    0.0934    0.1201    0.5716    0.2786   -0.0007    0.2396   -0.0204
    0.0437   -0.4303   -0.3670    0.1883   -0.4637    0.0737   -0.1112   -0.0214
    0.4748   -0.0568   -0.1135    0.0129    0.1138    0.0437   -0.2919    0.6951
   -0.5931   -0.2959   -0.0422    0.0822    0.2301   -0.0605    0.0484    0.2293
    0.0942    0.2751    0.3008   -0.2672    0.3586   -0.0500   -0.0143    0.0525
   -0.1104    0.6230   -0.5553    0.0254   -0.0624    0.2188    0.0219   -0.1136
    0.0533    0.0036   -0.0249    0.6407    0.1513    0.0716    0.1757   -0.0068
   -0.0492    0.1564    0.4044    0.2865   -0.2767   -0.3733   -0.5198   -0.2608
    0.1068   -0.1448    0.0390    0.0596   -0.1165    0.0367    0.0040   -0.0236


Extracted Eigenvalues Matrix

    7.6403         0         0         0         0         0         0
         0    2.6049         0         0         0         0         0
         0         0    1.6241         0         0         0         0
         0         0         0    1.1416         0         0         0
         0         0         0         0    0.7605         0         0
         0         0         0         0         0    0.5915         0
         0         0         0         0         0         0    0.2483

Extracted Eigenvectors Matrix
```

```
   -0.3505   -0.0855   -0.1013   -0.0599   -0.0488    0.1423   -0.0203
   -0.0327   -0.2370    0.4132   -0.5123    0.4906   -0.2970   -0.3564
   -0.3389   -0.1455   -0.0746   -0.1099    0.0767    0.1084   -0.0458
   -0.3397    0.0004   -0.1799   -0.0619    0.0262    0.2429   -0.1422
   -0.0480    0.3675   -0.5189   -0.3053    0.2015    0.1049   -0.2194
   -0.3002   -0.0573    0.0193   -0.2858    0.1177   -0.1613    0.8156
   -0.1713    0.4242    0.2845    0.1840    0.3979   -0.0360   -0.0047
   -0.1895    0.4420    0.2072   -0.1968   -0.3033   -0.0610   -0.0306
   -0.3295   -0.2103   -0.0101   -0.0084    0.0559    0.1194    0.0690
   -0.2099   -0.2702    0.3595    0.3682   -0.1908   -0.0830   -0.1241
   -0.2102    0.3117    0.1814   -0.2440   -0.5227   -0.3167   -0.1105
   -0.3413   -0.0631   -0.0589    0.1059   -0.0280    0.0335   -0.2974
    0.1741   -0.4003   -0.1535   -0.4258   -0.3475   -0.0607   -0.0847
   -0.3450   -0.1372   -0.0182    0.0701   -0.0650    0.1484   -0.0634
   -0.1597   -0.0648   -0.4449    0.2792    0.0843   -0.7940   -0.0627

The Components Matrix becomes as follows

    0.9689    0.1380    0.1291    0.0641    0.0426   -0.1094    0.0101
    0.0903    0.3825   -0.5266    0.5474   -0.4279    0.2284    0.1776
    0.9366    0.2348    0.0951    0.1174   -0.0669   -0.0833    0.0228
    0.9389   -0.0006    0.2292    0.0662   -0.0229   -0.1868    0.0709
    0.1327   -0.5932    0.6613    0.3262   -0.1757   -0.0807    0.1093
    0.8299    0.0924   -0.0246    0.3054   -0.1027    0.1240   -0.4065
    0.4735   -0.6847   -0.3626   -0.1966   -0.3470    0.0277    0.0023
    0.5239   -0.7134   -0.2640    0.2103    0.2645    0.0470    0.0152
    0.9107    0.3395    0.0129    0.0089   -0.0488   -0.0918   -0.0344
    0.5802    0.4360   -0.4582   -0.3934    0.1664    0.0639    0.0618
    0.5810   -0.5031   -0.2312    0.2607    0.4558    0.2436    0.0550
    0.9434    0.1019    0.0751   -0.1131    0.0244   -0.0258    0.1482
   -0.4811    0.6461    0.1956    0.4550    0.3030    0.0467    0.0422
    0.9537    0.2214    0.0232   -0.0749    0.0566   -0.1141    0.0316
    0.4415    0.1046    0.5669   -0.2983   -0.0735    0.6107    0.0313

Number of Selected PCA Variables
     7

New Variables via PCA
   1.0e+12 *

    0.1740   -0.0001    0.0425    0.0123   -0.0042   -0.0346    0.0131
    0.3315   -0.0002    0.0809    0.0234   -0.0081   -0.0660    0.0250
    0.1144   -0.0001    0.0279    0.0081   -0.0028   -0.0228    0.0086
    0.1400   -0.0001    0.0342    0.0099   -0.0034   -0.0279    0.0106
    0.0903   -0.0001    0.0220    0.0064   -0.0022   -0.0180    0.0068
```

```
    0.7020   -0.0004    0.1714    0.0495   -0.0171   -0.1397    0.0530
    1.1529   -0.0007    0.2815    0.0812   -0.0281   -0.2294    0.0870
    0.0874   -0.0001    0.0213    0.0062   -0.0021   -0.0174    0.0066
    0.0068   -0.0000    0.0017    0.0005   -0.0002   -0.0014    0.0005
    0.5438   -0.0003    0.1328    0.0383   -0.0132   -0.1082    0.0410
    0.0139   -0.0000    0.0034    0.0010   -0.0003   -0.0028    0.0010
    0.4837   -0.0003    0.1181    0.0341   -0.0118   -0.0963    0.0365
    0.1732   -0.0001    0.0423    0.0122   -0.0042   -0.0345    0.0131
    0.0849   -0.0001    0.0207    0.0060   -0.0021   -0.0169    0.0064
    0.0649   -0.0000    0.0158    0.0046   -0.0016   -0.0129    0.0049
    0.0293   -0.0000    0.0071    0.0021   -0.0007   -0.0058    0.0022
    0.3949   -0.0002    0.0964    0.0278   -0.0096   -0.0786    0.0298
    0.1807   -0.0001    0.0441    0.0127   -0.0044   -0.0360    0.0136
    0.2477   -0.0002    0.0605    0.0175   -0.0060   -0.0493    0.0187
    0.7366   -0.0005    0.1798    0.0519   -0.0179   -0.1466    0.0556

Original Data Correlation Matrix

  Columns 1 through 7

    1.0000    0.0666    0.9653    0.9626    0.1494    0.8068    0.2879
    0.0666    1.0000    0.1981   -0.0173   -0.3040    0.2935    0.0153
    0.9653    0.1981    1.0000    0.9229    0.0939    0.8036    0.2481
    0.9626   -0.0173    0.9229    1.0000    0.3299    0.7456    0.3493
    0.1494   -0.3040    0.0939    0.3299    1.0000    0.1250    0.2106
    0.8068    0.2935    0.8036    0.7456    0.1250    1.0000    0.3085
    0.2879    0.0153    0.2481    0.3493    0.2106    0.3085    1.0000
    0.3938   -0.0719    0.3205    0.4310    0.3371    0.4056    0.6956
    0.9422    0.1978    0.9683    0.8646   -0.0930    0.7717    0.2207
    0.5303    0.2057    0.5296    0.4157   -0.5903    0.4064    0.1442
    0.4760   -0.0055    0.3780    0.4583    0.2048    0.4794    0.5067
    0.9309    0.0316    0.8852    0.8891    0.0976    0.7063    0.3614
   -0.3161    0.2353   -0.2489   -0.3915   -0.2250   -0.2501   -0.9239
    0.9710    0.0748    0.9291    0.9224   -0.0091    0.7585    0.2880
    0.4283   -0.2063    0.4198    0.4166    0.2331    0.3340    0.0355

  Columns 8 through 15

    0.3938    0.9422    0.5303    0.4760    0.9309   -0.3161    0.9710    0.4283
   -0.0719    0.1978    0.2057   -0.0055    0.0316    0.2353    0.0748   -0.2063
    0.3205    0.9683    0.5296    0.3780    0.8852   -0.2489    0.9291    0.4198
    0.4310    0.8646    0.4157    0.4583    0.8891   -0.3915    0.9224    0.4166
    0.3371   -0.0930   -0.5903    0.2048    0.0976   -0.2250   -0.0091    0.2331
    0.4056    0.7717    0.4064    0.4794    0.7063   -0.2501    0.7585    0.3340
    0.6956    0.2207    0.1442    0.5067    0.3614   -0.9239    0.2880    0.0355
```

1.0000	0.2235	0.0861	0.8912	0.3711	-0.5838	0.3177	-0.0389
0.2235	1.0000	0.6029	0.3164	0.8803	-0.2278	0.9419	0.3981
0.0861	0.6029	1.0000	0.1982	0.6084	-0.2140	0.6721	0.1803
0.8912	0.3164	0.1982	1.0000	0.4675	-0.3817	0.4249	0.1096
0.3711	0.8803	0.6084	0.4675	1.0000	-0.4120	0.9386	0.4788
-0.5838	-0.2278	-0.2140	-0.3817	-0.4120	1.0000	-0.3306	-0.1606
0.3177	0.9419	0.6721	0.4249	0.9386	-0.3306	1.0000	0.4054
-0.0389	0.3981	0.1803	0.1096	0.4788	-0.1606	0.4054	1.0000

TOTAL VARIANCE EXPLAINED

Initial Eigenvalues

Total_Initial_Eigenvalues	Percentage_Variance	Cumulative_Variance
7.6403	0.50935	0.50935
2.6049	0.17366	0.68302
1.6241	0.10827	0.79129
1.1416	0.076108	0.8674
0.76047	0.050698	0.91809
0.59146	0.039431	0.95753
0.24835	0.016557	0.97408
0.1797	0.01198	0.98606
0.096565	0.0064377	0.9925
0.057219	0.0038146	0.99631
0.021671	0.0014447	0.99776
0.016161	0.0010774	0.99884
0.010906	0.00072704	0.99956
0.0043493	0.00028995	0.99985
0.002203	0.00014686	1

Extraction Sums of Squared Loadings

New_Total_Initial_Eigenvalues	New_Percentage_Variance	New_Cumulative_Variance
7.6403	0.50935	0.50935
2.6049	0.17366	0.68302
1.6241	0.10827	0.79129
1.1416	0.076108	0.8674
0.76047	0.050698	0.91809
0.59146	0.039431	0.95753
0.24835	0.016557	0.97408

```
Original Data Covariance Matrix

   1.0e+23 *

  Columns 1 through 7

    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000   -0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    1.0271    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000    0.0000
   -0.0000    0.0000   -0.0000   -0.0000   -0.0000   -0.0000   -0.0000
    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000   -0.0000    0.0000    0.0000    0.0000    0.0000    0.0000

  Columns 8 through 15

    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
   -0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000    0.0000   -0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000   -0.0000   -0.0000    0.0000    0.0000   -0.0000   -0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000   -0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
   -0.0000   -0.0000   -0.0000   -0.0000   -0.0000    0.0000   -0.0000   -0.0000
    0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
   -0.0000    0.0000    0.0000    0.0000    0.0000   -0.0000    0.0000    0.0000
```

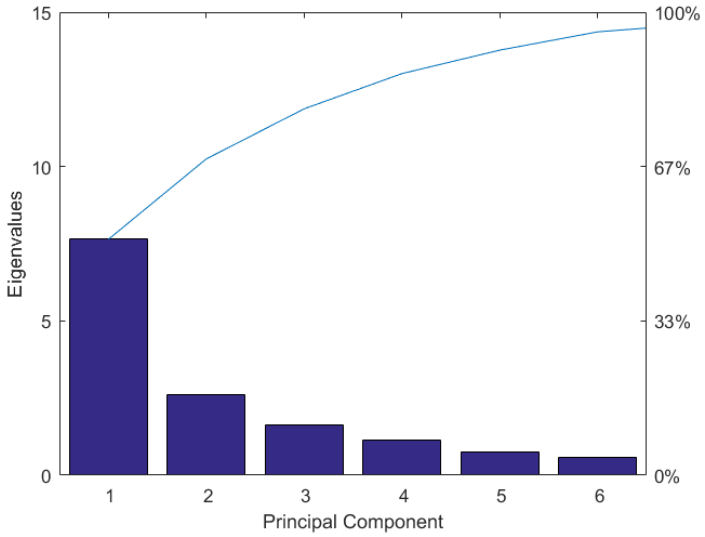


Figure B.2: Scree Plot

B.1.3 2011-2015

```
Eigenvalues Matrix

  Columns 1 through 7

    7.4229         0         0         0         0         0         0
         0    2.3454         0         0         0         0         0
         0         0    1.9027         0         0         0         0
         0         0         0    0.9233         0         0         0
         0         0         0         0    0.7648         0         0
         0         0         0         0         0    0.6461         0
         0         0         0         0         0         0    0.4345
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0
         0         0         0         0         0         0         0

  Columns 8 through 15

         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
         0         0         0         0         0         0         0         0
    0.2630         0         0         0         0         0         0         0
         0    0.1396         0         0         0         0         0         0
         0         0    0.0809         0         0         0         0         0
         0         0         0    0.0460         0         0         0         0
         0         0         0         0    0.0260         0         0         0
         0         0         0         0         0    0.0027         0         0
         0         0         0         0         0         0    0.0015         0
         0         0         0         0         0         0         0    0.0008


Eigenvectors Matrix

  Columns 1 through 7

   -0.3606   -0.0062   -0.0908    0.1122   -0.0458    0.0205   -0.0520
```

```
    0.0643   -0.5335   -0.2994    0.0708    0.2276   -0.0708    0.1532
   -0.3407   -0.0400   -0.1801    0.1721    0.0212    0.0218    0.0519
   -0.3515    0.0562    0.0358    0.2029   -0.0197   -0.0745   -0.0674
   -0.1348    0.0177    0.5147    0.4134    0.4410   -0.1774    0.1873
   -0.2983   -0.1198   -0.0850   -0.0900    0.2821   -0.3582    0.3565
    0.0697    0.2983   -0.4208    0.4838   -0.1111    0.3618    0.4429
   -0.1452   -0.4260    0.3608    0.1794   -0.2713    0.1318    0.1915
   -0.3321   -0.0189   -0.2793    0.0268   -0.0683    0.0068   -0.0176
   -0.2774    0.2215    0.2014    0.2787   -0.0623    0.1561   -0.4998
   -0.1743   -0.3792    0.2173   -0.2204   -0.1912    0.5758    0.1389
   -0.3229   -0.0264   -0.1443   -0.3580   -0.0555   -0.1922   -0.1491
    0.1307   -0.4003   -0.2665    0.2721    0.3924    0.1961   -0.5177
   -0.3492   -0.0082   -0.1722   -0.0512   -0.1346   -0.0118   -0.0338
   -0.1794    0.2668    0.0414   -0.3668    0.6041    0.4990    0.1039

  Columns 8 through 15

   -0.0428   -0.0314    0.0174    0.1194    0.2772    0.1391   -0.8478   -0.1072
   -0.3265   -0.4430    0.4724    0.0191   -0.0342   -0.0652    0.0095    0.0666
   -0.3316    0.1360   -0.1743   -0.4239    0.2056   -0.1495    0.2495   -0.5927
   -0.0732   -0.2323   -0.2569    0.2721   -0.6527   -0.4361   -0.0595   -0.0360
    0.0679   -0.2756   -0.2093   -0.0635    0.3178    0.0400    0.0945    0.2113
    0.4955    0.3145    0.2246   -0.1092   -0.2968    0.1768   -0.0185   -0.1408
    0.2825   -0.0184    0.0981    0.1225    0.1141   -0.1588    0.0707    0.0869
   -0.1897    0.5341    0.1557    0.3457    0.0800   -0.1335    0.0892    0.1013
   -0.2248    0.1799   -0.1755   -0.3856   -0.1033    0.1561    0.0283    0.7123
    0.1108   -0.0446    0.6333   -0.1720   -0.0816    0.0796    0.1408   -0.0005
    0.3482   -0.3194   -0.1413   -0.2964   -0.0855    0.0408   -0.0259   -0.0432
    0.2894   -0.0846    0.0670    0.1276    0.4599   -0.5512    0.1584    0.1780
    0.2829    0.2345   -0.2589    0.1189    0.0206   -0.0023    0.0231   -0.0082
   -0.0146   -0.2224   -0.1356    0.4981    0.0879    0.5907    0.3822   -0.0863
   -0.2494    0.1480    0.1019    0.1910   -0.0299   -0.0535    0.0083    0.0287


Extracted Eigenvalues Matrix

    7.4229         0         0         0         0         0         0
         0    2.3454         0         0         0         0         0
         0         0    1.9027         0         0         0         0
         0         0         0    0.9233         0         0         0
         0         0         0         0    0.7648         0         0
         0         0         0         0         0    0.6461         0
         0         0         0         0         0         0    0.4345

Extracted Eigenvectors Matrix
```

```
   -0.3606   -0.0062   -0.0908    0.1122   -0.0458    0.0205   -0.0520
    0.0643   -0.5335   -0.2994    0.0708    0.2276   -0.0708    0.1532
   -0.3407   -0.0400   -0.1801    0.1721    0.0212    0.0218    0.0519
   -0.3515    0.0562    0.0358    0.2029   -0.0197   -0.0745   -0.0674
   -0.1348    0.0177    0.5147    0.4134    0.4410   -0.1774    0.1873
   -0.2983   -0.1198   -0.0850   -0.0900    0.2821   -0.3582    0.3565
    0.0697    0.2983   -0.4208    0.4838   -0.1111    0.3618    0.4429
   -0.1452   -0.4260    0.3608    0.1794   -0.2713    0.1318    0.1915
   -0.3321   -0.0189   -0.2793    0.0268   -0.0683    0.0068   -0.0176
   -0.2774    0.2215    0.2014    0.2787   -0.0623    0.1561   -0.4998
   -0.1743   -0.3792    0.2173   -0.2204   -0.1912    0.5758    0.1389
   -0.3229   -0.0264   -0.1443   -0.3580   -0.0555   -0.1922   -0.1491
    0.1307   -0.4003   -0.2665    0.2721    0.3924    0.1961   -0.5177
   -0.3492   -0.0082   -0.1722   -0.0512   -0.1346   -0.0118   -0.0338
   -0.1794    0.2668    0.0414   -0.3668    0.6041    0.4990    0.1039

The Components Matrix becomes as follows

    0.9824    0.0096    0.1252   -0.1078    0.0401   -0.0165    0.0343
   -0.1751    0.8171    0.4130   -0.0680   -0.1990    0.0569   -0.1010
    0.9282    0.0613    0.2484   -0.1654   -0.0185   -0.0175   -0.0342
    0.9576   -0.0860   -0.0493   -0.1950    0.0172    0.0599    0.0444
    0.3671   -0.0271   -0.7100   -0.3972   -0.3857    0.1426   -0.1235
    0.8129    0.1835    0.1172    0.0865   -0.2467    0.2879   -0.2350
   -0.1899   -0.4569    0.5804   -0.4649    0.0971   -0.2908   -0.2919
    0.3957    0.6524   -0.4977   -0.1724    0.2372   -0.1060   -0.1262
    0.9047    0.0289    0.3852   -0.0257    0.0598   -0.0055    0.0116
    0.7558   -0.3392   -0.2778   -0.2678    0.0545   -0.1254    0.3295
    0.4748    0.5808   -0.2997    0.2118    0.1672   -0.4628   -0.0915
    0.8798    0.0404    0.1991    0.3440    0.0485    0.1544    0.0983
   -0.3560    0.6131    0.3677   -0.2614   -0.3431   -0.1577    0.3412
    0.9515    0.0126    0.2375    0.0492    0.1177    0.0095    0.0223
    0.4888   -0.4085   -0.0571    0.3525   -0.5283   -0.4011   -0.0685

Number of Selected PCA Variables
     7

New Variables via PCA
   1.0e+12 *

    0.2020   -0.0181   -0.0104   -0.0411    0.0036    0.0126    0.0094
    0.3952   -0.0355   -0.0204   -0.0805    0.0071    0.0247    0.0183
    0.1435   -0.0129   -0.0074   -0.0292    0.0026    0.0090    0.0067
    0.1536   -0.0138   -0.0079   -0.0313    0.0028    0.0096    0.0071
```

```
    0.0977   -0.0088   -0.0050   -0.0199    0.0018    0.0061    0.0045
    0.8118   -0.0729   -0.0418   -0.1653    0.0146    0.0507    0.0376
    1.3824   -0.1242   -0.0712   -0.2815    0.0249    0.0864    0.0641
    0.0718   -0.0064   -0.0037   -0.0146    0.0013    0.0045    0.0033
    0.0072   -0.0006   -0.0004   -0.0015    0.0001    0.0004    0.0003
    0.5469   -0.0491   -0.0282   -0.1114    0.0098    0.0342    0.0254
    0.0175   -0.0016   -0.0009   -0.0036    0.0003    0.0011    0.0008
    0.5779   -0.0519   -0.0298   -0.1177    0.0104    0.0361    0.0268
    0.2237   -0.0201   -0.0115   -0.0456    0.0040    0.0140    0.0104
    0.0832   -0.0075   -0.0043   -0.0169    0.0015    0.0052    0.0039
    0.0803   -0.0072   -0.0041   -0.0164    0.0014    0.0050    0.0037
    0.0314   -0.0028   -0.0016   -0.0064    0.0006    0.0020    0.0015
    0.3836   -0.0345   -0.0197   -0.0781    0.0069    0.0240    0.0178
    0.2125   -0.0191   -0.0109   -0.0433    0.0038    0.0133    0.0099
    0.3656   -0.0328   -0.0188   -0.0744    0.0066    0.0229    0.0170
    0.8249   -0.0741   -0.0425   -0.1680    0.0148    0.0516    0.0383

Original Data Correlation Matrix

  Columns 1 through 7

    1.0000   -0.1114    0.9615    0.9542    0.2943    0.7737   -0.0711
   -0.1114    1.0000    0.0203   -0.2379   -0.2524    0.0861   -0.0945
    0.9615    0.0203    1.0000    0.8977    0.2320    0.7483    0.0020
    0.9542   -0.2379    0.8977    1.0000    0.4685    0.7216   -0.1163
    0.2943   -0.2524    0.2320    0.4685    1.0000    0.3319   -0.3230
    0.7737    0.0861    0.7483    0.7216    0.3319    1.0000   -0.2132
   -0.0711   -0.0945    0.0020   -0.1163   -0.3230   -0.2132    1.0000
    0.3605    0.2190    0.3321    0.3593    0.4315    0.3079   -0.5032
    0.9416    0.0139    0.9722    0.8500    0.0353    0.7427    0.0335
    0.7471   -0.5398    0.6295    0.8122    0.5035    0.3950   -0.0676
    0.4180    0.1871    0.3459    0.3536    0.1891    0.3537   -0.4268
    0.8561   -0.0902    0.7771    0.7668    0.0458    0.8158   -0.2742
   -0.2729    0.7091   -0.1797   -0.3676   -0.2366   -0.1556    0.0544
    0.9687   -0.0766    0.9214    0.9067    0.1240    0.7593   -0.0680
    0.4174   -0.3613    0.3947    0.4020    0.2335    0.3522   -0.0349

  Columns 8 through 15

    0.3605    0.9416    0.7471    0.4180    0.8561   -0.2729    0.9687    0.4174
    0.2190    0.0139   -0.5398    0.1871   -0.0902    0.7091   -0.0766   -0.3613
    0.3321    0.9722    0.6295    0.3459    0.7771   -0.1797    0.9214    0.3947
    0.3593    0.8500    0.8122    0.3536    0.7668   -0.3676    0.9067    0.4020
    0.4315    0.0353    0.5035    0.1891    0.0458   -0.2366    0.1240    0.2335
    0.3079    0.7427    0.3950    0.3537    0.8158   -0.1556    0.7593    0.3522
```

-0.5032	0.0335	-0.0676	-0.4268	-0.2742	0.0544	-0.0680	-0.0349
1.0000	0.2189	0.2430	0.7318	0.1821	0.0155	0.2725	-0.1519
0.2189	1.0000	0.5683	0.3158	0.8443	-0.1811	0.9470	0.3829
0.2430	0.5683	1.0000	0.2329	0.5298	-0.4031	0.6384	0.4240
0.7318	0.3158	0.2329	1.0000	0.4087	0.0226	0.4148	0.1570
0.1821	0.8443	0.5298	0.4087	1.0000	-0.2943	0.9151	0.4098
0.0155	-0.1811	-0.4031	0.0226	-0.2943	1.0000	-0.2936	-0.3313
0.2725	0.9470	0.6384	0.4148	0.9151	-0.2936	1.0000	0.3957
-0.1519	0.3829	0.4240	0.1570	0.4098	-0.3313	0.3957	1.0000

TOTAL VARIANCE EXPLAINED

Initial Eigenvalues

Total_Initial_Eigenvalues	Percentage_Variance	Cumulative_Variance
7.4229	0.49486	0.49486
2.3454	0.15636	0.65122
1.9027	0.12684	0.77806
0.92328	0.061552	0.83961
0.76478	0.050985	0.8906
0.64606	0.04307	0.93367
0.4345	0.028967	0.96264
0.26296	0.017531	0.98017
0.13959	0.0093057	0.98947
0.080934	0.0053956	0.99487
0.046006	0.0030671	0.99794
0.025988	0.0017325	0.99967
0.0027124	0.00018083	0.99985
0.0014972	9.9813e-05	0.99995
0.00075952	5.0635e-05	1

Extraction Sums of Squared Loadings

New_Total_Initial_Eigenvalues	New_Percentage_Variance	New_Cumulative_Variance
7.4229	0.49486	0.49486
2.3454	0.15636	0.65122
1.9027	0.12684	0.77806
0.92328	0.061552	0.83961
0.76478	0.050985	0.8906
0.64606	0.04307	0.93367
0.4345	0.028967	0.96264

Original Data Covariance Matrix

1.0e+23 *

Columns 1 through 7

0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
-0.0000	0.0000	0.0000	-0.0000	-0.0000	0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	-0.0000	0.0000	1.3451	0.0000	0.0000	-0.0000
0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
-0.0000	-0.0000	0.0000	-0.0000	-0.0000	-0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
-0.0000	0.0000	-0.0000	-0.0000	-0.0000	-0.0000	0.0000
0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
0.0000	-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000

Columns 8 through 15

0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
0.0000	0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
-0.0000	0.0000	-0.0000	-0.0000	-0.0000	0.0000	-0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
0.0000	-0.0000	-0.0000	0.0000	-0.0000	0.0000	-0.0000	-0.0000
0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000
-0.0000	0.0000	0.0000	0.0000	0.0000	-0.0000	0.0000	0.0000

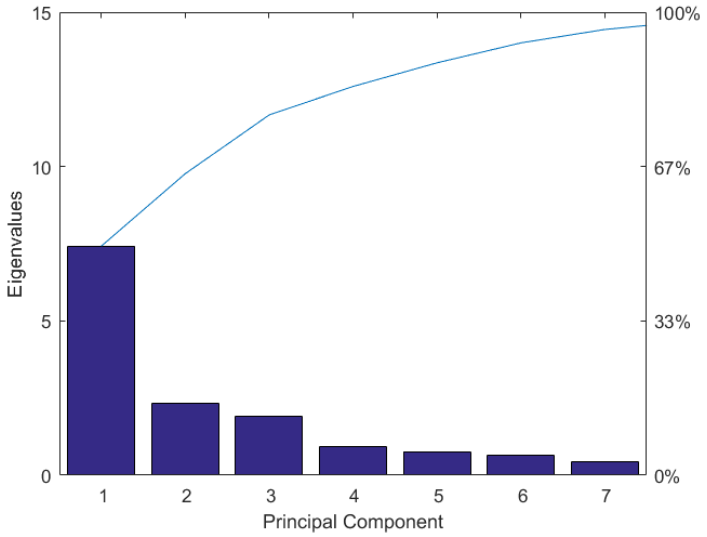


Figure B.3: Scree Plot

B.2 Linear Regression

B.2.1 2001-2005

Descriptive_Statistics

Maximum_ Value	Minimum_ Value	Average_ Value	Median_ Value	Standard_ Deviation	Variance
5.4617	2.8511	4.3491	4.4736	0.75937	0.57665
6.9473e+11	4.2651e+09	1.7729e+11	9.4322e+10	1.9183e+11	3.6798e+22
9.2181e+10	5.6591e+08	2.3524e+10	1.2515e+10	2.5453e+10	6.4786e+20
1.7674e+11	1.0851e+09	4.5102e+10	2.3996e+10	4.8801e+10	2.3816e+21
4.4219e+10	2.7149e+08	1.1284e+10	6.0039e+09	1.221e+10	1.4908e+20
5.3176e+10	3.2646e+08	1.357e+10	7.2197e+09	1.4683e+10	2.1559e+20
3.2456e+10	1.9926e+08	8.2823e+09	4.4065e+09	8.9617e+09	8.0311e+19
3.6168e+10	2.2205e+08	9.2296e+09	4.9106e+09	9.9867e+09	9.9735e+19

First row corresponds to the dependent variable followed by the independent ones, in order

Linear_Model

Linear regression model:
y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	0.87432	0.88044	0.99305	0.34029
x1	0.0060387	0.0036301	1.6635	0.12208
x2	-0.016255	0.0074334	-2.1867	0.0493
x3	-0.014525	0.013117	-1.1074	0.28984
x4	-0.019487	0.0087773	-2.2202	0.046423
x5	-0.005766	0.013921	-0.4142	0.68603
x6	0.03228	0.010483	3.0791	0.0095537
x7	-0.00025197	0.0045291	-0.055633	0.95655

Number of observations: 20, Error degrees of freedom: 12
Root Mean Squared Error: 0.35
R-squared: 0.866, Adjusted R-Squared 0.788
F-statistic vs. constant model: 11.1, p-value = 0.000215

```
ANOVA
                SumSq     DF    MeanSq       F         pValue
                ------    --    -------    ------    ----------

    Total       10.956    19    0.57665
    Model       9.4859     7     1.3551    11.059    0.00021478
    Residual    1.4704    12    0.12254

Coefficient_Confidence_Intervals

   -1.0440    2.7926
   -0.0019    0.0139
   -0.0325   -0.0001
   -0.0431    0.0141
   -0.0386   -0.0004
   -0.0361    0.0246
    0.0094    0.0551
   -0.0101    0.0096


Kolmogorov-Smirnov Normality test
Test Hypothesis:
Ho: The Residuals come from a standard normal distribution
Ha: The Residuals do not come from such a distribution
Kolm_Smirnov_pvalue
    0.0125

The test rejects the null hypothesis at the chosen significance level


Durbin Watson autocorrelation test
Test Hypothesis:
Ho: The residuals are not autocorrelated
Ha: The residuals are autocorrelated

Durbin_Watson_pvalue =

    0.5858

The test fail to reject the null hypothesis at the Assumptions
Significance Level


Runs test for independence
```

```
Test Hypothesis:
Ho: The values of the residuals come in random order
Ha: The values of the residuals do not come in random order
Runs_Test_pvalue
    0.7899

The test fail to reject the null hypothesis at the Assumptions
Significance Level


Collinearity test

VIF:

   1.0e+16 *

    0.5205    0.0112    1.3170   -0.2379    0.7684    0.1537   -0.1595

The VIF expresses the rate at which the estimator's variance increases
when collinearity exist

If VIF > 10 then multicollinearity is high

Variance Decomposition

 sValue   var1    var2    var3    var4    var5    var6    var7    var8
-------------------------------------------------------------------------
 2.7637   0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.6014   0.0103  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0001   0.0001  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0000   0.0000  0.0000  0.0001  0.0000  0.0003  0.0001  0.0002  0.0000
 0.0000   0.0004  0.0001  0.0001  0.0001  0.0032  0.0002  0.0145  0.0233
 0.0000   0.0049  0.0003  0.0030  0.0006  0.0292  0.0041  0.0080  0.9045
 0.0000   0.5949  0.0131  0.1411  0.0204  0.0669  0.4434  0.5082  0.0203
 0.0000   0.3894  0.9866  0.8557  0.9789  0.9004  0.5523  0.4691  0.0520

Condition_Indices

   1.0e+10 *

    0.0000
    0.0000
    0.0000
    0.0171
    0.0872
```

```
    0.2345
    0.9729
    4.3467

If Condition Index < 10, multicollinearity is small
If 10 <= Condition Index <= 30, multicollinearity is medium
If Condition Index > 30, multicollinearity is high


Graphs Explanation

For Normality:
An equal distribution of the residuals around their median suggests the
existence of normal distribution

For Homoscedasticity:
The increase in the variance as the fitted values increase suggests
possible heteroscedasticity
On the contrary, the decrease in the variance as the fitted values
increase suggests possible homoscedasticity

For Autocorrelation:
A trend among the residuals indicates a possible correlation between them

For Independence
If the data are normally distributed and uncorrelated then they are
independent

For Linearity
If the data do not exhibit a pattern, then linearity could be accepted
```

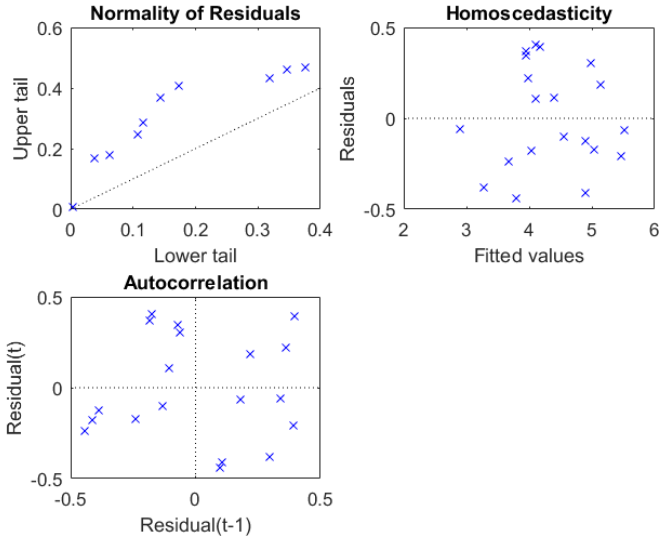


Figure B.4: Assumptions

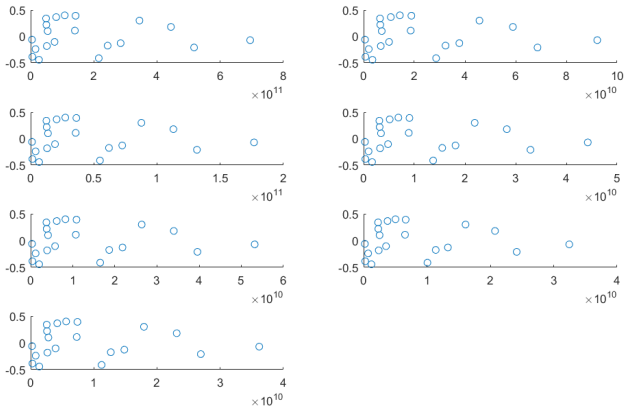


Figure B.5: Linearity assumption

After the extraction of *X5* and *X7*, the results are:

```
Descriptive_Statistics

 Maximum_     Minimum_       Average_      Median_        Standard_     Variance
Value         Value          Value         Value          Deviation
----------    -----------    ---------    ------------    ----------    ----------

 5.5292        3.0269         4.5573        4.6681         0.70763       0.50074
 6.9473e+11    4.2651e+09     1.7729e+11    9.4322e+10     1.9183e+11    3.6798e+22
 9.2181e+10    5.6591e+08     2.3524e+10    1.2515e+10     2.5453e+10    6.4786e+20
 1.7674e+11    1.0851e+09     4.5102e+10    2.3996e+10     4.8801e+10    2.3816e+21
 4.4219e+10    2.7149e+08     1.1284e+10    6.0039e+09     1.221e+10     1.4908e+20
 3.2456e+10    1.9926e+08     8.2823e+09    4.4065e+09     8.9617e+09    8.0311e+19

First row corresponds to the dependent variable followed by the independent ones, in order.

Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                   Estimate        SE         tStat       pValue
                   ---------    ---------    -------    ----------

    (Intercept)       1.1459      0.63003     1.8188      0.090389
    x1             0.0070111    0.0015125     4.6353    0.00038568
    x2             -0.017578    0.0038019    -4.6234    0.00039453
    x3             -0.018936    0.0041295    -4.5856    0.00042396
    x4             -0.021714    0.0047161    -4.6042    0.00040919
    x5              0.032552    0.0073719     4.4158    0.00058674


Number of observations: 20, Error degrees of freedom: 14
Root Mean Squared Error: 0.295
R-squared: 0.872,  Adjusted R-Squared 0.826
F-statistic vs. constant model: 19, p-value = 8.56e-06

ANOVA
                SumSq     DF    MeanSq       F          pValue
                ------    --    -------    ------    ----------

    Total       9.5141    19    0.50074
    Model       8.2926     5     1.6585    19.009    8.5633e-06
    Residual    1.2215    14    0.08725
```

```
Coefficient_Confidence_Intervals
   -0.2054    2.4972
    0.0038    0.0103
   -0.0257   -0.0094
   -0.0278   -0.0101
   -0.0318   -0.0116
    0.0167    0.0484


Kolmogorov-Smirnov Normality test
Test Hypothesis:
Ho: The Residuals come from a standard normal distribution
Ha: The Residuals do not come from such a distribution
Kolm_Smirnov_pvalue
    0.0189

The test rejects the null hypothesis at the chosen significance level

Durbin Watson autocorrelation test
Test Hypothesis:
Ho: The residuals are not autocorrelated
Ha: The residuals are autocorrelated

Durbin_Watson_pvalue =

    0.3825

The test fail to reject the null hypothesis at the Assumptions Significance Level
Runs test for independence
Test Hypothesis:
Ho: The values of the residuals come in random order
Ha: The values of the residuals do not come in random order
Runs_Test_pvalue
     1

The test fail to reject the null hypothesis at the chosen significance level.

Collinearity test

VIF:

  1.0e+15 *

    2.9062    0.7369    2.1223    0.9621   -0.5903
```

```
The VIF expresses the rate at which the estimator's variance increases
when collinearity exist.
If VIF > 10 then multicollinearity is high

Variance Decomposition

 sValue      condIdx      var1    var2    var3    var4    var5    var6
------------------------------------------------------------------------
 2.3740       1          0.0001  0.0000  0.0000  0.0000  0.0000  0.0000
 0.6033      3.9348      0.0086  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0001  45851.9044      0.0001  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0000      1.8665e+08  0.0001  0.0000  0.0001  0.0000  0.0005  0.0007
 0.0000      8.4759e+08  0.0009  0.0003  0.0010  0.0012  0.0016  0.0101
 0.0000      2.8145e+10  0.9902  0.9997  0.9989  0.9988  0.9980  0.9892

Condition_Indices

   1.0e+10 *

    0.0000
    0.0000
    0.0000
    0.0187
    0.0848
    2.8145

If Condition Index < 10, multicollinearity is small
If 10 <= Condition Index <= 30, multicollinearity is medium
If Condition Index > 30, multicollinearity is high


Graphs Explanation

For Normality:
An equal distribution of the residuals around their median suggests
the existence of normal distribution

For Homoscedasticity:
The increase in the variance as the fitted values increase suggests
possible heteroscedasticity
On the contrary, the decrease in the variance as the fitted values
increase suggests possible homoscedasticity

For Autocorrelation:
```

```
A trend among the residuals indicates a possible correlation between them

For Independence
If the data are normally distributed and uncorrelated then they are independent

For Linearity
If the data do not exhibit a pattern, then linearity could be accepted
```

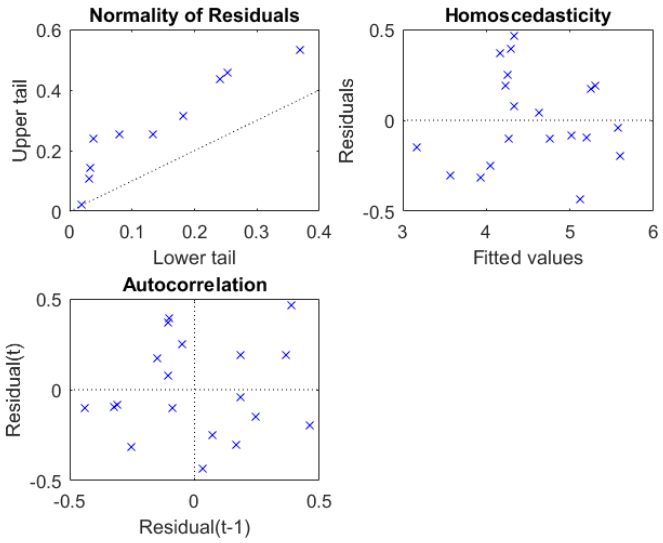


Figure B.6: Assumptions

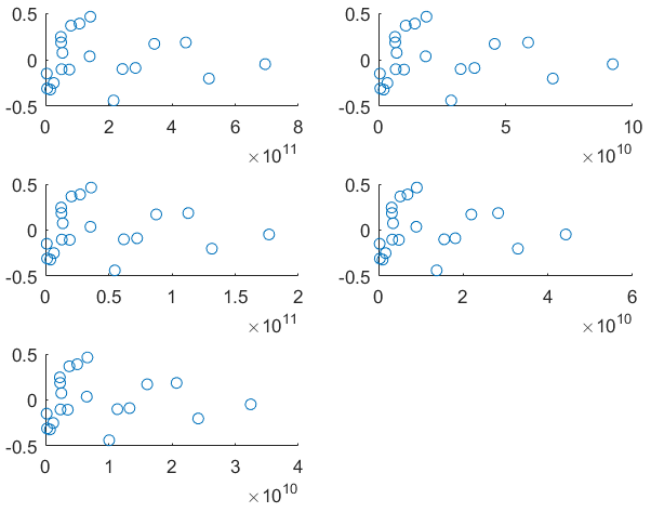


Figure B.7: Linearity assumption

B.2.2 2006-2010

Descriptive_Statistics

Maximum_ Value	Minimum_ Value	Average_ Value	Median_ Value	Standard_ Deviation	Variance
5.4896	2.9187	4.47	4.5881	0.72085	0.51963
1.1529e+12	6.8372e+09	2.8764e+11	1.7362e+11	3.0093e+11	9.0559e+22
-4.2125e+06	-7.2116e+08	-1.7975e+08	-1.0751e+08	1.8815e+08	3.5401e+16
2.8146e+11	1.6695e+09	7.0222e+10	4.2386e+10	7.3465e+10	5.3971e+21
8.1236e+10	4.8151e+08	2.0268e+10	1.2234e+10	2.1204e+10	4.4961e+20
-1.6657e+08	-2.8081e+10	-7.006e+09	-4.2291e+09	7.3296e+09	5.3722e+19
-1.36e+09	-2.2942e+11	-5.7238e+10	-3.4549e+10	5.9882e+10	3.5858e+21
8.7003e+10	5.1596e+08	2.1707e+10	1.3102e+10	2.2709e+10	5.1571e+20

First row corresponds to the dependent variable followed by the independent ones, in order.

Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-1.2595	1.3848	-0.90953	0.38098
x1	0.010589	0.0030526	3.4689	0.0046407
x2	-0.029168	0.0082644	-3.5294	0.0041511
x3	-0.022032	0.0061539	-3.5802	0.0037806
x4	-0.0086152	0.0025609	-3.3641	0.005631
x5	-0.028823	0.007279	-3.9598	0.0018941
x6	0.012096	0.0044256	2.7333	0.018156
x7	-0.038647	0.010643	-3.6311	0.0034435

Number of observations: 20, Error degrees of freedom: 12
Root Mean Squared Error: 0.333
R-squared: 0.865, Adjusted R-Squared 0.787
F-statistic vs. constant model: 11, p-value = 0.000219

ANOVA

	SumSq	DF	MeanSq	F	pValue

```
                ------    --    -------    ------    ----------

    Total        9.873    19    0.51963
    Model       8.5431     7     1.2204    11.013    0.00021921
    Residual    1.3298    12    0.11082

Coefficient_Confidence_Intervals

   -4.2768    1.7577
    0.0039    0.0172
   -0.0472   -0.0112
   -0.0354   -0.0086
   -0.0142   -0.0030
   -0.0447   -0.0130
    0.0025    0.0217
   -0.0618   -0.0155


Kolmogorov-Smirnov Normality test
Test Hypothesis:
Ho: The Residuals come from a standard normal distribution
Ha: The Residuals do not come from such a distribution
Kolm_Smirnov_pvalue
    0.0096

The test rejects the null hypothesis at the chosen significance level


Durbin Watson autocorrelation test
Test Hypothesis:
Ho: The residuals are not autocorrelated
Ha: The residuals are autocorrelated

Durbin_Watson_pvalue =

    0.9483

The test fail to reject the null hypothesis at the
Assumptions Significance Level


Runs test for independence
Test Hypothesis:
Ho: The values of the residuals come in random order
Ha: The values of the residuals do not come in random order
```

```
Runs_Test_pvalue
     1

The test fail to reject the null hypothesis at the chosen significance level


Collinearity test

VIF:
   1.0e+15 *

    1.4912    0.0001   -0.9144    0.5415    2.1893    1.3277    1.3199

The VIF expresses the rate at which the estimator's variance increases when
collinearity exist.

If VIF > 10 then multicollinearity is high

Variance Decomposition

 sValue   var1    var2    var3    var4    var5    var6    var7    var8
-------------------------------------------------------------------------
 2.7645   0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.5977   0.0089  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0039   0.0001  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0000   0.0024  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0000   0.0000  0.0000  0.0001  0.0000  0.0002  0.0010  0.0000  0.0000
 0.0000   0.0115  0.0001  0.0004  0.0110  0.0196  0.0484  0.0000  0.0431
 0.0000   0.0380  0.0001  0.0014  0.0033  0.7005  0.3951  0.2063  0.0401
 0.0000   0.9392  0.9998  0.9981  0.9857  0.2798  0.5555  0.7937  0.9168

Condition_Indices
   1.0e+10 *

    0.0000
    0.0000
    0.0000
    0.0000
    0.0120
    0.2887
    0.5739
    4.0833

If Condition Index < 10, multicollinearity is small
If 10 <= Condition Index <= 30, multicollinearity is medium
```

```
If Condition Index > 30, multicollinearity is high


Graphs Explanation

For Normality:
An equal distribution of the residuals around their median suggests
the existence of normal distribution

For Homoscedasticity:
The increase in the variance as the fitted values increase suggests
possible heteroscedasticity
On the contrary, the decrease in the variance as the fitted values
increase suggests possible homoscedasticity

For Autocorrelation:
A trend among the residuals indicates a possible correlation between them

For Independence
If the data are normally distributed and uncorrelated then they are independent

For Linearity
If the data do not exhibit a pattern, then linearity could be accepted
```

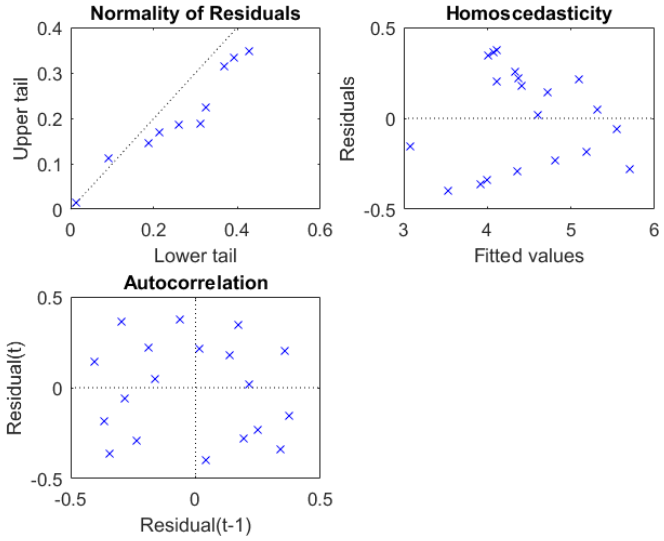


Figure B.8: Assumptions

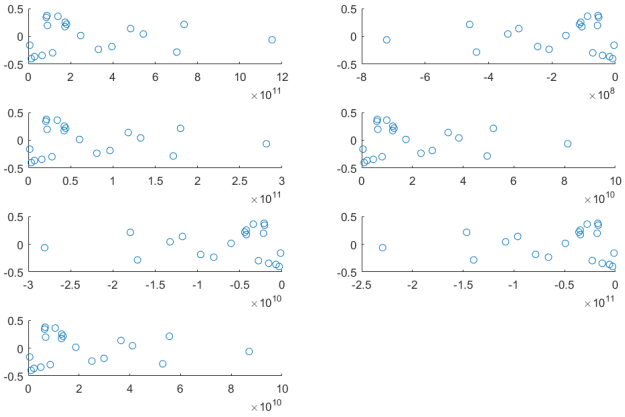


Figure B.9: Linearity assumption

B.2.3 2011-2015

```
Descriptive_Statistics

 Maximum_    Minimum_     Average_     Median_     Standard_   Variance
 Value       Value        Value        Value       Deviation
 ---------  ----------    --------    --------    ----------  ----------

 5.5292       3.0269      4.5573       4.6681      0.70763     0.50074
 1.3824e+12   7.1995e+09  3.3063e+11   2.0724e+11  3.5122e+11  1.2335e+23
-6.4727e+08  -1.2422e+11 -2.9709e+10  -1.8622e+10  3.1559e+10  9.9594e+20
-3.708e+08   -7.1192e+10 -1.7027e+10  -1.0673e+10  1.8087e+10  3.2714e+20
-1.4656e+09  -2.815e+11  -6.7325e+10  -4.2198e+10  7.1517e+10  5.1147e+21
 2.4859e+10   1.2894e+08  5.9453e+09   3.7262e+09  6.3158e+09  3.9889e+19
 8.642e+10    4.4964e+08  2.0669e+10   1.2955e+10  2.1956e+10  4.8207e+20
 6.4117e+10   3.3383e+08  1.5335e+10   9.6117e+09  1.629e+10   2.6535e+20

First row corresponds to the dependent variable followed by the independent ones,
in order.

Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5 + x6 + x7

Estimated Coefficients:
                    Estimate          SE          tStat       pValue
                   -----------    ----------    ---------    --------

    (Intercept)         4.0111        1.3751       2.9169    0.012913
    x1              0.00014486    0.00060448      0.23965     0.81465
    x2              0.00024837     0.0073778     0.033665      0.9737
    x3             -0.00015032     0.0032925    -0.045656     0.96434
    x4              0.00074675     0.0034409      0.21702     0.83184
    x5              -0.0001515      0.012717    -0.011913     0.99069
    x6              0.00092311      0.010518     0.087764     0.93151
    x7             -0.00071601      0.016952    -0.042238       0.967


Number of observations: 20, Error degrees of freedom: 12
Root Mean Squared Error: 0.484
R-squared: 0.705,  Adjusted R-Squared 0.533
F-statistic vs. constant model: 4.1, p-value = 0.0159

ANOVA
                SumSq     DF    MeanSq        F         pValue
```

```
                ------    --    -------    ------    --------

    Total       9.5141    19    0.50074
    Model       6.7086     7    0.95837    4.0993    0.015891
    Residual    2.8055    12    0.23379

Coefficient_Confidence_Intervals

    1.0150    7.0072
   -0.0012    0.0015
   -0.0158    0.0163
   -0.0073    0.0070
   -0.0068    0.0082
   -0.0279    0.0276
   -0.0220    0.0238
   -0.0377    0.0362


Kolmogorov-Smirnov Normality test
Test Hypothesis:
Ho: The Residuals come from a standard normal distribution
Ha: The Residuals do not come from such a distribution
Kolm_Smirnov_pvalue
    0.0306

The test rejects the null hypothesis at the chosen significance level


Durbin Watson autocorrelation test
Test Hypothesis:
Ho: The residuals are not autocorrelated
Ha: The residuals are autocorrelated

Durbin_Watson_pvalue =

    0.6136

The test fail to reject the null hypothesis at the Assumptions Significance Level

Runs test for independence
Test Hypothesis:
Ho: The values of the residuals come in random order
Ha: The values of the residuals do not come in random order
Runs_Test_pvalue
    0.8281
```

```
The test fail to reject the null hypothesis at the chosen significance level

Collinearity test

VIF:
   1.0e+15 *

   -4.3475   -3.6351   -0.1814   -0.2229   -0.2294   -7.1007    5.2092

The VIF expresses the rate at which the estimator's variance increases when
collinearity exist.

If VIF > 10 then multicollinearity is high

Variance Decomposition

 sValue   var1    var2    var3    var4    var5    var6    var7    var8
-------------------------------------------------------------------------
 2.7619   0.0001  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.6097   0.0277  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0001   0.0005  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0000   0.0073  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000  0.0000
 0.0000   0.0007  0.0000  0.0017  0.0028  0.0014  0.0021  0.0016  0.0009
 0.0000   0.0125  0.0012  0.0072  0.0013  0.0136  0.0001  0.0106  0.0099
 0.0000   0.0403  0.8371  0.1371  0.9548  0.0111  0.2209  0.4963  0.1445
 0.0000   0.9107  0.1617  0.8539  0.0411  0.9740  0.7769  0.4916  0.8447

Condition_Indices
   1.0e+10 *

    0.0000
    0.0000
    0.0000
    0.0000
    0.0450
    0.1197
    0.8440
    1.0806

If Condition Index < 10, multicollinearity is small
If 10 <= Condition Index <= 30, multicollinearity is medium
If Condition Index > 30, multicollinearity is high
```

```
Graphs Explanation
For Normality:
An equal distribution of the residuals around their median suggests
the existence of normal distribution

For Homoscedasticity:
The increase in the variance as the fitted values increase suggests
possible heteroscedasticity
On the contrary, the decrease in the variance as the fitted values
increase suggests possible homoscedasticity

For Autocorrelation:
A trend among the residuals indicates a possible correlation between them

For Independence
If the data are normally distributed and uncorrelated then they are
independent

For Linearity
If the data do not exhibit a pattern, then linearity could be accepted
```

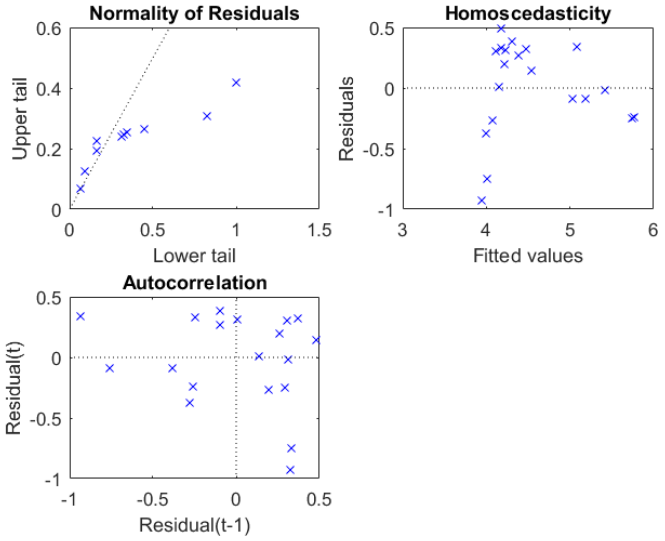


Figure B.10: Assumptions

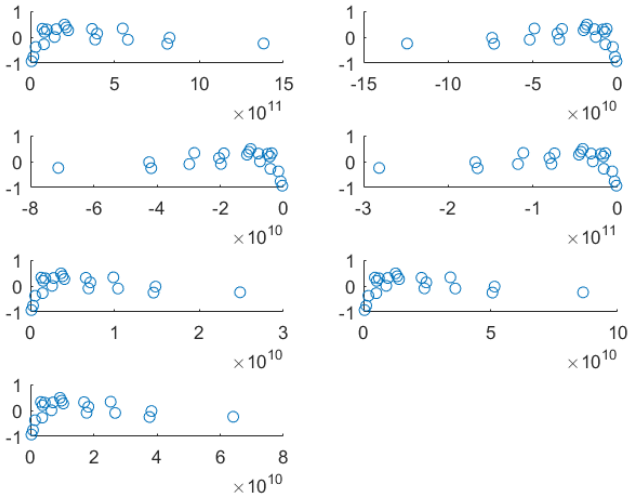


Figure B.11: Linearity assumption

B.2.4 Overall Model

```
Descriptive_Statistics

Maximum_      Minimum_       Average_      Median_       Standard_     Variance
 Value         Value          Value         Value         Deviation
----------    -----------    ---------     ----------    -----------    ----------

 5.4935        2.9323         4.4588        4.5779        0.72861        0.53087
 1.0767e+12    6.1006e+09     2.6519e+11    1.6227e+11    2.8062e+11     7.8749e+22
 1.0714e+08   -1.0919e+10    -2.1216e+09   -1.2573e+09    2.631e+09      6.9223e+18
 1.29e+11      7.9458e+08     3.2766e+10    1.8437e+10    3.473e+10      1.2062e+21
-2.3752e+08   -5.2013e+10    -1.1924e+10   -7.2719e+09    1.2864e+10     1.6548e+20
-2.3703e+08   -3.6847e+10    -9.4289e+09   -5.4016e+09    9.7304e+09     9.4681e+19

First row corresponds to the dependent variable followed by the independent ones, in order

Linear_Model

Linear regression model:
    y ~ 1 + x1 + x2 + x3 + x4 + x5

Estimated Coefficients:
                    Estimate          SE         tStat        pValue
                   -----------    ----------    --------    ----------

    (Intercept)         3.8466       0.17297      22.239    2.5401e-12
    x1              -1.8586e-07    3.3977e-07    -0.54701       0.59298
    x2              -2.5713e-07     5.127e-07    -0.50152        0.6238
    x3                1.119e-06    2.0533e-06     0.54496       0.59436
    x4              -1.0628e-06    1.9306e-06    -0.55053       0.59063
    x5               6.3235e-08    1.4799e-07     0.42729       0.67567


Number of observations: 20, Error degrees of freedom: 14
Root Mean Squared Error: 0.475
R-squared: 0.687,  Adjusted R-Squared 0.575
F-statistic vs. constant model: 6.13, p-value = 0.00328

ANOVA
                SumSq     DF    MeanSq       F       pValue
                ------    --    -------    -----    --------

    Total       10.086    19    0.53087
    Model       6.9249     5      1.385    6.133    0.003275
    Residual    3.1616    14    0.22583
```

```
Coefficient_Confidence_Intervals

    3.4756    4.2176
   -0.0000    0.0000
   -0.0000    0.0000
   -0.0000    0.0000
   -0.0000    0.0000
   -0.0000    0.0000

Kolmogorov-Smirnov Normality test
Test Hypothesis:
Ho: The Residuals come from a standard normal distribution
Ha: The Residuals do not come from such a distribution
Kolm_Smirnov_pvalue
    0.0450

The test rejects the null hypothesis at the chosen significance level

Durbin Watson autocorrelation test
Test Hypothesis:
Ho: The residuals are not autocorrelated
Ha: The residuals are autocorrelated

Durbin_Watson_pvalue =

    0.6206

The test fail to reject the null hypothesis at the Assumptions Significance Level.

Runs test for independence
Test Hypothesis:
Ho: The values of the residuals come in random order
Ha: The values of the residuals do not come in random order
Runs_Test_pvalue
    0.4768

The test fail to reject the null hypothesis at the chosen significance level

Collinearity test
VIF:
   1.0e+11 *

    7.6513    0.0015    4.2802    0.5191    0.0017
```

The VIF expresses the rate at which the estimator's variance increases when collinearity exist

If VIF > 10 then multicollinearity is high

Variance Decomposition

sValue	condIdx	var1	var2	var3	var4	var5	var6
2.3364	1	0.0066	0.0000	0.0000	0.0000	0.0000	0.0000
0.5951	3.9260	0.6715	0.0000	0.0000	0.0000	0.0000	0.0000
0.4274	5.4668	0.0422	0.0000	0.0000	0.0000	0.0000	0.0000
0.0665	35.1531	0.0559	0.0000	0.0000	0.0000	0.0000	0.0000
0.0000	49402.6145	0.0092	0.0000	0.0617	0.0002	0.0017	0.2250
0.0000	3.6581e+06	0.2145	1.0000	0.9383	0.9998	0.9983	0.7750

Condition_Indices

1.0e+06 *

0.0000
0.0000
0.0000
0.0000
0.0494
3.6581

If Condition Index < 10, multicollinearity is small
If 10 <= Condition Index <= 30, multicollinearity is medium
If Condition Index > 30, multicollinearity is high

Graphs Explanation

For Normality:
An equal distribution of the residuals around their median suggests the existence of normal distribution

For Homoscedasticity:
The increase in the variance as the fitted values increase suggests possible heteroscedasticity
On the contrary, the decrease in the variance as the fitted values increase suggests possible homoscedasticity

For Autocorrelation:
A trend among the residuals indicates a possible correlation between them

```
For Independence:
If the data are normally distributed and uncorrelated then they are independent

For Linearity:
If the data do not exhibit a pattern, then linearity could be accepted
```

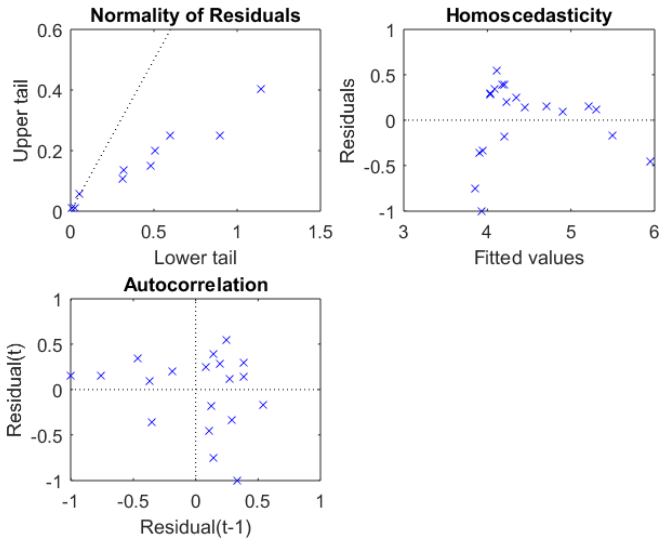


Figure B.12: Assumptions

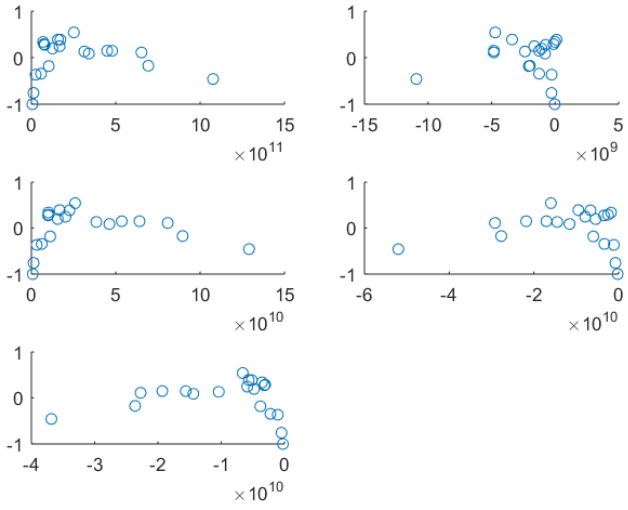


Figure B.13: Linearity assumption

Bibliography

[1] See Wikipedia, Descriptive statistics, `https://en.wikipedia.org/wiki/Descriptive_statistics (as of Mar. 27, 2018, 12:31 GMT).`

[2] Wasserman, L.: *All of Statistics: A Course in Statistical Inference*, Springer, New York (2004).

[3] See Creative Research Systems, Correlation, `https://www.surveysystem.com/correlation.htm (as of Mar. 27, 2018, 12:35 GMT).`

[4] Jolliffe, I.T.: *Principal Components Analysis, 2nd ed.*., Springer (2002).

[5] Haining, R.: *Spatial Data Analysis, Theory and Practice*, Cambridge University Press (2003).

[6] Cowpertwait, P.-S.P., Metcalfe, A.-V.: *Introductory Time Series with R*, Springer (2009).

[7] Christensen, L.: *Log-linear Models and Logistic Regression*, Springer (1997).

[8] Cox, D. D., Snell, E. J.: *The Analysis of Binary Data, 2nd ed.*., Chapman and Hall/CRC (1989).

[9] Hosmer, S., Lemeshow, D. W.: *Applied Logistic Regression*, Wiley (2000).

[10] McCulloch, C.-E., Searle, S.-R.: *Generalized, Linear, and Mixed Models*, Wiley (2001).

[11] Pagès, J.: *Multiple Factor Analysis by Example Using R*, Taylor & Francis Group/CRC (2015).

[12] Agresti, A.: *Analysis of Ordinal Categorical Data, 2nd ed..*, Wiley (2010).

[13] Holzmann, R.: *Aging Population, Pension Funds, and Financial Markets: Regional Perspectives and Global Challenges for Central, Easter, and Southern Europe*, The World Bank, 47687 (2009).

[14] Gill, I., Packard, T., Yermo, J.: *Keeping the promise of social security in Latin America*, The International Bank for Reconstruction and Development / The World Bank, 34433 (2005).

[15] Shiller, R.: *Finance and the Good Society*, Princeton University Press (2012).

[16] Beltrametti, L., Della Valle, M.: *The Implicit Pension Debt: its meaning and an international comparison*, Economia Internazionale / International Economics, Vol. 65, Issue 1, Pages:15-38 (2012).

[17] Wang, Y., Xu, D., Wang, A., Zhai, F.: *Implicit Pension Debt, Transition Cost, Options and Impact of China's Pension Reform: A Computable General Equilibrium Analysis*, The World Bank Policy Research Working Paper (2001-02).

[18] Knoema database :
`https://knoema.com/`.

[19] Organisation for Economic Co-operation and Development (OECD) database:
`https://data.oecd.org/`.

[20] OECD: Pension Markets in Focus, 2017 edition:
`http://www.oecd.org/finance/private-pensions/pensionmarketsinfocus.htm`.

[21] IMF World Economic Outlook(WEO) database:
`http://www.imf.org/en/Data`.

[22] The World Bank database:
`https://data.worldbank.org/`.

[23] United Nations, Department of economic and social Affairs database:
`http://www.un.org/en/development/desa/population/publications/database/index.shtml`.

[24] Mathworks:
`https://www.mathworks.com/?s_tid=gn_logo`.

[25] Matlab Gui:
`https://www.mathworks.com/discovery/matlab-gui.html`.

[26] IBM:
`https://www.ibm.com/analytics/hadoop/big-data-analytics`.

[27] BusinessDictionary:
`http://www.businessdictionary.com/definition/stem-leaf-plot.html`.

[28] Cheever, E.:
`http://lpsa.swarthmore.edu/MtrxVibe/EigMat/MatrixEigen.html`,
Swarthmore College, Department of Engineering.

[29] International Organisation of Pension Supervisors (IOPS): *Toolkit for Risk-based pension supervision*, Module 3, Identifying Risks (2012).

[30] Edited by Mészáros J.: *PENSION ADEQUACY AND SUSTAINABILITY*, Report of the conference of the Central Administration of National Pension Insurance. Organized in cooperation with the Ministry for National Economy held in Budapest, Hungary on 20th September 2013.

[31] Christensen, R.: *Plane Answers in Complex Questions: The Theory of Linear Models, 4th ed.*., Springer (2011).

[32] Beale, E.M.L., Kendall, M.G., Mann, D.W.: *The discarding of variables in multivariate analysis*, Biometrika, Vol. 54, Issue 3 & 4, Pages:357-366 (1967).

[33] Lay, C.D., Lay, S.R., McDonald, J.J.: *Linear Algebra and its Applications*, Pearson Education (2015).

[34] Jolliffe, I.T.: *Discarding Variables in a Principal Component Analysis. I: Artificial Data*, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 21, No. 2, Pages: 160-173 (1972).

[35] Draper, N.R., Smith, H.: *Applied Regression Analysis, Third Edition*, John Wiley & Sons (1998).

[36] Sheather, S.: *A Modern Approach to Regression with R*, Springer Science & Business Media (2009).

[37] Gujarati, D.N., Porter, D.C.: *Basic Econometrics*, McGraw-Hill Education (India) Pvt Limited (2009).

[38] Freedman, D.A.: *Statistical Models: Theory and Practice*, Cambridge University Press (2009).

[39] Scheffé, H.: *The Analysis of Variance*, ohn Wiley & Sons (1999).

[40] Kozak M, Piepho H-P.: *What's normal anyway? Residual plots are more telling than significance tests when checking ANOVA assumptions. J Agro Crop Sci. 2018;204:86–98.*

[41] Karagrigoriou, A.: *Course notes in Statistical Packages and Data Analysis*, University of the Aegean, Department of Statistics and Actuarial - Financial Mathematics.

[42] Dimitrakopoulou, T.: *Course notes in Statistics*, University of the Aegean, Department of Statistics and Actuarial - Financial Mathematics.

[43] Milionis, A.: *Course notes in Econometrics*, University of the Aegean, Department of Statistics and Actuarial - Financial Mathematics.

[44] Anderson , D.W.: *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons (2009).

[45] Pearson, K.: *On lines and planes of closest fit to systems of points in space*, The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Vol. 2, No.11, pages: 559-572 (1901).

[46] Hotelling, H.: *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, Vol. 24, Pages: 417–441 and 498–520 (1933).

[47] Hotelling, H.: *Relations between two sets of variates*, Biometrika, Vol. 28, No. 3/4, Pages: 321–377 (1936).

[48] Cattell, R.: *The scree test for the number of factors*, Multivariate Behavioral Research, Vol. 1, No. 2, Pages: 245–76 (1966).

[49] Woods, H., Steinour, H.H., Starke, H.R.: *Effect of Composition of Portland Cement on Heat Evolved during hardening*, Industrial and Engineering Chemistry, Vol. 24, No. 11, Pages:1207-1214 (1932).

[50] King, J.R., Jackson, D.A.: *Variable Selection in Large Environmental Data Sets Using Principal Components Analysis*, Environmetrics, Vol. 10, Pages: 67-77 (1999).

[51] Geary, R.C., Leser, C.E.V.: *Significance Tests in Multiple Regression*, The American Statistician, Vol. 22, No. 1, Pages:20-21 (1968).

[52] Largey, A., Spencer, J.E.: *F- and T-Tests in Multiple Regression: The Possibility of 'Conflicting' Outcomes*, Journal of the Royal Statistical Society. Series D (The Statistician), Vol. 45, No. 1, Pages: 105-109 (1996).

[53] Galton, F.: *Regression Towards Mediocrity in Hereditary Stature*, The Journal of the Anthropological Institute of Great Britain and Ireland, Vol. 15 , Pages: 246-263 (1886).

[54] Tukey, J.: *Exploratory Data Analysis*, Addison-Wesley Publishing Company (1977).

[55] Fisher, R.A.: *The Correlation Between Relatives on the Supposition of Mendelian Inheritance*, Philosophical Transactions of the Royal Society of Edinburgh, Vol. 52, Pages: 399–433 (1918).

[56] Fisher, R.A.: *On the "Probable Error" of a Coefficient of Correlation Deduced from a Small Sample*, Metron, Vol. 1 (1921).

[57] Cortes, C., Vapnik, V.: *Support-Vector Networks*, Machine Learning, Vol. 20, No. 3, Pages: 273-297 (1995).