Evaluation of Models for predicting the average monthly Euro versus Norwegian krone exchange rate from financial and commodity information

Raju Rimal

A Dissertation

Presented to the Faculty

of Norwegian University of Life Sciences

in Candidacy for the Degree

of Masters of Bioinformatics and Applied Statistics

Recommended for Acceptance

by the Department of

IKBM

Supervisor: Ellen Sandeberg and Trygve Almoe

Dec 2014

Abstract

This is Abstract File ....

Acknowledgements

I would like to express my gratitude to my supervisors Ellen Sandberg and Trygve Almoe for their guidance and invaluable suggestions. I am greatful to Prof. Solve Saebo for his helpful advice and instructions. I want to thank Statistic Norway and Norges Bank for making valuable data easily availiable without which it is impossible to complete my thesis.

Being a student of a developing country, I was unaware of programming, mordern statistical methods and academic writing. With the encouragement of Prof. Solve Saebo and my supervisors, I have completed by thesis with extensive use of R programming and modern statistical tools.

Finally, my special thanks also goes to my families and friends for their continious support and encouragement.

To my parents.

Contents

List of Tables

List of Figures

Abbreviations and Symbols

Abbreviations and their full forms used in this Thesis

Short.Form	Full.Form
PCA	Principal Component Analysis
PLS	Partial Least Square
PCR	Princiapal Component Regression
AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
Cp	Mallows'sCp
VIF	Variance Inflation Factor
RMSE	Root-Mean-Square Error
RMSEP	Root-Mean-Square Errorof Prediction
RMSECV	Root-Mean-Square Errorof Cross-validation
R2pred	PredictedR-squared
VAR	Vector Autoregression
ARIMA	Autoregressive Integrated Moving Average
ADL	Autoregressive Distributed Lag
NOK	Norwegian Krone
USD	United State Dollor

Symbols and their meaning used in this Thesis

Symbols	Meaning
Bold Symbols like,	
$\mathbf{X}$, $\mathbf{Y}$	Matrices and Vectors
Sigma ($\mathbf{S}$)	Popularion (Sample) variance-covariance matrix
R^2 adj	Adjusted coefficient of determination
CVadj	RMSECV adjusted for bias
`cp.model`	Subset of linear model selected with minimum Mallow's Cp Criteria
`r2.model`	Subset of linear model selected with maximum R^2 adjusted Criteria
`aic.model`	Subset of linear model selected with minimum AIC Criteria
`bic.model`	Subset of linear model selected with minimum BIC Criteria
`forward.model`	Subset of linear model selected based on F-test Criteria using forward selection procedure
`backward.model`	Subset of linear model selected based on F-test Criteria using backward elimination procedure
train	Training Dataset (From Jan 2000 to Dec 2012)
test	Test Dataset (From Jan 2013 to Nov 2014)
λ	Ridge Regression Parameter
Q^2	R^2 predicted
`PerEURO`	Exchange Rate of Norweian Krone Per Euro (Response Variable)

Chapter 1

Introduction

Initial Paragraph

Trading started from the very beginning of human civilization. People used to trade with goods at the time but with advancement of development people started using gold, silver and finally money. The process is not restricted within a country. Some countries are powerful and some are not so as their currencies. Currency of another country becomes essential to buy things from that country. Here comes the roal of exchange rate. Buying powerful currencies requires large sum of weak currencies.

Any international trade is conducted through more than one currencies. Participants in the international trade require to exchange their currency which is performed by foreign exchange market. "The foreign exchange market (ForEx) is the mechanism that brings together buyers and sellers of different currencies" (Appleyard, Field, and Cobb, 2014).

As any other commodity, exchange rate is also determined from its demand and supply in money market. All those economic activities that exist between countries

create demand and supply of the currencies which consequently determine the exchange rate. The economic activities betwen countries are recorded as balance of payment account. Thus the balance of payment captures all the demand and supply of foreign currency (Fang and Kwong, 1991). When the domestic demand for foreign currency exceeds the foreign demand of domestic currency i.e. a deficit in the balance of payment, the domestic currency depreciate (*Balance of Payments Deficits and Surpluses*).

Foreign currencies are involved in various activities such as, (a) imports and exports of goods and services, (b) interest and devidends payed to foreign investment in domestic market, (c) interst and devidends earned from investments made on foreign market, (d) all the currencies that enter into and leave from a country as income and expenditure.

Three factors affecting exchange rate are considred in this thesis. Primarily, total monthly imports and exports of goods are considered. Ships, oil platform, chemicals and food stuffs are major imports of Norway and petroleum products, machinery, equipments, chemicals and fishes are major exports. Since the economy of norway highly depend on petrolium productes, apart from imports and exports, the second component considered is the spot oil price. Third factor is the financial variables such as interest rate and consumer price index are considered. In interest rate - (a) key interest rate of Norway, (b) Loan interest rate (c) key interest rate of euro area are taken into account as factors affecting interest rate.

1.1 Methods opted for analysis

Univariate time series analysis is very common in Econometrics where Autoregressive (AR), Moving Average (MA) and Autoregressive integrated Moving average (ARIMA) are used. However, dealing a timeseries data with many predictor variable using latent variable and principal component methods is unconventional. This thesis aims to analysis a timeseries data with financial and commondity data, as discussed before, using statistical regression methods such as - Multipal Linear Regression, Ridge Regression, Principal Component Regression (PCR) and Partial Least Square (PLS) Regression. Apart from these, a subset models which selected from the Multipal Linear Regression using various criteria are also used. An application of PCR and PLS on time series data makes this thesis distinct.

1.2 Sources of Dataset

Data related to balance of payment such as import, export and trade balance used in this thesis are obtained from Statistics Norway. Consumer price index is also obtained from the same source. Interest rate variable related to Norway are obtained from Norges' Bank and the key interest rate for euro zone is obtained from Euro Bank while the oil spot price is obtained from US Energy information system. The average monthly spot price for brent oil was on Doller per Barrel unit which was converted into NOK using NOK per USD exchange rate for that month.

1.3 Objective of thesis

There are three main objective of this thesis-

1. To analyze the relationship of foreign exchange rate with teh financial (price, indices and exchange rate) and commodity (imports, exports and trade balance) information
2. Prediction of external sample (Exchange Rate) using various models
3. Comparision of the Models considered on the basis of goodness of their fit and their predictive ability

Chapter 2

Data and Material

Prediction of dynamics of Exchange Rate through Economic and Financial indicators is the main aim of this thesis. From these two broad categories, only those factors were considered which are believed to be useful to understand the exchange rate dynamics.

2.1 ForeX Market

Foreign Exchange(Fx) Market is the most traded and liquid financial market where individuals, firms and banks buy and sell foreign currencies. Forex market constitute of monetary counters connected electronically which are in constant contact forming a single international financial market. The market remains open 24 hr a day for five working days of a week (*Introduction to the Forex Market*).

Currencies are exchanged for activities like trade, tourism and investments in another countries. For instance, a person visiting France needs euro since euro is accepted in France. On returning back from the visit (s)he might want to exchange

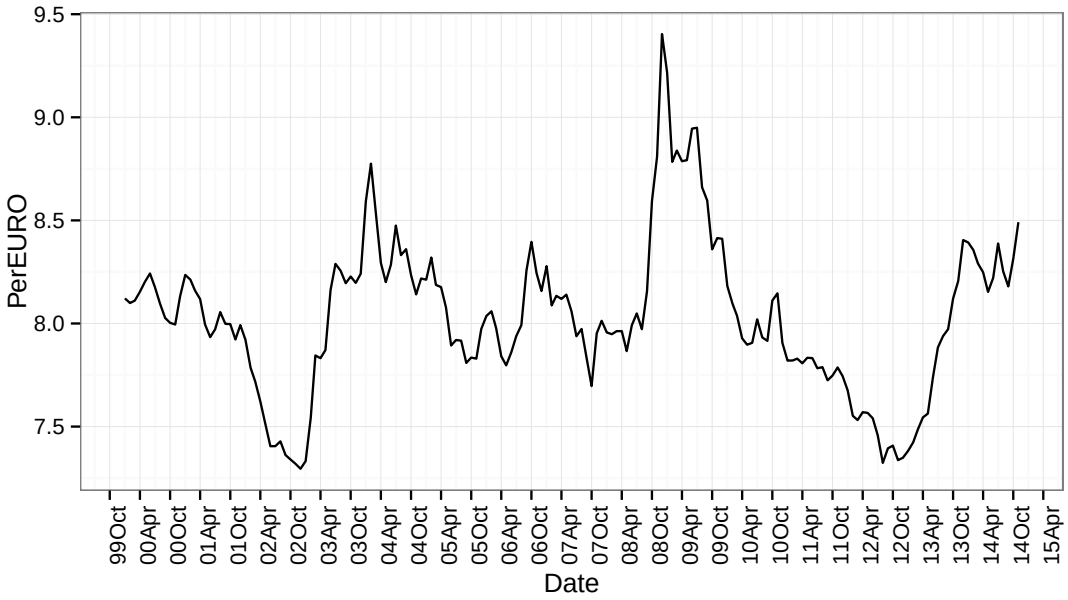


Fig 2.1: Exchange rate of Norwegian Krone per Euro

back those Euros to Norwegian Krone. This transaction is affected by the exchange rate of Norwegian Krone per Euro. The exchange rate of NOK per Euro over time is plotted in figure-2.1.

Exchange rate can be set according to different macroeconomic variables, such as interest rate, price index, balance of payment etc. Such exchange rate determined by ForEx market transaction is called Floating exchange rate. Some country fix exchange rate while others pegged with other currency. Norway has a floating exchange rate.

2.2 The Norwegian krone (NOK)

After introduction of Krone in April 1875 (*Brief History Of Norges Bank* 2014-11), Norway was pushed to join the Scandinavian Monetary Union established on

1873 (*Norwegian Kroner* 2014/12). Although the Union was formally abolished on 1972, Norway decided to keep the names of its currencies. In December 1982, due to heavy speculation, Norges Bank (Central Bank of Norway) decided to fix Norwegian Krone which later floated on 1992 (*Brief History Of Norges Bank* 2014-11).

2.3 EURO

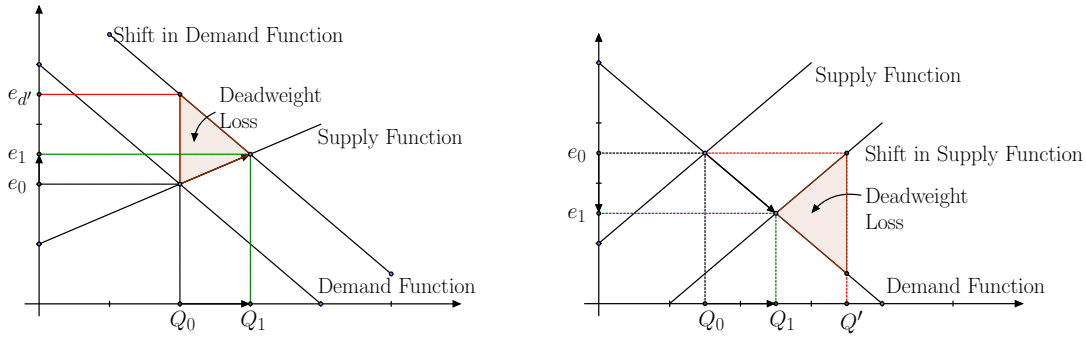
Euro, the official currency in the Eurozone, was introduced as a virtual currency in 1999 and later as physical in 2002. It is the single currency shared by 19[1] of the European Union's Member States of Euro Area. Although European Central Bank (ECB) manages Euro, the fiscal policy (public revenue and expenditure) are in the hands of individual national authorities. The single currency market throughout the euro zone not only makes traveling across the countries easier but also helps the member country to keep their economy sound and stable. This situation removes currency exchange cost, smooth international trade and consequently gives them more powerful voice in the world. A stable economy and larger area protects euro zone from external economic fluctuation, instability in currency market and unpredictable rise in oil prices.(*The euro* 2015)

2.4 Factors influencing Exchange Rate

The demand of any currency relative to its supply determines its price, just like any other commodity. For each possible price of a Norwegian Krone, there is

[1] `https://www.ecb.europa.eu/euro/intro/html/index.en.html`

a corresponding demand and supply to be exchanged with euro in the money market. When demand of krone equals its supply, the price it exhibit at some specific time is called its equilibrium exchange rate. Factors like inflation, interest rates, expectation and government policy affects the demand for any currency. But the supply is mostly in control of central bank. In a floating exchange rate regime, the shift in demand (fig-2.2a) and supply(fig-2.2b) function determines equilibrium exchange rate of any currency.



(a) Demand Shift and Exchange Rate Equilibrium

(b) Supply Shift and Exchange Rate Equilibrium

Fig 2.2: Effect of shifts on demand and supply of currencies on their Exchange rates

In case of demand shift, being currency supply constant, the exchange rate will suddenly rise to e_d creating deadweight loss (also known as excess burden or allocative inefficiency[2]) which consequently pushes the supply from Q_0 to Q_1 creating a new equilibrium exchange rate at e_1. In the similar fashion, if the market is over flooded with currency, shifting the supply function and creating deadweight loss, the exchange rate is pressed from e_0 to create a new equilibrium at e_1. In both the situation, the quantity supplied althought being increased, the first one leads to a rise in exchange rate while the other leads to its fall.

[2] `http://www.princeton.edu/~achaney/tmve/wiki100k/docs/Deadweight_loss.html`

Madura (2012, p. 103) suggested an equation consisting those macroeconomic factors that can affect the demand and supply of any currency and consequently the exchange rate as,

$$e = f(\Delta INF, \Delta INT, \Delta INC, \Delta GC, \Delta EXP) \tag{2.1}$$

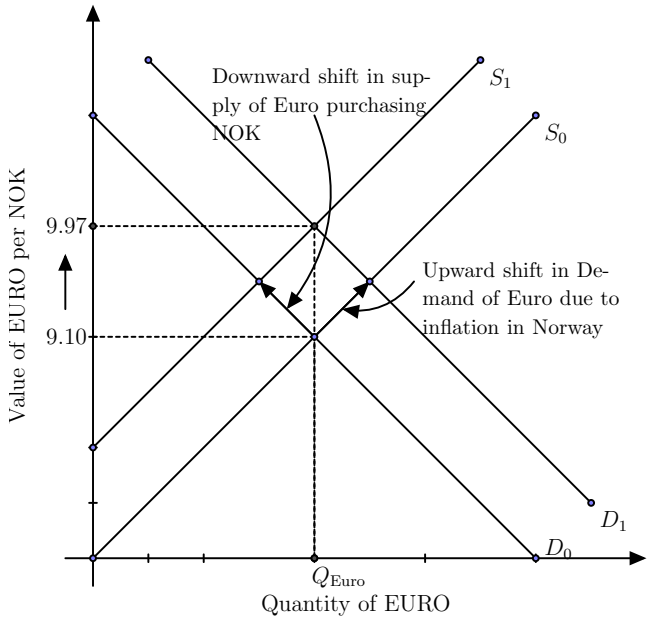
where,

e:	percentage chage in spot exchange rate
Δ INF:	change in inflation differential between two countries (currencies)
Δ INT:	change in interest rate differential between two countries
Δ INC:	change in the income level differential between two countries
Δ GC:	change in government control
Δ EXP:	change in currency value expectations

2.4.1 Inflation

Inflation is the steady rise in overall price level, i.e. a decrease in the value of currency. In other words, more amount of money is needed to buy same goods than previous. Relative change in inflation has effect on exchange rate. For instance, an abrupt rise in the inflation in Norway relative to the Eurozone, Norwegian products becomes relatively expensive in terms of Norwegian Currency. On one hand, this would increase the demands for Eurozone goods, and consequently the demand for euro increases in the short run. On the other hand, expensive Norwegian goods becomes less attractive in Eurozone and therefore reduce the supply of euro purchasing Norwegian kroner. In figure -2.3, the demand function of Euro

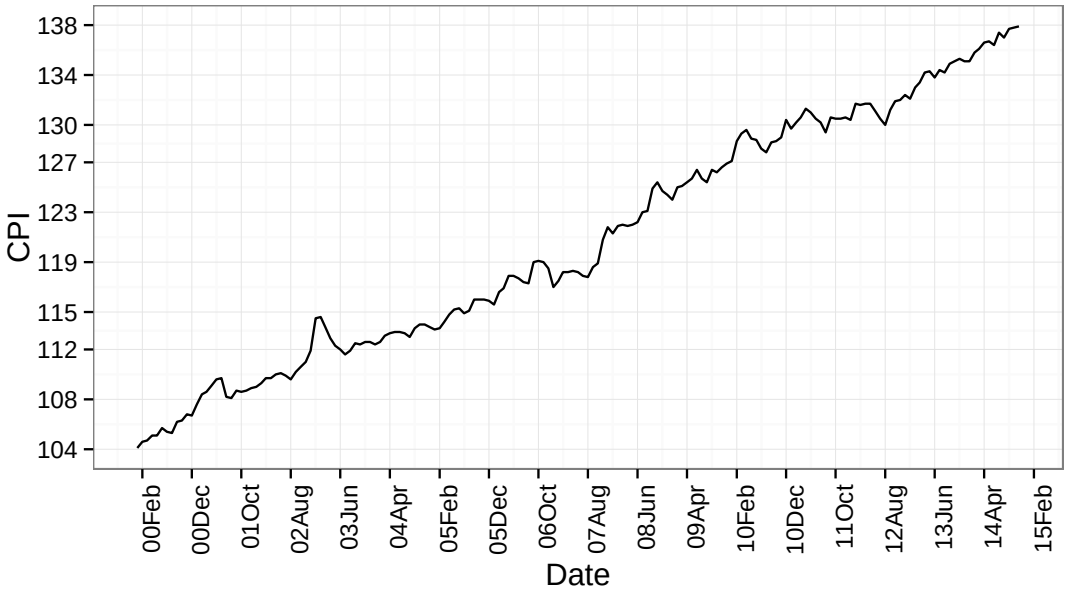
shift upward due to inflation of NOK, i.e. Eurozone goods are more attractive than Norwegian goods and the downward shift on supply function occurs as the customers are less interested in Norwegian products. As a result the value of Euro per NOK increases from 9.10 to 9.97, i.e Norwegian Krone deprecates against the Euro (Madura, 2012, p. 104).



Source: Madura, 2012

Fig 2.3: Effect of inflation on Exchange Rate Equilibrium

Statistics Norway prepares and publishes the official figures for inflation, the consumer price index (CPI) with base year at 1998. Since the real value of money is constantly decling, high inflation means that storing money is expensive. while low and stable inflation contributes to an efficient distribution of resources in a market economy (*FAQ: Monetary Policy, Inflation and Interest Rates* 2007). Data for CPI is obtained for this thesis from Norges bank, since this is an inportant factor



Source: Norges Bank

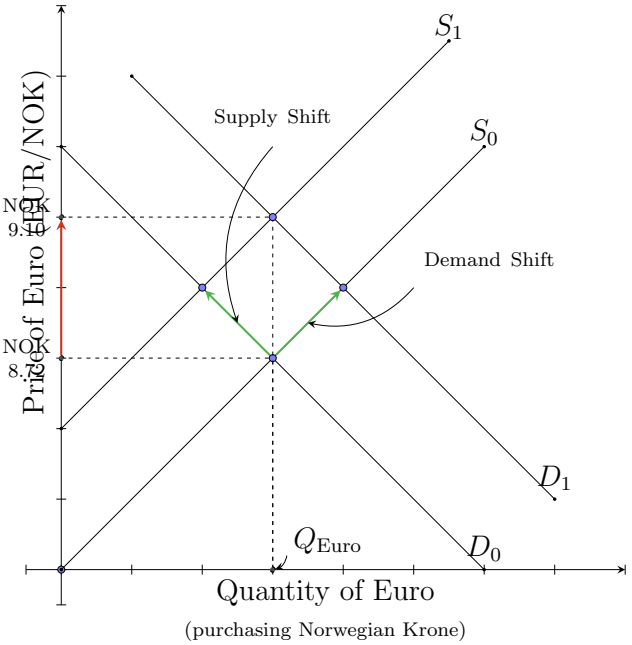
Fig 2.4: Time Series plot of Consumer Price Index (CPI)

that can influence exchange rate. The time-series plot for CPI in figure-**??** shows an steady increment over the time.

2.4.2 Interest Rate

Since Interest rate has impact on inflation and currency values, by manipulating it, central banks exert influence over both inflation and exchange rates. For example, a sudden increase in interest rate in Norway relative to Eurozone could have increase on investment of Eurozone in Norway with interest-bearing securities. The Eurozone investors wants to invest more in Norway which increases the demand for NOK in Eurozone. Due to stronger incentives, Norwegians also increase their domestic investment, as a result, the supply of NOK in currency market will re-

duce. The increase in Demand of NOK and decrease in its supply results a shift in exchange rate to lower level. The process is illustrated in figure - 2.5.

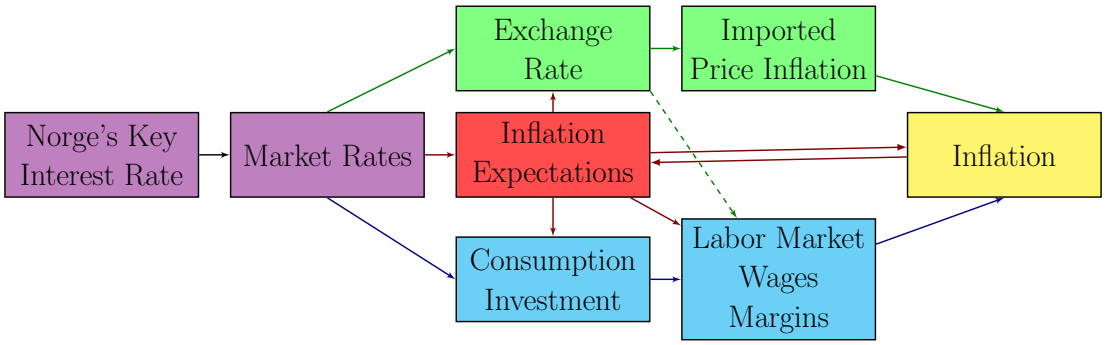


Source: Madura, 2012

Fig 2.5: Effect of Interest Rate change on Exchange Rate includes (a) Demand Shift: Due to increased interest rate in Norway, demand of Norwegian Krone increases creating a demand shift in demand function and (b) Supply Shift: The supply of Krone decrease as Norwegian increase their domestic investment creating a sortage of NOK in market.

The influence of market interest rate flows through multiple channel such as demand channel, exchange Rate channel and expectation Channel as shown in figure-2.6 (*Effect of Interest Rate Changes* 2004).

According to Madura (2012), change in interest rate in third country can also affect the exchange rates between NOK and Euro. For instance, the sudden increase of interest rate in US would shift the european investment from Norway to



Source: *Effect of Interest Rate Changes* 2004

Fig 2.6: Market Rate influence on demand channel, exchange rate channel and expectation channel

US which consequently reduce the demand of NOK resulting a downward pressure on its exchange rate with Euro.

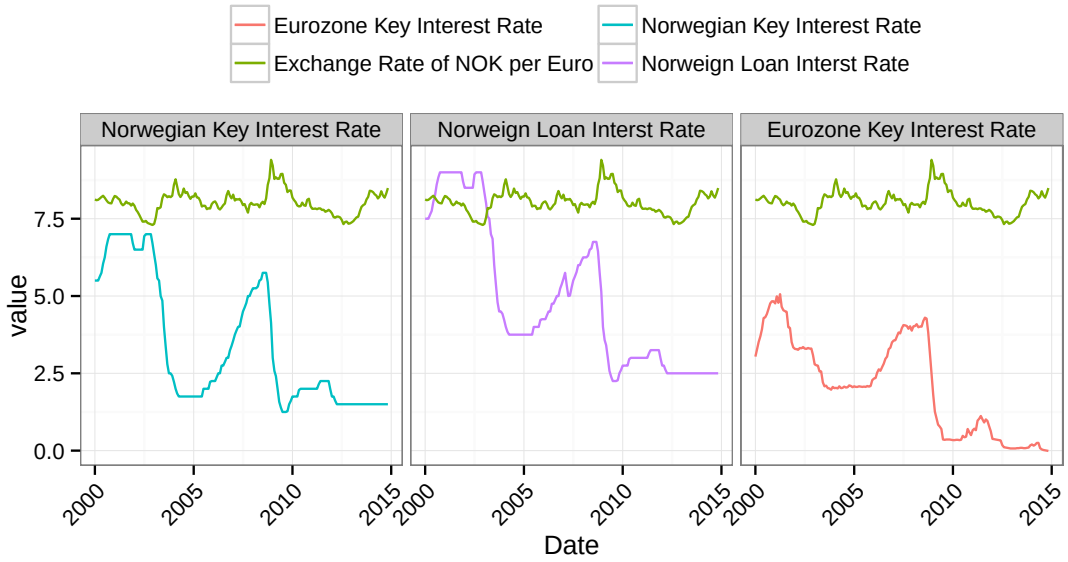


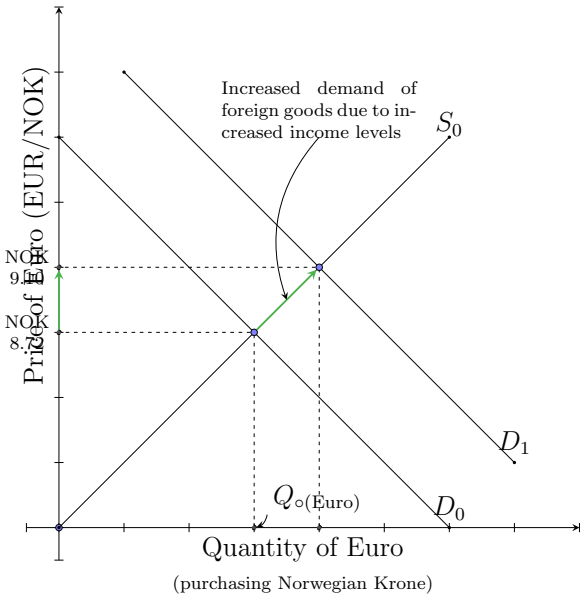
Fig 2.7: Interest Rates from Norway and Eurozone and their comparision with Exchange Rate showing a distinct inverse relationship

Since the interest rate is a key factor influencing exchange rate, the key interest rate of Norway and Eurozone along with the loan interest rate of Norway is

considered in this thesis. The time series plot of these variables are in figure - 2.7. Due to simultenious act of other variables, the plot does not exhibit any discrete relationship. However, the model fitted by the data collected suggest some indepth understanding of this relationship which is analysed and presented in chapter-4.

2.4.3 Income Levels

The rise in real income level increases the consumption level. Relative income levels of a country is another factor which can affect the demand of imported goods which consequently affect exchange rate (Madura, 2012). For instance, if the income levels of people of eurozone rises, other factor being constant, the demand for foreign goods in eurozone may increase which can shift the demand function outward and subsequently increase the exchange rate (figure-2.8).



Source: Madura, 2012

Fig 2.8: Effect of change in relative income levels on exchange rate *ceteris paribus*.

The example considered above is on the assumption of *ceteris paribus*, which in reality is not usual. The change in exchange rate due to income levels is also guided through the effect of income levels on interest rates and inflation. The increased income levels increase the consumption cause the economy to overheat. Central banks could increase interest rates to prevent overheating and increased inflation. Thus the relative change in income levels can affect exchange rates directly and indirectly (Madura, 2012, p. 106).

2.4.4 Government Control

Government Control is the fourth factor Madura (2012) has considered that can influence foreign exchange rate. Government can influence exchange rate in many ways including, (a) imposing foreign exchange barriers, (b) imposing foreign trade barriers, (c) intervening (buying and selling currencies) in the foreign exchange markets, and (d) affecting macro variables such as inflation, interest rates, and income levels. Norges Bank could force the currency to depreciate by flooding the market with NOK (i.e increasing supply) if Norway wants to boost its exports. Similarly, the bank could used their foreign currency reserve to purchase NOK to rise its value. Such direct interventions make considerable impact on the exchange rate. As a indirect intervention, the government can influening the underlying macroeconomic factors like inflation, interest rate and income level (Madura, 2012, p. 107).

2.4.5 Expectations

Response to new information in foreign exchange market is similar to other financial market. The current expectation for the future value is reflected in the exchange rate changes. Like in stock market, when a company publishes its prosperous financial statement, the stock price suddenly rises; the forex market also exhibit similar performance. For example, a news of increasing inflation in Norway cause currency traders to sell Norwegian Krone expecting a decrease in its future value. This expectation is immediately seen as a downward pressure on Norwegian Krone. The similar effect is obtained when speculator expects the currency to depreciate (Madura, 2012, p. 107).

A person of one country need the currency of another country for various purposes such as trade of goods and services, foreign investment and travelling. The actual flow of currency from one country to another is in these forms of activities. The volume of transaction of trade in terms of goods and services between specific countries is kept recorded as a form of balance of payment which can even have signal of possible shifts in exchange rate.

2.5 Balance of Payment

Although international trade is possessed in various forms, the transaction of multiple currency is common in each of them. A country keep these transactions with other countries as a form of Balance of Payments. A balance of payment account maintain a systematic records of these transactions conducted at some specific time period between a home country and others (those countries with which the transactions are made). A balance of payment account of a country exhibit the

size of its economic activities with rest of the world (Appleyard, Field, and Cobb, 2014, p. 462).

Since Balance of Payment is a bookkeeping system for inter countries economic activities, the items with payments inward to the home country are credited while payments outward from the home country are debited. Exports, inflow of foreign investment, interest and devidents obtained from the investment made on foreign country by the home country are considered as credited items as they increase the inward flow of currency. Similary, Imports, investment made on foreign countries, interest and devidents paied to foreign countries for their investment in home country are the items to be debited (Appleyard, Field, and Cobb, 2014, p. 465).

Table 2.2: Two components of Balance of Payments and their subdivision

Balance of Payment

Current Account	Capital Account
• Payments for Merchandise and Services • Factor Income Payments • Transfer Payments • Examples of Payment Entries • Actual Current Account Balance	• Direct Foreign Investment • Portfolio Investment • Other Capital Investment • Errors and Omissions and Reserves

Source: Madura, 2012

Balance of payment can be classified into two broad categories - (a) Current Account and (b) Capital Account. The items that lies in these subcategories are illustrated in table-2.2.

2.5.1 Current Account

Current account measures net imports and exports of a country. Imports and exports are divied into three sub categories - (a) Trade of goods, (b) Trade of services and (c) Income which includes the interest and dividend payed to international firms operating within home country and interest and dividends earned from domestically owned firms abroad (Krugman and Obstfeld, 2006).

The current account balance is the difference between export and import. When export of a country exceed its import, there is current account surplus and when import exceed export there is a current account deficit.

$$\text{Current Account} = \text{Total Exports} - \text{Total Imports} \tag{2.2}$$

Above equation can also be expressed as a form of income and expenditure (2.3) as the difference between Total National Income and Total Domestic consumption (Krugman and Obstfeld, 2006).

$$\text{Current Account Balance} = \underbrace{\text{Y}}_{\text{GNP}} - \underbrace{(\text{C} + \text{I} + \text{G})}_{\substack{\text{Total Domestic} \\ \text{Consumption}}} \tag{2.3}$$

where,

C = Consumption

I = Investment

G = Government Purchases

Current account incorportates a wide range of international transactions so there is a vital roal of exchange rate in each of those transaction. This thesis has considered the monthly data for imports and exports of goods which is availiable

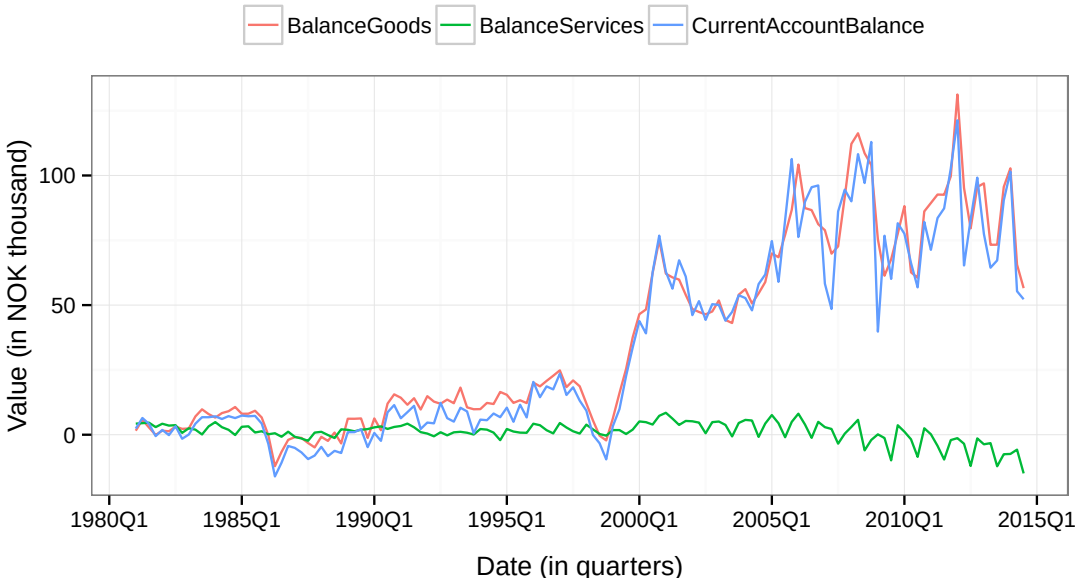


Fig 2.9: Current Account Balance prepared from quartely data from the year 1981 to 2014

from Statistics Norway. In Norway, current balance is highly influence by the balance in goods. Figure-2.9 shows that the balance in services in Norway is decreasing while the balance in Goods has boost up around after 1998. Further, the balance in services ploted in the same figure from the quartely data exhibit a seasonal trend which is usual in Norway.

Imports

Machinery & equipments, chemicals, metals and food stuffs are major imports of Norway. Sweden (13.6%), Germany (12.4%), China (9.3%), Denmark (6.3%), UK (6.1%) and US (5.4%) are major import partners [3]. The monthly imports of New Ships (`ImpNewShip`), Oil Platform (`ImpOilPlat`), Old Ship (`ImpOldShip`) and all

[3] `https://www.cia.gov/library/publications/the-world-factbook/geos/no.html`

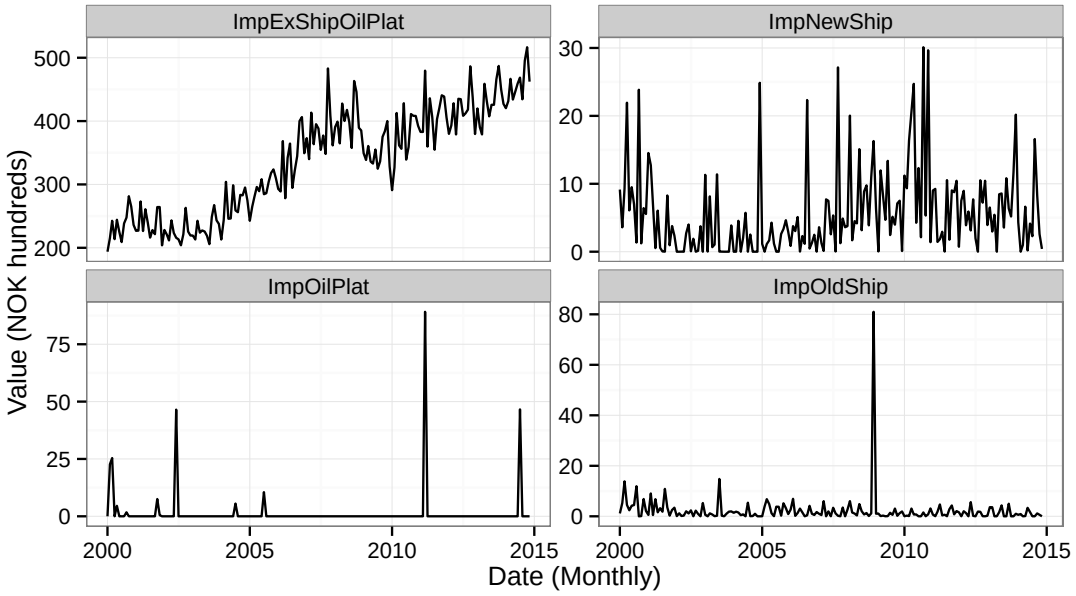


Fig 2.10: Time Series plot of major imports of Norway

other items excluding Ship and Oil Platform (`ImpExShipOilPlat`) are considered as predictor variable in data analysis of this thesis. The time-series plot for these variables are presented in figure-2.10

Exports

Norway is richly endowed with natural resources - petroleum, hydropower, fish, forests, and minerals but the economy is highly dependent on the petroleum sector [3]. Petroleum products, machinery and equipments, metals, chemicals, ships and fishes are major exports of Norway [3]. The monthly time series for the Export of condensed fuel (`ExpCond`), crude oil (`ExpCrdOil`), natural gas (`ExpNatGas`), new ships (`ExpNewShip`), oil platform (`ExpOilPlat`), old ships (`ExpOldShip`) and all other exports excluding ships and oil platform (`ExpExShipOilPlat`) are presented in figure-2.11.

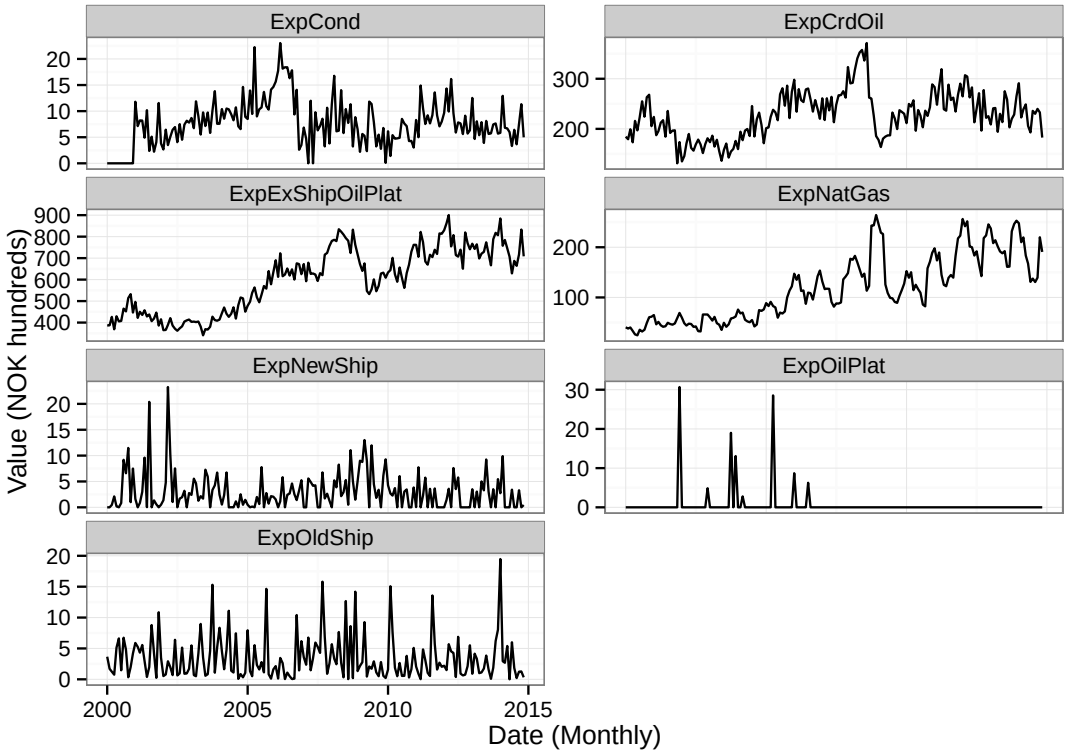


Fig 2.11: Time Series plot of major exports of Norway

2.5.2 Capital and Financial Accounts

The following text of capital and finacial accounts are adapted from *International financial management* by Madura (2012). A capital account includes transaction of inter-country transfer of financial assets due to immigration and non-financial assets such as buying and selling of patents and trademarks. These transaction are relatively minor in comparison to the items of financial accounts. The key elements of financial account are,

- **Direct Foreign Investment** includes investment in fixed assets in foreign countries.

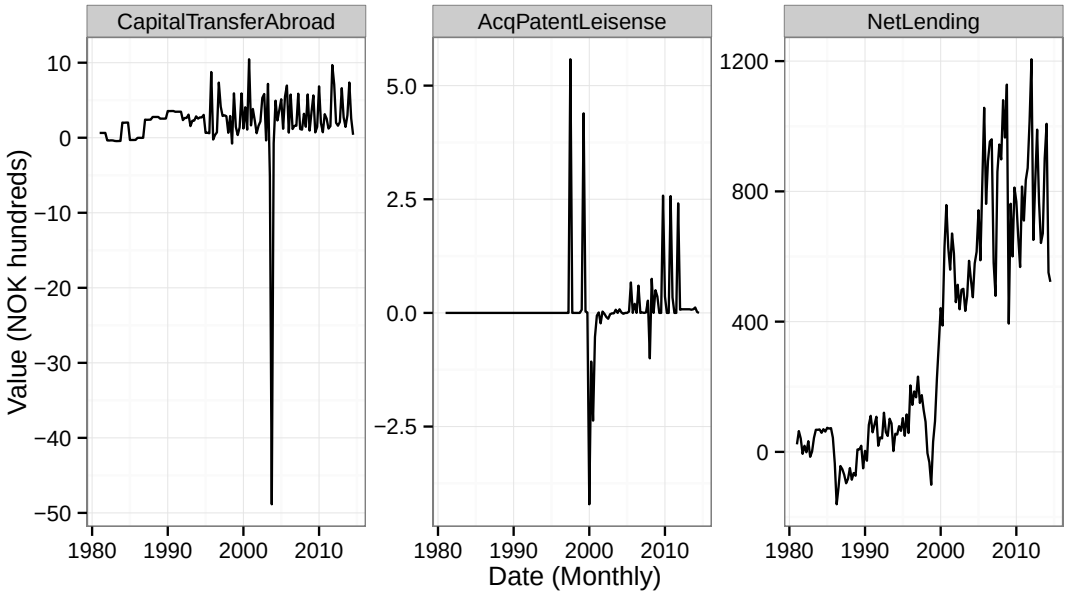


Fig 2.12: Time Series plot of variables related to capital account

- **Portfolio Investment** includes transaction of long term financial assets such as bonds and stocks.
- **Other Capital Investment** includes short term financial assets such as money market securities.
- **Errors, Ommissions and Reserves** includes adjustment for negative balance in current account.

Due to unavailability of monthly data for capital accounts, this thesis has not included the data in the analysis. The time series plot from quartely totals for the variables related to capital account are plotted in the figure-2.12. The figure shows that the economy of Norway has drastically heated after the year around 1998.

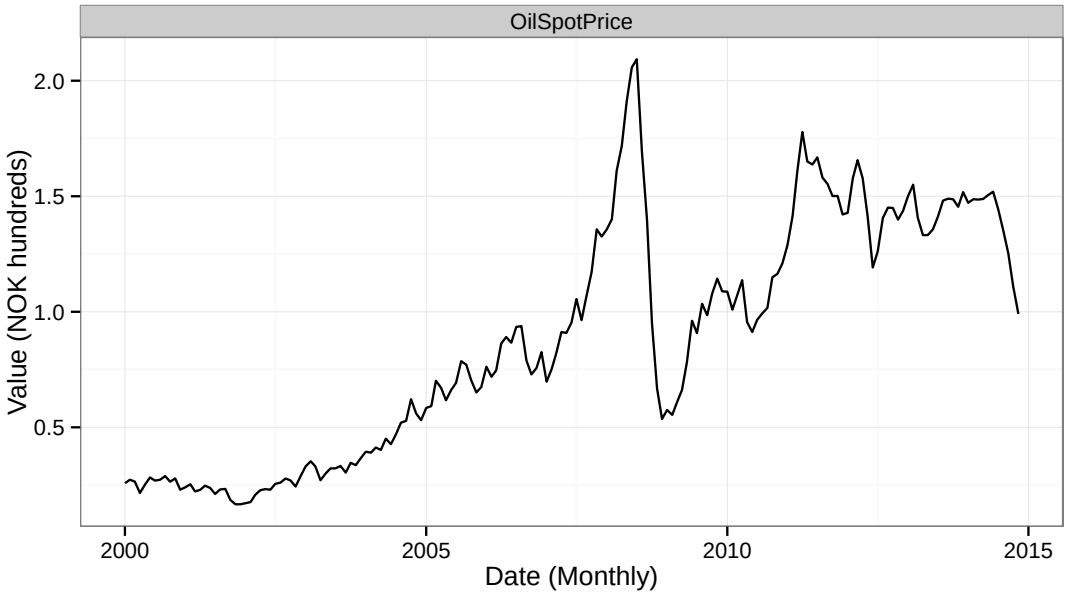


Fig 2.13: Time Series plot of oil spot price from Jan 2000

2.6 Oil Spot Price

After the discovery of oil in the North Sea in late 1969, economy of Norway has transformed completely (*Norway The rich cousin* 2013). Since the economy of Norway is highly depended on its petrolium related products, oil spot price also has influence on foreign exchange rate of Norway. However, Ferraro, Rogoff, and Rossi (2012) argued that the predictive ability of exchange rate from oil price is more effective at a daily frequency and is hardly visiable at monthly frequencies. Oil spot price is also considered as predictive variable in this thesis. The heavy fluctuation in the oil spot price shown in time series plot (fig-2.13) is due to the financial crisis of 2007-2008.

2.7 Lagged response variable as predictor

Exchange rate, being a time-series variable, contains autocorrelation which can be checked out by including the lagged variables of the response as predictor. Further, the correlation of response (`PerEURO`) with its first lag and second lag are 0.94 and 0.86 respectively. In addition, two spikes which are significant in the partial autocorrelation function as plotted in figure-2.14 also indicate for the use of auto-regressive terms in the model. This thesis has included the first and second lag of response variable as a predictor.

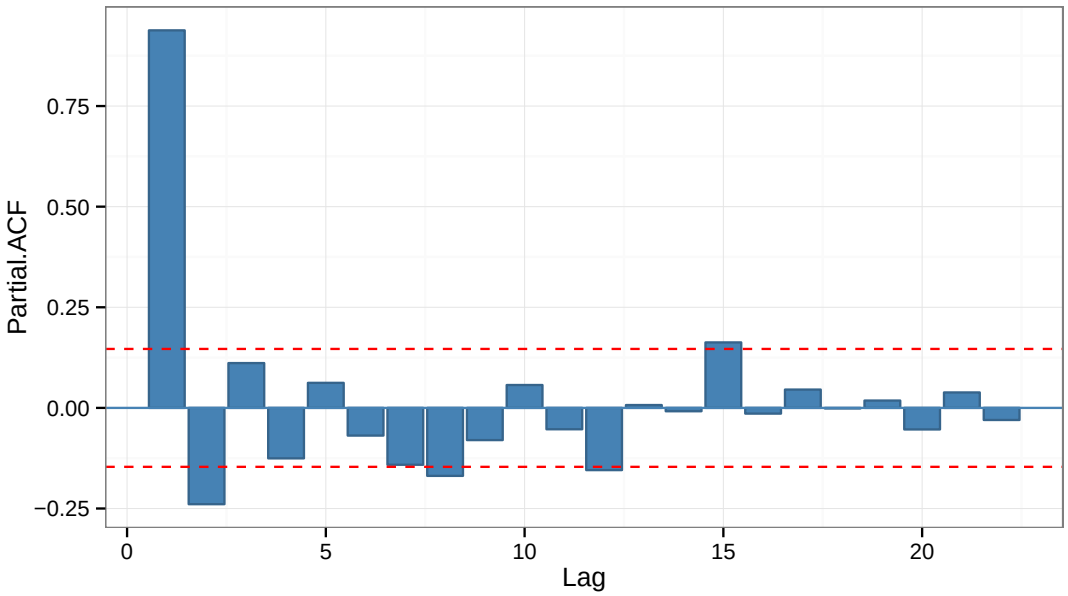


Fig 2.14: Partial autocorrelation function for Exchange Rate of NOK per Euro. The red dashed line denotes the 95% level of significance.

2.8 Effect of Crisis period

Chapter 3

Models and Methods

3.1 A statistical Model

A statistical model describes the relationship between a cause and its effect. A vector $\mathbf{y}$ contains n number of responses. Let $\mathbf{X}$ be a $n \times p$ matrix whose columns are predictor variables and each of them have n observations. These variables in $\mathbf{X}$ can affect $\mathbf{y}$ so, the relationship between $\mathbf{X}$ and $\mathbf{y}$ can be written in a functional form as,

$$\mathbf{y} = f(\mathbf{X}) + \epsilon \tag{3.1}$$

where, ϵ is a vector of unknown errors usually referred as 'white noise' when dealing with time-series data which is assumed to have zero mean, constant variance and no autocorrelation.

3.2 Linear Regression Model

The linear regression model with a single response ($\mathbf{Y} = y_{t1}, y_{t2}, \ldots, y_{tp}$) and p predictor variable $X_1, X_2, \ldots, X_p$ has form,

$$\underbrace{\mathbf{Y}}_{\text{Response}} = \underbrace{\beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \ldots + \beta_p X_{tp}}_{\text{Mean Response explained by predictors only}} + \underbrace{\epsilon}_{\text{Error Term}} \tag{3.2}$$

The model - 3.2 is linear function of $p+1$ unknown parameters $\beta_\circ, \beta_1, \beta_2, \ldots, \beta_p$ which is generally referred as regression coefficients. In matrix notation, equation-(3.2) becomes,

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times(p+1)}{\mathbf{X}} \underset{(p+1)\times 1}{\boldsymbol{\beta}} + \underset{n\times 1}{\boldsymbol{\epsilon}} \tag{3.3}$$

3.2.1 Least Square Estimation

The estimate of the unknown parameter vector β in (3.3) is obtained by minimizing the sum of square of residuals, The sum of square of residuals is,

$$\boldsymbol{\epsilon}^t\boldsymbol{\epsilon} = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) \tag{3.4}$$

On minimizing eq - 3.4, we get the OLS estimate of $\boldsymbol{\beta}$ as,

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \tag{3.5}$$

For ordinary least square estimation, following basic assumptions (Wooldridge, 2012) are required,

1. Linear in parameter

2. Absence of Multicollinearity

3. No correlation between Error terms and predicotor variable, mathematically,

$$E(\epsilon_i|\mathbf{X}) = 0, t = 1, 2, \ldots, n$$

   The equation implies that the error term at time t should be uncorrelated with each explanatory variable in every time period

4. Homoskedastic Error terms, i.e,

$$\text{var}(\boldsymbol{\epsilon_t}|\mathbf{X}) = \text{var}(\epsilon_\mathbf{t}) = \sigma^2\mathbf{I}$$

5. No serial correlation (autocorrelation) in error terms, i.e,

$$\text{corr}(\boldsymbol{\epsilon_t}, \epsilon_\mathbf{s}) = 0, \forall t \neq s$$

For Hypothesis testing and inference using t and F test, an additional assumption of normality is needed, i.e

$$\epsilon_t \sim N(\mathbf{0}, \sigma^2\mathbf{I})$$

Under the assumption from 1 to 5, the OLS estimate obtained from eq-3.5 is best linear unbiased estimator of β.

3.2.2 Prediction

Using $\hat{\boldsymbol{\beta}}$ obtained in eq-3.5, following two matrices can be obtained,

$$\text{Predicted Values:}\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X^t X})^{-1}\mathbf{X^t Y} \tag{3.6a}$$

$$\text{Residuals:}\hat{\boldsymbol{\epsilon}} = \mathbf{Y} - \hat{\mathbf{Y}} = [\mathbf{I} - \mathbf{X}(\mathbf{X^t X})^{-1}\mathbf{X^t}]\mathbf{Y} \tag{3.6b}$$

Here eq-3.6a gives predicted values of $\mathbf{Y}$ which on subtracting from observed value give the predicted error terms as is presented in eq-3.6b. Eq-3.6a can also be written as,

$$\hat{\mathbf{Y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{HY} \tag{3.7}$$

Here, $\mathbf{H}$ is called Hat matrix and is the orthogonal projection of y onto the space spanned by the columns of $\mathbf{X}$.

3.3 Variable selection

Although including many variables in the model can add information, they are also the source of unnecessary noise. In addition, many variables in a model is also the cause of multicollinearity. So, a model that is simple yet contain usefull information is always diserable. Variable selection is intended for selecting best subset of predictor variables. Some of the criteria for variable selection as discribed in (Weisberg, 2005) are discussed below:

3.3.1 Criteria for variable selection

Suppose X_s is selected set of variable which gives the predicted output of,

$$\hat{Y} = E\left(Y|X_s - x_s\right) = \beta_s' x_s \tag{3.8}$$

If X_s misses important variables, the residual sum of squares of fitted model in equation-3.8 will be larger than the full model. Lack of fit for selecting the set X_s is measured by its Error sum of square.

Model statistic Approach

When a model is fitted, various statistics such as R^2, R^2-adj, F-statistic are obtained which measures the quality of that model. Based on these statistic, a model is selected as better than other.

Information Criteria

Another common criterion, which balances the size of the residual sum of squares with the number of parameters in the model (Johnson and Wichern, 2007, p. 386), for selecting subset of predictor variable is AIC (*Akaike Information Criterion*). It is given as,

$$\text{AIC} = n\log(\text{RSS}_s/n) + \text{k} \tag{3.9}$$

where, RSS=Residual Sum of Square, n =number of observation and k =Number of variables included in the model

A model with smaller value of AIC obtained from eq-3.9 is better better than other with larger AIC. An alternative to AIC is its Bayesian analogue, also

known as Schwarz or Bayesian information criteria. Bayesian Information Criteria provides between model complexity and lack of fit. Smaller value of BIC is better.

$$\text{BIC} = n\log(\text{RSS}_s/n) + \text{k}\log(n) \tag{3.10}$$

A third criterion that balances the complexity and lack of fit of a model is Mallows C_p (Mallows, 1973), where the subscript p is the number of variables in the candidate model. The formula for this statistic is given in equation-3.11,

$$\text{Mallows } C_p = \frac{\text{RSS}}{\hat{\sigma}^2} + 2kn \tag{3.11}$$

Where, $\hat{\sigma}^2$ is from the full model. A plot of C_p vs k for each subset of predictors indicate models that predict the responses well. Better models usually lie near the 45° line of the plot.

3.3.2 Computational procedure for variable selection

When a model is large, fitting all possible subsets is not feasible. Furnival and Wilson (1974) suggested several algorithm to calculate residual sum of square of all possible regression called leap and bound technique which has been widely implemented in statistical software. However, this method is not appropriate for criteria based on model statistic where Stepwise methods can be used. Stepwise methods has three basic variation (Weisberg, 2005, p. 221).

Forward selection procedure

Model is started without any variable and in each step a variable is added and the model is fitted. The variable is left in the model if the subset minimizes

the criterion of interest . Similar process is repeated for other predictor variable.

Backward elimination procedure

This process is like the reverse of Forward selection procedure. In this process, the model is fitted with all the predictor variable and variables are removed one at a time except those that are forced to be in the model. The model is examined against the considered criteria. Usually, the term with smallest t-value is removed since this give rise to the residual sum of square.

Stepwise procedure

This combines both Forward selection procedure and Backward elimination procedure. In each step, a predictor variable is either deleted or added so that resulting model minimizes the criterion function of interest.

3.4 Principal Component Analysis

The purpose of PCA is to express the information in $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ by a less number of variables $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_q); q < p$ called principal components of $\mathbf{X}$ (Martens and Naes, 1992). These principal components are orthogonal and linearly uncorrelated. Since they are computed from the linear combinations of $\boldsymbol{X}$ variables, the variation in $\mathbf{X}$ variables are compressed in first few principal components. In other words, the first principal components is the direction along which the $\boldsymbol{X}$ variables have the largest variance (Massart, 1998). In this situation, the multicollinearity in $\boldsymbol{X}$ is not a problem any more.

The principal components can be performed on Covariance or Correlation matrix. If the variables are of same units and their variances do not differ much, a covariance matrix can be used. However the population correlation matrix is unknown, its estimate can be used. In this thesis, sample correlation matrix is used to compute sample principal components. Construction of principal components requires following steps,

1. Estimate the correlation matrix $\boldsymbol{A}$ of $\boldsymbol{X}$ as,

$$\text{corr}(\mathbf{X}) = (\text{diag}(\mathbf{\Sigma}))^{-\frac{1}{2}} \, \Sigma \, (\text{diag}(\mathbf{\Sigma}))^{-\frac{1}{2}} \tag{3.12}$$

   Using sample observation, equation-3.12 can be estimated as,

$$\mathbf{A} = \text{corr}(\mathbf{X}) = (\text{diag}(\mathbf{S}))^{-\frac{1}{2}} \, \mathbf{S} \, (\text{diag}(\mathbf{S}))^{-\frac{1}{2}} \tag{3.13}$$

   Where $\mathbf{S}$ is the sample estimate of covariance matrix $\mathbf{\Sigma}$,

$$\mathbf{S} = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}[\mathbf{X}]) \, (\mathbf{X} - \mathbb{E}[\mathbf{X}])^{\mathrm{T}} \right] \tag{3.14}$$

2. Calculate eigenvalue and eigenvector of the correlation matrix obtained in eq-3.12. An eigenvalue $\mathbf{\Lambda}$ of a square matrix $\mathbf{A}$ of rank p is a diagonal matrix of order p which satisfies,

$$\mathbf{AE} = \mathbf{E\Lambda} \tag{3.15}$$

   where,

$$\mathbf{\Lambda} = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_p) \tag{3.16}$$

In PCA these eigenvalues are arranged in descending order, i.e. $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. For each eigen values there is an eigenvector. Let $\mathbf{E} = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p)$ be the matrix of eigen vector so that the correlation matrix $\mathbf{A}$ can be decomposed and expressed as,

$$\mathbf{A} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{-1} = \mathbf{E}\mathbf{\Lambda}\mathbf{E}^{T} \tag{3.17}$$

Equivalently, $|\mathbf{A} - \lambda_i \mathbf{I}_n|\mathbf{E} = 0$ which can only be realized if $\mathbf{A} - \lambda_i \mathbf{I}_n$ is singular, i.e.,

$$|\mathbf{A} - \lambda_i \mathbf{I}_n| = 0 \tag{3.18}$$

Eq-3.18 is called the characteristic equation where, $\mathbf{A}$ is the correlation matrix obtained from eq-3.12. The root of the equation is called eigenvalues (Seber, 2008) and the vector $\mathbf{E}_i$ is called eigenvector corresponding to the eigenvalue λ_i. The eigenvector obtaind from eq-3.15 are then normalized, i.e. $||\mathbf{E}_i||^2 = 1$.

3. Since, the variation explained in data are accumulated in first few principal components, only k eigenvalues are considered. The corresponding eigenvectors of those eigenvalues is called projection matrix. The projection matrix is,

$$\mathbf{P} = \begin{pmatrix} \mathbf{E}_1^T & \mathbf{E}_k^T & \dots & \mathbf{E}_k^T \end{pmatrix}^T \tag{3.19}$$

The projection matrix in eq-3.19 projects the datamatrix into lower dimensional subspace $\mathbf{Z}_i$. i.e.,

$$\mathbf{Z} = \mathbf{PX} \tag{3.20}$$

The columns vectors of matrix Z obtained from 3.20 are the orthogonal projections of data matrix $\mathbf{X}$ into k dimensional subspace. These components are the linear combination of the rows of matrix $\mathbf{X}$ such that the most variance is explained by the first column vector of $\mathbf{Z}$ and second one has less variance than the first one and so on. Here,

$$\text{var}(\mathbf{Z}_i) = \lambda_i \text{ and}$$
$$\text{cov}(\mathbf{Z}_i\mathbf{Z}_j) = 0 \text{ for } i \neq j$$

3.5 Principal Component Regression

The components of Principal Component Analysis (PCA) accumulate the variation in predictor variables on first few components. A linear regression fitted with only those components can give a similar results as the full linear model. However, Jolliffe (1982) in his paper "A note on the use of principal components in regression", has given many examples taken from different papers of various fields where the components with low variance are also included in regression equation

in order to explain most variation in the response variable. Following are the steps to perform Principal Component Regression. These steps are based on the paper "A comparison of partial least squares regression with other prediction methods" by Yeniay and Goktas, 2002.

1. First principal components are obtained for $\mathbf{X}$ as explained in section-3.4. The PCs obtained are orthogonal to each other.

2. Suppose m PC which are supposed to influence the response are taken and a regression model is fitted as,

$$\mathbf{Y} = \mathbf{Z_m}\alpha_m + \epsilon \tag{3.21}$$

3. Here, $\alpha_m = \left(\mathbf{Z}_m^T\mathbf{Z}_m\right)^{-1}\mathbf{Z}_m^T\mathbf{Y}$ are the coefficients obtained from OLS methods. Using this alpha, one can obtain the estimate of $\boldsymbol{\beta}$ as,

$$\hat{\boldsymbol{\beta}}_{\text{PCR}} = \boldsymbol{P}\left(\boldsymbol{P}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{P}\right)^{-1}\boldsymbol{P}^T\boldsymbol{X}^T\boldsymbol{Y} \tag{3.22}$$

   Here, $\boldsymbol{P}$ is a projection matrix defined in equation-3.19.

Since, PCR includes only m componets, the estimate obtained are biased. ;The number of components m can be chosen by cross-validation the prediction mean squared error (RMSEP). If all the components are included in the model, estimates obtained from PCR, i.e. $\boldsymbol{\beta}_{\text{PCR}}$ are identical to the estimates of OLS ($\boldsymbol{\beta}_{\text{OLS}}$).

3.6 Partial Least Square Regression

Partial Least Square Regession (PLS) is relatively new methods and it can be used for both univariate and multivariate regression. It constructs a new set of variables called latent variable (or factor or components) from the linear combination of predictor variables $X_1, X_2, \ldots, X_n$ (Garthwaite, 1994) as in the case of principal components, however PCR construct components (factors) maximizing the variation of data matrix(X) while PLS construct them using the variation in both X and Y (Yeniay and Goktas, 2002). The intension of PLS is to create latent variables (components) that caputre most of the information in the X variables that is useful for predicting $Y_1, Y_2, \ldots, Y_p$, while reducing the dimensionality of the regression problem by using fewer components than the number of X-variables (Garthwaite, 1994). Partial least square regression can be performed using following steps. These steps are adapted from the paper "PLS-regression: a basic tool of chemometrics" from Wold, Sjöström, and Eriksson (2001). The X and Y matrices are column centered for the ease of computation.

1. PLS estimates the latent variables also called X-scores denoted by $t_a, (a = 1, 2, \ldots, A)$, where A is the number of Components a model has considered. These X-scores are used to predict both X and Y, i.e. both X and Y are assumed to be modelled by the same latent variable. The X-scores are estimated as linear combination of original variables with the coefficients $W(w_{ka})$ as in equation-3.23, i.e,

$$t_{ia} = \sum_{k=1}^{p} W^*_{ka} X_{ik} \quad (\boldsymbol{T} = \boldsymbol{X}\boldsymbol{W}^*) \tag{3.23}$$

Where, $\boldsymbol{W}^*$ is a vector of weights w_a^* of $\boldsymbol{X}$. It is obtained as in equation-3.24 below as a normalized coefficients obtained on regressing X on a column of Y.

$$\boldsymbol{W}^* = \frac{\boldsymbol{X}^t \boldsymbol{y}^{(i)}}{\|\boldsymbol{X}^t \boldsymbol{y}^{(i)}\|} \tag{3.24}$$

Here, $\boldsymbol{y}^{(i)}$ is any column of response matrix $\boldsymbol{Y}$.

2. The x-scores (T) are used to summarize $\boldsymbol{X}$ as in the equation-3.25. Since the summary of $\boldsymbol{X}$ explained most of the variations, the residuals ($\boldsymbol{E}$) are small.

$$X_{ik} = \sum_a t_{ia} P_{ak} + e_{ik}; \quad (\boldsymbol{X} = \boldsymbol{T}\boldsymbol{P}' + \boldsymbol{E}) \tag{3.25}$$

A similar setup can be used to have the summary for Y-matrix as in equation-3.26,

$$Y_{im} = \sum_a u_{ia} q_{am} + g_{im}; \quad (\boldsymbol{Y} = \boldsymbol{U}\boldsymbol{Q}' + \boldsymbol{G}) \tag{3.26}$$

where, $\boldsymbol{U} = \boldsymbol{Y}\boldsymbol{Q}$ and $\boldsymbol{Q} = \boldsymbol{T}^t\boldsymbol{Y}$

3. The X-scores ($\boldsymbol{T}_\circ$) are also good predictor of $\boldsymbol{Y}$, i.e.,

$$y_{im} = \sum_a q_{ma} t_{ia} + f_{im} \quad (\boldsymbol{Y} = \boldsymbol{T}\boldsymbol{C}^t + \boldsymbol{F}) \tag{3.27}$$

Here, $\boldsymbol{F}$ is the deviation between the observed and modelled response.

4. **Coefficients Estimates:**

   Equation(3.27) can also be written as,

$$\begin{aligned} y_{im} &= \sum_a q_{ma} \sum_k w^*_{ka} x_{ik} + f_{im} \\ &= \sum_k b_{mk} x_{ik} + f_{im} \end{aligned}$$

   In matrix notation this can be written as,

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{W}^*\boldsymbol{C}^t + \boldsymbol{F} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{F} \tag{3.28}$$

   Thus, the estimates of PLS coefficients are obtained as,

$$\hat{b}_{mk} = \sum_a q_{ma} w^*_{ka} \tag{3.29}$$

$$i.e., \boldsymbol{B}_{\text{PLS}} = \boldsymbol{W}^*\boldsymbol{C}^t \tag{3.30}$$

Above process is repreated for each components (a), the matrix $\boldsymbol{X}$ and $\boldsymbol{Y}$ are "deflated" by subtracting their best summaries ($\boldsymbol{T}\boldsymbol{P}^t$ for $\boldsymbol{X}$ and $\boldsymbol{Q}\boldsymbol{C}^t$ for $\boldsymbol{Y}$). The Residuals obtained are used as new $\boldsymbol{X}$ and $\boldsymbol{Y}$ in the computation process for new component. However, the deflation of $\boldsymbol{Y}$ is not necessary since the result is equivalent with or without the deflation (Wold, Sjöström, and Eriksson, 2001, p. 5).

Various algorithm exist to perform PLS regression among which NIPLS and SIMPLS are in fashion. This thesis has opted NIPLS (Nonlinear Iterative Partial Least Square) regression which is performed by `oscores` method of `pls` package in R. Appendix - D shows the flowchart for the algorithm. In the algorithm, the

first weight vector ($\boldsymbol{w}_1$) is the first eigenvector of the combined variance-covariance matrix $\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t\boldsymbol{X}$ and the following weight vectors are computed using the deflated version. Similarly, the first score vector ($\boldsymbol{t}_1$) is computed as the first eigenvector of $\boldsymbol{X}\boldsymbol{X}^t\boldsymbol{Y}\boldsymbol{Y}^t$ and the following x-scores uses the deflated version of the matrices.

3.7 Ridge Regression

When the minimum eigenvalue of $\boldsymbol{X}^t\boldsymbol{X}$ matrix is very much smaller than unity (i.e. $\lambda_{\min} << 1$), the least square estimate obtained from equation-3.5 are larger than average (Marquardt and Snee, 1975). Estimates based on $[\boldsymbol{X}^t\boldsymbol{X} + \lambda\boldsymbol{I}_p], \lambda \geq 0$ rather than $\boldsymbol{X}^t\boldsymbol{X}$ can solve these problems. A.E. Hoel first suggests that to control instibility of the least square estimate, on the condition above, can be;

$$\begin{aligned}\hat{\boldsymbol{\beta}}^*_{\text{ridge}} &= \left[\boldsymbol{X}^t\boldsymbol{X} + \lambda\boldsymbol{I}\right]^{-1}\boldsymbol{X}^t\boldsymbol{Y};\ \lambda \geq 0 \\ &= \boldsymbol{W}\boldsymbol{X}^t\boldsymbol{Y} \end{aligned} \tag{3.31}$$

The analysis build around equation-3.31 is called "ridge equation". The relationship of ridge estimate with ordinary least square is,

$$\begin{aligned}\boldsymbol{\beta}_{\text{ridge}} &= \left[\boldsymbol{I}_p + \lambda\left(\boldsymbol{X}^t\boldsymbol{X}\right)^{-1}\right]^{-1}\hat{\beta}_{\text{OLS}} \\ &= \boldsymbol{Z}\hat{\boldsymbol{\beta}}_{\text{OLS}} \end{aligned} \tag{3.32}$$

Here, as $\lambda \to 0, \hat{\beta}_{\text{ridge}} = \hat{\beta}_{\text{OLS}}$ and $\lambda \to \infty, \hat{\beta}_{\text{ridge}} = 0$ Further, the hat matrix for Ridge regression is given as,

$$\boldsymbol{H}_{\text{ridge}} = \boldsymbol{X}\left(\boldsymbol{X}^t\boldsymbol{X} + \lambda\boldsymbol{I}\right)^{-1}\boldsymbol{X}^t \tag{3.33}$$

All the theory behind Ridge Regression described above are cited from "Ridge regression: Biased estimation for nonorthogonal problems" by Hoerl and Kennard (1970).

3.8 Comparision Criteria

After fitting models with various methods, it becomes necessary to test their validity for their results to be trusted. Models react differently for the new information during prediction as the quality of model hightly depends on their estimates. Since the purpose of this thesis is to compare different models on the quality of their prediction, models are compared on the basis of their (a) Goodness of fit and (b) Predictability.

3.8.1 Goodness of fit

A model is assumed to follow some hypothetical state of being ideal. Setting up this state as null hypothesis ($H_\circ$), in many situations, the test of goodness of fit for a model construct an alternative hypothesis simply stating that the model gives little or no information about the distribution of the data. However in other situation, such as testing for no effect of some specific variable in the model, rejection of

$H_\circ$ indicate that the variable is useful in the model (D'Agostino, 1986, p. 1). A goodness of fit for a model depends on many aspects such as,

Residual obtained after the model fit

Residuals obtained from the fitted model are assumed to be random and normal considering that no useful information are still content on them.

Outlier

Outliers can distort the analysis toward unintential direction creating false estimates. Models without such outliers are considered better.

Variance explained by the model

The variance explained by the model is generally measured by R^2 or R^2 adj in linear models. More the variation contained in the data is explained by the model, better the model is considered. In the case of PLS and PCR, the residuals contains very little information left on the ignored components.

Relative value of Information Criteria such as AIC and BIC

AIC (Akaike information criterion) and BIC (Bayesian information criterion or Schwarz criterion) measures relative quality of model. Although, it is not an absolute measure of the model quality, helps to select a better model among others. AIC is defined as in equation - 3.34 which is free from the ambiguities present in the convenential hypothesis testing system (Akaike, 1974).

$$\text{AIC} = (-2)\log(\mathcal{L}) + 2(k) \tag{3.34}$$

where, $\mathcal{L}$ = maximum likelihood and k = number of independently adjusted parameters within the model For least square case, above formula resembles to equation - 3.9 (Hu, 2007).

3.8.2 Predictability

Prediction is highly influenced by the model in used. So, prediction strongly depends on the estimates of a model. False and unstable estimates makes the prediction poor and unreliable. On one side, providing more information (variable) can well train the model resulting more precise prediction. On the other hand, overfitting, which attempts to explain idiosyncrasies in the data, leads to model complexity reducing the predictive power of a model. In the case of PLS and PCR, adding more components results in including noise in the model.

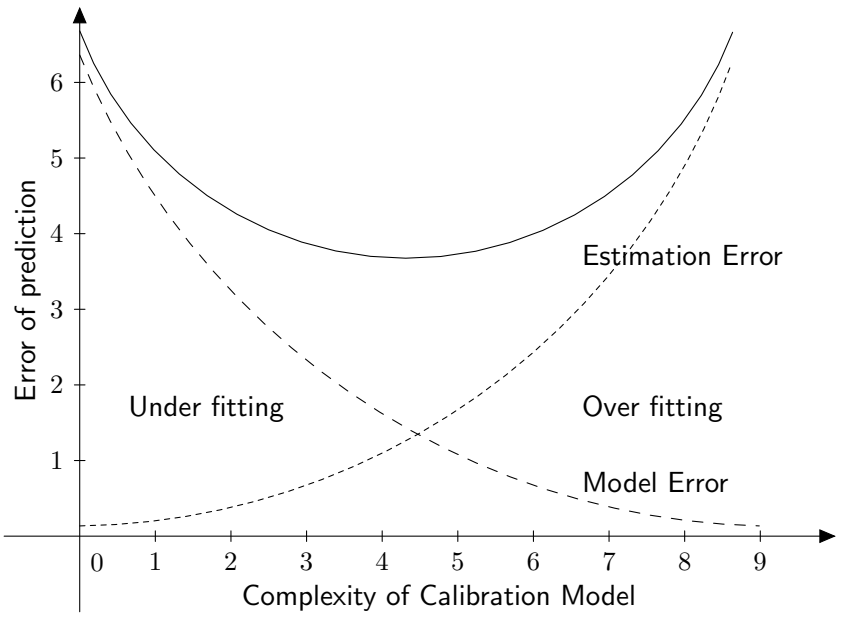


Fig 3.1: Model Error - Estimation Error and Prediction Error

The relationship between the model complexity and the prediction error is presented in figure-3.1 with the case of under-fitting and over-fitting of a model.

Furthermore, a model exhibits an *external validity* if it closely predicts the observations that were not used to fit the model parameters (Lattin, Carroll, and Green, 2003, p. 72). An overfitted model model fails to perform well for those observation that are not included during model parameter estimation. The dataset in this thesis is divided into two parts. The first part includes the observations from Jan 2000 to December 2012 and the second one includes observation onward till November 2014. A cross-validation approach is utilized on the first set of observation to train the model. The model is used to predict the the exchange rate of NOK per Euro from the predictors of the second set of observations. Figure - 3.2 shows the procedure adopted for prediction in this thesis.

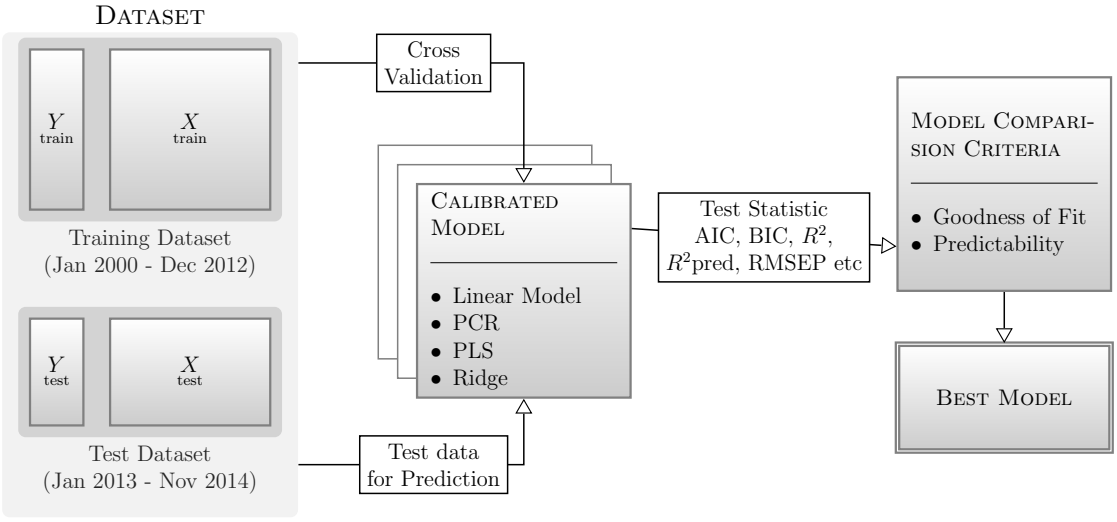


Fig 3.2: Procedure adopted in the thesis for model comparision.A cross-validation technique is used to validate the trained dataset. The trained model is used to predict the test response from with prediction errors are obtained.

Cross-Validation

There are various cross-validation techniques among which two were described below;

K-Fold Cross-validation:

The dataset are splitted into k equal parts. For each $i = 1, 2, \ldots, k$, a model is fitted leaving out the i^{th} portion. A prediction error is calculated for this model. The process is repeated for all i. THe prediction error for K-fold cross validation is obtained by averaging the prediction error of each of the model fitted.

Leave-one-out cross validation:

This is a special case of $k-$ fold cross-validation where $k = n$ (number of observation), i.e, each time one observation is removed and the model is fitted.

Prediction Error

Prediction of a model becomes precise if the error is minimum. Models can be compared according to their predictability. Understanding of different measures of prediction error is necessary acknoledge their predictability and eventually perform model comparison.

Root Mean Square Error (RMSE)

RMSE is the measure of how well the model fit the data.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{3.35}$$

Where,

$\hat{y}_i$ are predicted values for y_i and

n is the number of observation

Root Mean Square Error of Cross-Validation (RMSECV)

RMSECV gives the models ability to predict new samples that were not present in the model during calibration. It is obtained as,

$$\text{RMSECV} = \sqrt{\frac{\text{PRESS}}{n}} \tag{3.36}$$

Where,

$$\text{PRESS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)}\right)^2 \tag{3.37}$$

In the special case of leave one out cross validation, i represents each sample.

R-squared for Prediction

R-squared for prediction is analogus to the R-sq in the case of model estimation. In the case of cross-validation, it is also denoted by Q^2. It is obtained by subtracting the ratio of PRESS obtained from equation-3.37 to Total sum of square from 1. i.e,

$$R^2_{CV} = Q^2 = 1 - \frac{\text{PRESS}}{\text{TSS}} = 1 - \frac{\sum_{i=1}^{n} \left(y_i - \hat{y}_{(i)}\right)^2}{\sum_{i=1}^{n} \left(y_i - \bar{y}\right)^2} \tag{3.38}$$

Here, $Q^2 < 1$ and when prediction is very bad, PRESS may exceed TSS resulting negative value suggesting that the average value is better than the prediction using the model.

Chapter 4

Data Analysis

This chapter will present the analysis report obtained for different models considered in chapter-3. The analysis process includes following series of steps,

1. The model is trained from the observation from the trained period (Jan 2000 - Dec 2012) through cross validation.

2. Prediction on the average monthly exchange rate of Euro vs Norwegian Krone is made for the test period (Jan 2013 - Nov 2014)

3. Compare them on the basis of criteria discussed in section-3.8

The summary report of the variables that are in use during the analysis are in table (4.1),

Table 4.1: Summary Report of all the variables used in this report

	min	median	max	mean	stdev
PerEURO	7.30	8.00	9.40	8.03	0.37
KeyIntRate	1.25	2.25	7.00	3.39	2.05

Continued on next page

Table 4.1: Summary Report of all the variables used in this report

	min	median	max	mean	stdev
LoanIntRate	2.25	4.00	9.00	4.87	2.32
EuroIntRate	-0.01	2.07	5.06	2.10	1.57
CPI	104.10	118.60	137.90	120.71	9.65
OilSpotPrice	16.70	86.29	209.29	87.64	50.80
ImpOldShip	0.00	103.00	8099.00	229.56	641.49
ImpNewShip	0.00	377.00	3011.00	556.02	629.05
ImpOilPlat	0.00	0.00	8914.00	145.68	863.83
ImpExShipOilPlat	19381.00	34812.00	51660.00	33610.13	8437.76
ExpCrdOil	13125.00	22630.00	37132.00	22771.27	4676.88
ExpNatGas	2457.00	11341.00	26420.00	11883.05	6532.83
ExpCond	0.00	751.00	2305.00	768.94	452.03
ExpOldShip	0.00	213.00	1948.00	342.45	358.67
ExpNewShip	0.00	211.00	2326.00	299.54	363.54
ExpOilPlat	0.00	0.00	3069.00	63.65	364.35
ExpExShipOilPlat	34060.00	62457.00	90063.00	59912.43	14947.02
TrBal	10853.00	25001.00	48141.00	26076.72	8257.33
TrBalExShipOilPlat	11493.00	25331.00	47250.00	26302.36	8191.34
TrBalMland	-18150.00	-9308.00	-2766.00	-9120.96	3167.78
ly.var	7.30	8.00	9.40	8.03	0.37
l2y.var	7.30	8.00	9.40	8.03	0.37
l.CPI	103.60	118.50	137.80	120.52	9.65

The correlation between response variable and predictor variable helps us to determine their relationship. Figure -(4.1) shows that only few of the predictor variables have significant correlation with response variable. In the figure first and second lagged response variable have strong correlation with response while most of the others have low (weak) correlation. Although, being weak correlation, many of them are statistically significant. According to the paper "Interpretation of the correlation coefficient: a basic review" by Taylor (1990), the significance of the low correlation, which would have little practical importance, is due to the large number of observation. According to him, a correlation coefficients is an abstract

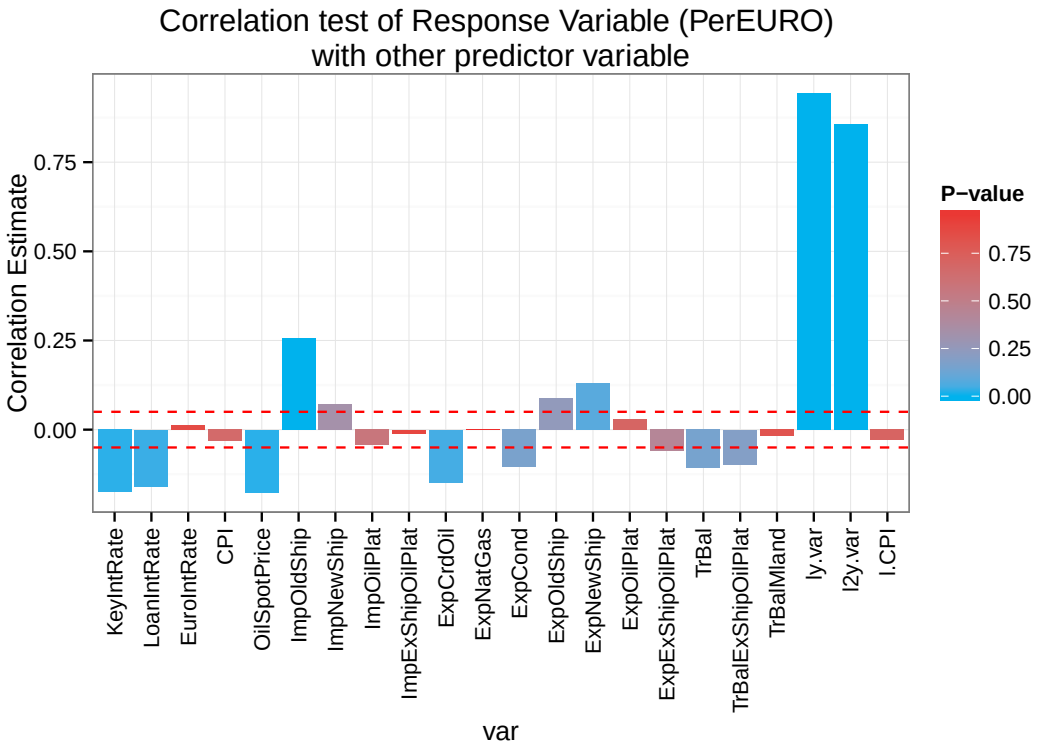


Fig 4.1: The bars represents the correlation between response variable (`PerEURO`) and other predictor variable. The bars are shaded with the p-value for their signifance test performed by `cortest` function. The red horizontal line is the critical value at 5 percent level of significance.

measure which does not give direct precise interpretations. A more useful measure can be obtained during the model fitting.

4.1 Multiple Linear Regression

The functional form for determining exchange rate of Krone per Euro can be written as,

$$\begin{aligned} \texttt{PerEURO} &= f(\texttt{interest Rate}, \texttt{Trade}, \texttt{Price}, \texttt{Expectation}) + \texttt{Error} \\ &= \alpha_0 + \alpha_1(\texttt{interest Rate}) + \alpha_2(\texttt{Trade}) \\ &\quad + \alpha_3(\texttt{Price}) + \alpha_4(\texttt{Expectation}) + \texttt{Error} \end{aligned} \tag{4.1}$$

Where, f is a linear function of regression coefficients $\boldsymbol{\alpha}$.

In equation-4.1, interest Rate include both interest rate of Norway and European Central Bank. Trade encorporates import, export and trade balance of Norway. Similarly, Price include Consumer price index and Oil price and Expectation variables are created with the lagged value of response, i.e. `PerEURO`. The observation for all the model fitting from this point onward are from the training dataset, i.e. from Jan 2000 to Dec 2012. The detail explanation for the variables are in Appendix A. As described in section - 3.2, the linear model is fitted. The results shows that variables in table-4.2 has significant effect on the Euro vs Norwegian Krone exchange rate.

Table 4.2: Variables significant at $\alpha = 0.05$ while fitting linear model

	Estimate	P-value
`EuroIntRate`	0.0599	0.0307
`ly.var`	1.0907	0.0000
`l2y.var`	-0.2358	0.0044

Since, there are a lot of variables that are not significant at 5% level of significance in the fitted linear model. So, it is suitable to use variable selection procedure as described in section-3.3.

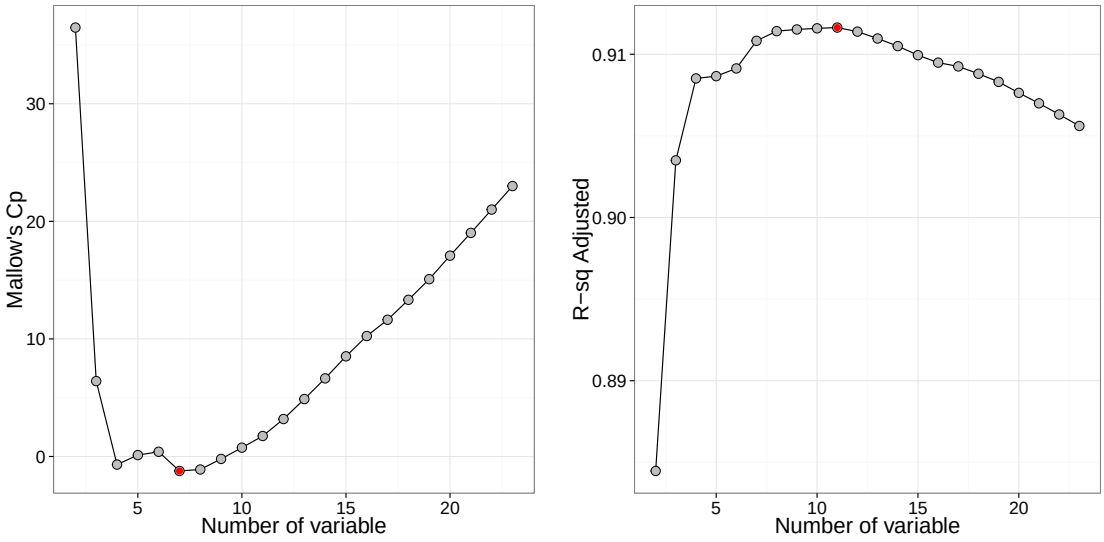
4.2 Variable Selection Procedure

Variable selection is based on criteria to choose best model form the possible subset. Linear model fitted above when exposed to the those criteria from subsection-3.3.1 for choosing best subset, following results are obtained.

4.2.1 Model selection using Mallows C_p and R^2 adjusted

The best subset is selected using (a) Mallows C_P and (b) Adjusted R^2. The number of variable vs these two criteria are plotted in figure-4.2. The plot in fig-4.2a, shows that including 7 variables which the algorithm has selected, minimize the Mallow's C_p while fig-4.2b suggest to include 11 variables including intercept to maximize the adjusted R^2.

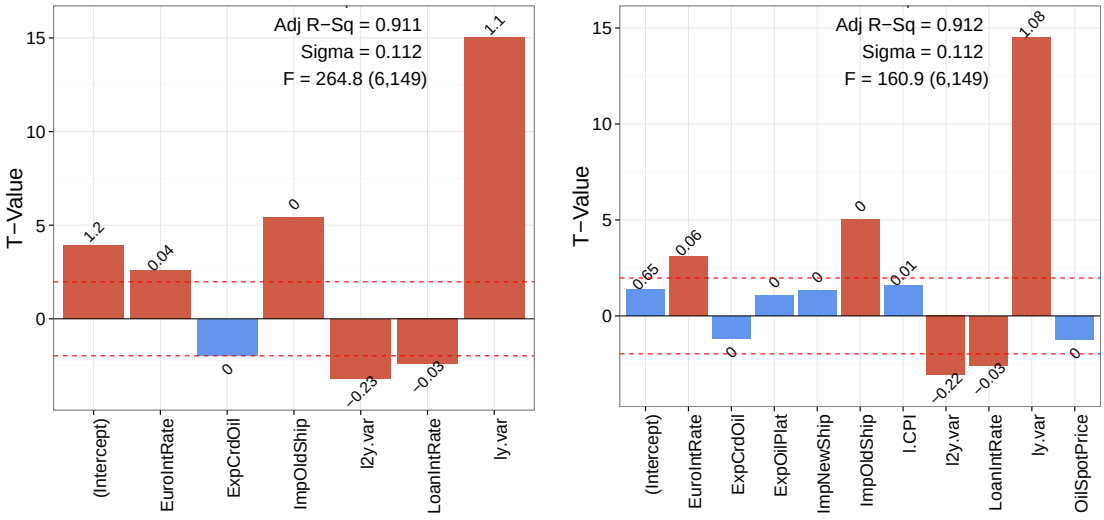
The models selected by these criteria when fitted results on few insignificant variables. The plot of the t-value in fig-4.3 has only one (for C_p criteria) and two (R^2adj criteria) are insignificant. With fewer variables than the full model, this model has described the variation almost equally as the the linear model with full variables (table-4.7).



(a) Mallows Cp vs no. of Variable

(b) R2 adjusted vs no. of variable

Fig 4.2: Number of variable against the criteria where the red dot corresponds the number of variable to acheave the criteria, i.e. minimum for Cp and maximum for R^2 adjusted



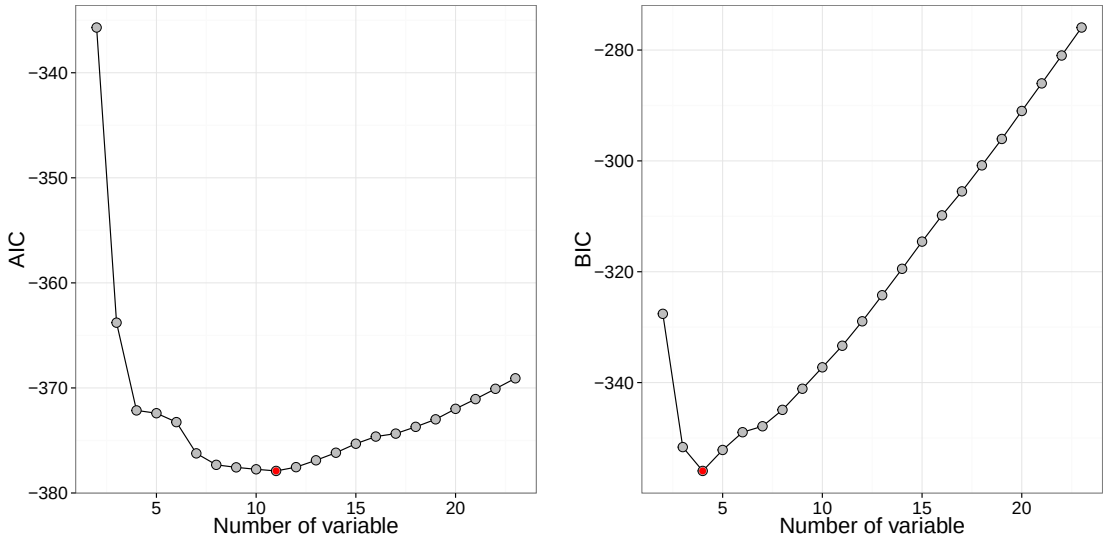
(a) Model selected from Mallows' C_p criteria

(b) Model selected from R^2 adjusted criteria

Fig 4.3: Model selected by C_p and R^2 adjusted criteria. Red and blue bars are significant and insignificant variables respectively. The estimates rounded at 2 decimals are given on top of the bars.

4.2.2 Model selection using AIC and BIC criteria

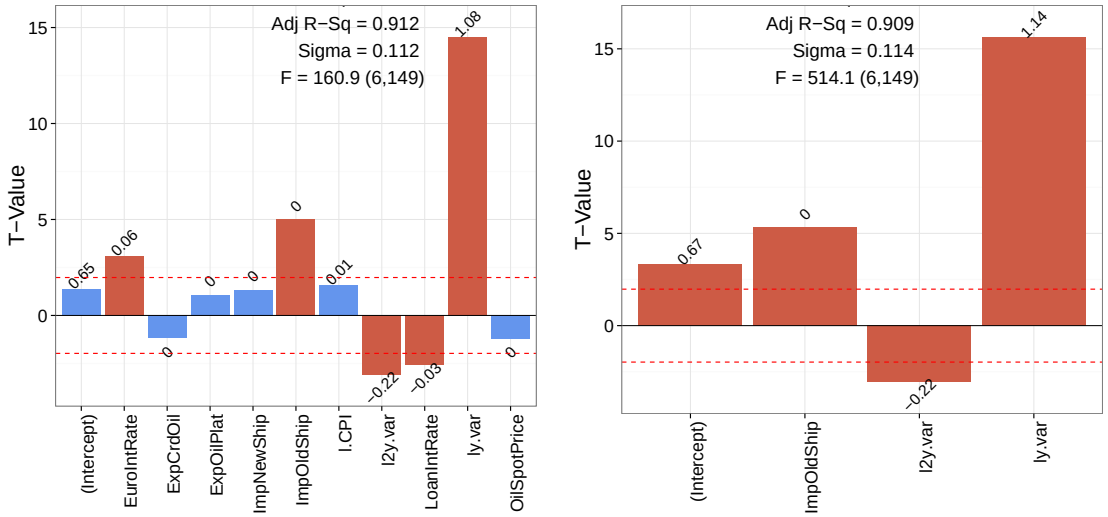
Applying AIC and BIC criteria to select best model, exhaustive search algorithm as used by `leaps` package (Lumley and Lumley, 2004) is used in this thesis. Number of variables required to minimize the information criteria is selected as guide by the plot in figure -4.4. For minimum AIC, 11 (fig-4.4a) variables are needed and for minimum BIC, 4(fig-4.4b) are needed to get the best subsetted model. The models suggested are fitted with results of few insignificant variables (fig-4.5). The summary statistic (table-4.7) shows that AIC model has larger R^2 adjusted than BIC model due to the addition of more variables.



(a) AIC vs no. of Variable

(b) BIC vs no. of variable

Fig 4.4: Number of variable against the AIC vs BIC criteria. The red dot corresponds to the number of variables that can minimize the criteria.



(a) Model selected from minimum AIC criteria

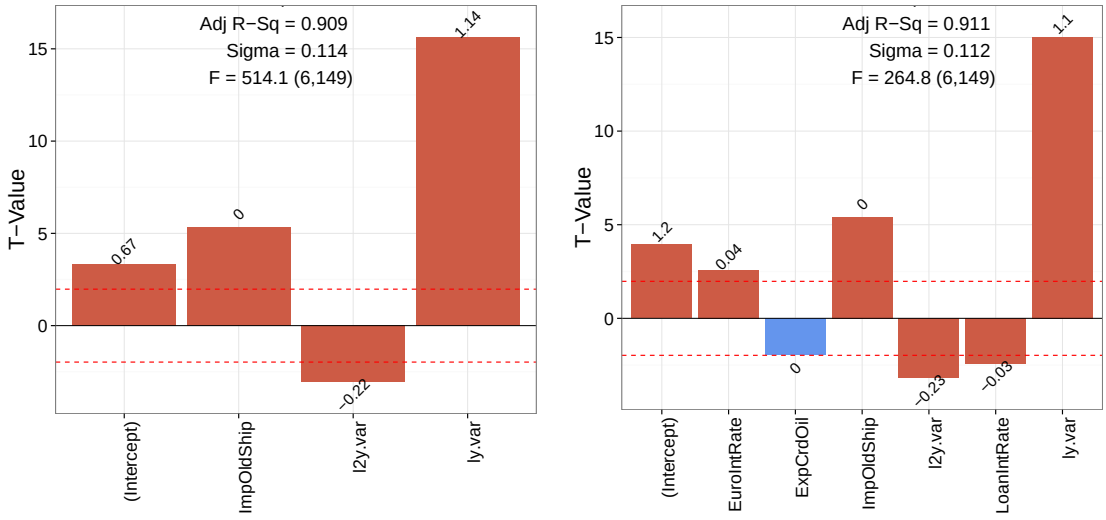
(b) Model selected from minimum BIC criteria

Fig 4.5: Best subset model selected by AIC and BIC criteria. Red and blue bars are significant and insignificant variables respectively. The estimates rounded at 2 decimals are given on top of the bars.

4.2.3 Stepwise procedures based on F-value

The models fitted in previous sub sections resulted with some insignificant variables because the criteria there was based on model statistics other than the p-value of the respective variables. The stepwise procedure based on the F test fit the model removing the insignificant variable one at a time in forward search and adding variable one at a time in backward search. The fitted results (fig-4.6) for the models fitted with backward (fig-4.6a) and forward (fig-4.6b) stepwise procedure show that all the variables are significant at 5 percent since the `alpha-to-remove` and `alpha-to-enter` criteria for the process are set at 0.05.

The models suggested by R^2 critera and AIC are same. Similary BIC and stepwise forward selection based on F-test also have suggested the same model. Despite of explaing enough variation in response, some of these models have severe



(a) Model selected from stepwise forward selection prcedure

(b) Model selected from stepwise backward elimination procedure

Fig 4.6: Best subset model selected by F-test based criteria. Red and blue bars are significant and insignificant variables respectively. The estimates rounded at 2 decimals are given on top of the bars.

multi-collinearity problem (Fig-4.7) since the VIF (Variance Inflation Factor) of some of the variables included in the model are much larger than 10 which is usually considered as rule of thumb (Oâbrien, 2007) for measuring multi-collinearity.

Multicollinearity in a model distorts the estimate and consequently distorts the prediction made by the model. An alternative solution for the multicollineary problem is using principal component related model such as PLS and PCR or one can use ridge regression as well.

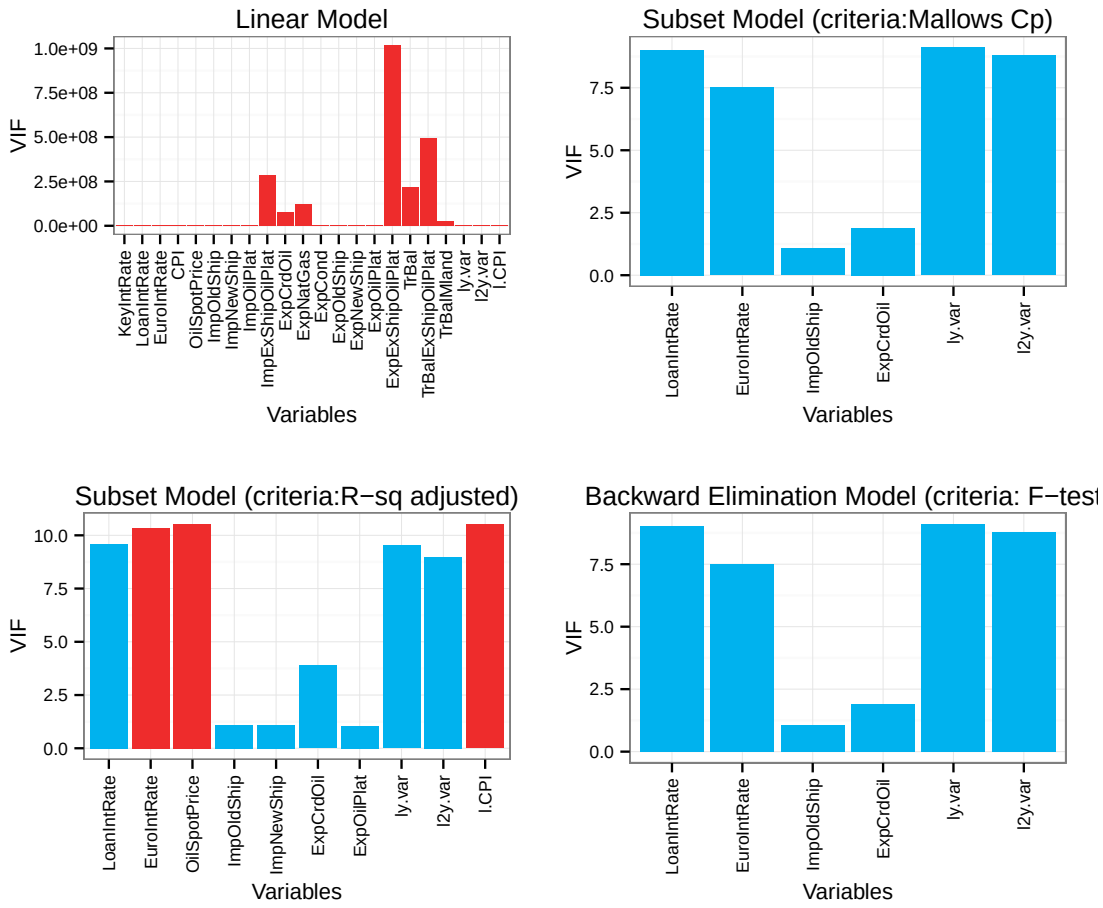


Fig 4.7: Variance Inflation Factor (VIF) of different models. The red bars represents the variables with VIF greater than 10.

4.3 Principal Component Analysis

Principal Component Analysis(PCA) creates a new set of mutually orthogonal and independent variables called components. The PCA analysis is done from full dataset (Jan 2000 - Nov 2014) which are first centered and scaled.

Since the standard deviation of first 6 principal components are greater than one (table-4.3), they are explaining the variation greater than the original variables. In addition around 99 percent of variation in x-variables are explained by

13 components of PCA which is seen on the cumulative proportaion of variation in the same table.

Table 4.3: Dispersion of data explained by principal components

Comp	Std.Dev	Var.Prop	Cum.Var.Prop	Comp	Std.Dev	Var.Prop	Cum.Var.Prop
1	3.018	0.414	0.414	8	0.958	0.042	0.867
2	1.602	0.117	0.531	9	0.891	0.036	0.903
3	1.376	0.086	0.617	10	0.848	0.033	0.936
4	1.216	0.067	0.684	11	0.787	0.028	0.964
5	1.054	0.051	0.734	12	0.620	0.017	0.981
6	1.023	0.048	0.782	13	0.446	0.009	0.990
7	0.978	0.044	0.825	14	0.274	0.003	0.994

4.4 Principal Component Regression

A prediction model based on the few components instead of all original variables, considered in PCA, not only remove the complexity of the model but also gives mutually orthogonal and uncorrelated components (new variables) which removes the multicollinearity problem during model fitting. A PCA model is fitted with observations in the training dataset (Jan 2000 - Dec 2012), the variation explained on both $\mathbf{X}$ and $\mathbf{Y}$ are presented in table-4.5.

The results shows that the first 6 components which explain larger variance than the actual variable, as seen in PCA, explain about 78 percent of variation in response. If 13 components are considered, the percentage of explained variation in response rises to almost 99 percent.

Table 4.5: Percentage of variation explained by PCR model in response and predictor

Comp	X	PerEURO	Comp	X	PerEURO	Comp	X	PerEURO

1	41.05	0.52	8	86.58	85.56	15	99.61	89.72
2	53.14	43.73	9	90.19	85.62	16	99.86	91.80
3	61.61	77.57	10	93.42	85.63	17	99.98	91.80
4	68.22	79.93	11	96.27	85.82	18	99.99	91.80
5	73.31	84.14	12	98.02	86.67	19	100.00	91.82
6	78.05	84.14	13	99.02	86.82	20	100.00	91.86
7	82.50	84.31	14	99.35	86.87	21	100.00	91.88

4.5 Partial Least Square Regression

Principal Component Regression aims to collect the variation present in predictor variables with its first few components but it does not give any consideration to the variation present in response. In many cases, PCA can caputre the variation present in response variable but in other situations, it fails or become slower (need more components) to explain it. In such case, Partial Least Square (PLS) regression can be a solution.

Partial Least Square (PLS) regression when fitted with seven components can explain more than 90 percent of variation in Exchange Rate while it explain only 76 percent of variation in predictor variable. Table-4.6 shows that the percentage of variation explained in Exchange rate has increased dramatically in first two components which then settled down. If all the components are considered in the model, the variation explained in the case resembles with the R^2 value of linear model. Since, the later components contains only residuals and have no useful information, the idea of including them only increases the model complexity and can leads to overfitting which is also true for PCR model.

Table 4.6: Percentage of variation Explained by PLS model in Response and Predictor

Comp	X	PerEURO	Comp	X	PerEURO	Comp	X	PerEURO
1	13.63	77.96	8	79.16	91.75	15	95.34	91.81
2	50.59	83.70	9	81.66	91.79	16	99.64	91.81
3	60.46	86.75	10	83.62	91.80	17	99.87	91.82
4	65.73	87.90	11	87.41	91.80	18	99.99	91.82
5	68.68	89.53	12	89.23	91.80	19	100.00	91.82
6	72.80	90.62	13	91.53	91.80	20	100.00	91.88
7	75.70	91.54	14	94.48	91.80	21	100.00	91.90

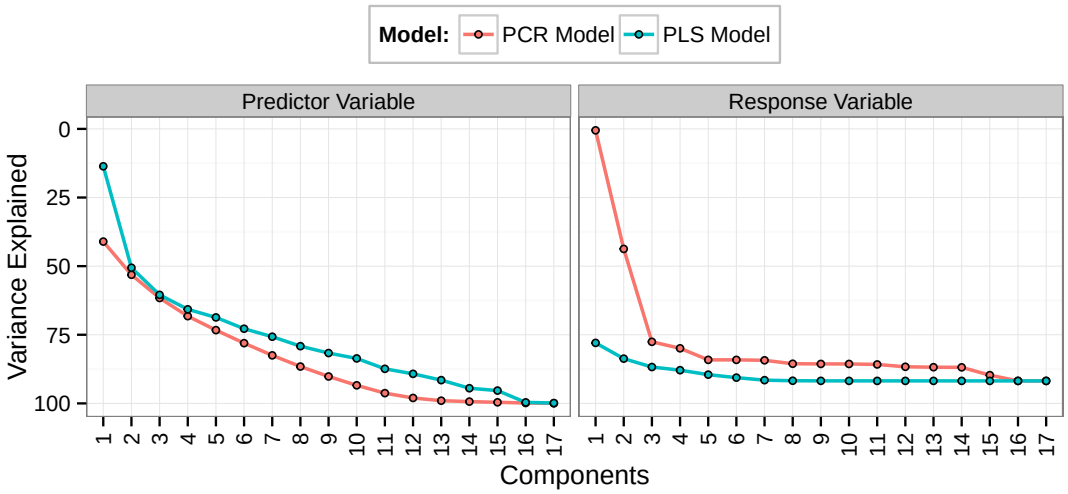


Fig 4.8: Variation Explained by PLS and PCR model on Predictor Variable and Response Variable

The actual difference between PLS and PCR model can also be observed from the variation explained plot in figure-4.8. The plot shows that PCR explain more of the predictor variation with few components while PLS explain more of the response variation with lesser components than PCR. However, on taking more components, both the models agrees at some point of variation explanation.

4.6 Ridge Regression

Ridge regression in this thesis is performed using `ridge` package. Although the package has implement semi-automatic method (Cule and De Iorio, 2012) to choose the ridge regression paramter(λ), this thesis has chosen λ from a range [0, 0.01] by implementing cross validation technique. The parameter is found to be 0.005 which can results minimum RMSECV. An alternative way is to choose λ by maximizing the R^2 predicted (fig-4.9). The parameter is also known as shrinkage parameter as it shrink the coefficients estimates which was enlarged by the Multicollinearity problem. Coefficient estimates plotted in figure -4.12 shows that the coefficients obtained from linear model has fluctuated due to the presence of multicolliniearity. In the figure, the coefficients obtained from ridge regression were pulled down towards zero.

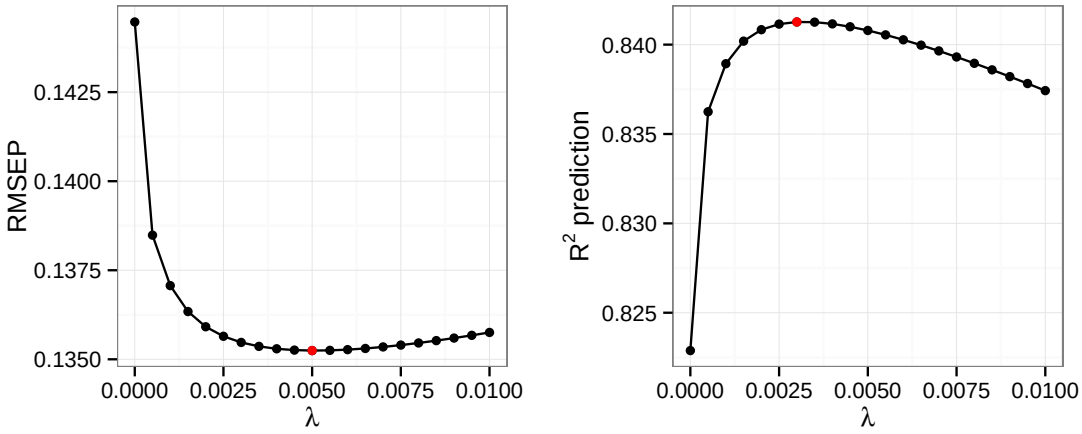


Fig 4.9: RMSE and R2pred plots for different ridge regression paramter λ. The red dots refers to the maximum R^2 pred and minimum RMSEP.

4.7 Cross Validation

Usually, a predictive model is expected to predict test responses not included in the sample. A model which can well predict the in-sampled observation may not perform well for out-of-sample observations. Cross-validation can verify the ablity of model during prediction of such observations. Since time-series has a sequential form of ordered by date, a random prediction is unsuitable. A cross-validation technique is appled to the training dataset dividing them into 12 consecutive segments. Each time a segment is removed from the fitted model which then predict the segment which was not included. The process is repeated for all the segments and RMSECV and R^2 prediction (Q^2) are computed using the equation-3.36 and equation-3.38 respectively. The validation is performed for all the models discussed above, from which RMSECV and R^2 predicted are computed as in table-4.9.

The table shows that PLS with 9 components and PCR with 16 components have least RMSECV and highest R^2 predicted. This also indicate that those models speaks better with the new observation, that are not included in the models, in compared with other linear models.

Further analysis is made on PLS and PCR models by computing the RMSECV and R^2 predicted, and plotted them against all the components. Figure-4.10 shows that the curve of RMSECV and R^2 predicted fluctuate over components in contrast to the results without cross-validation. In the case without cross-validation, RMSEP continually decreases initially and gets stable and R^2 predicted continually increases and gets stable.

In the plot, PLS model starts predicting better from very beginning while PCR meets the quality only after considering 16 components. From the results of

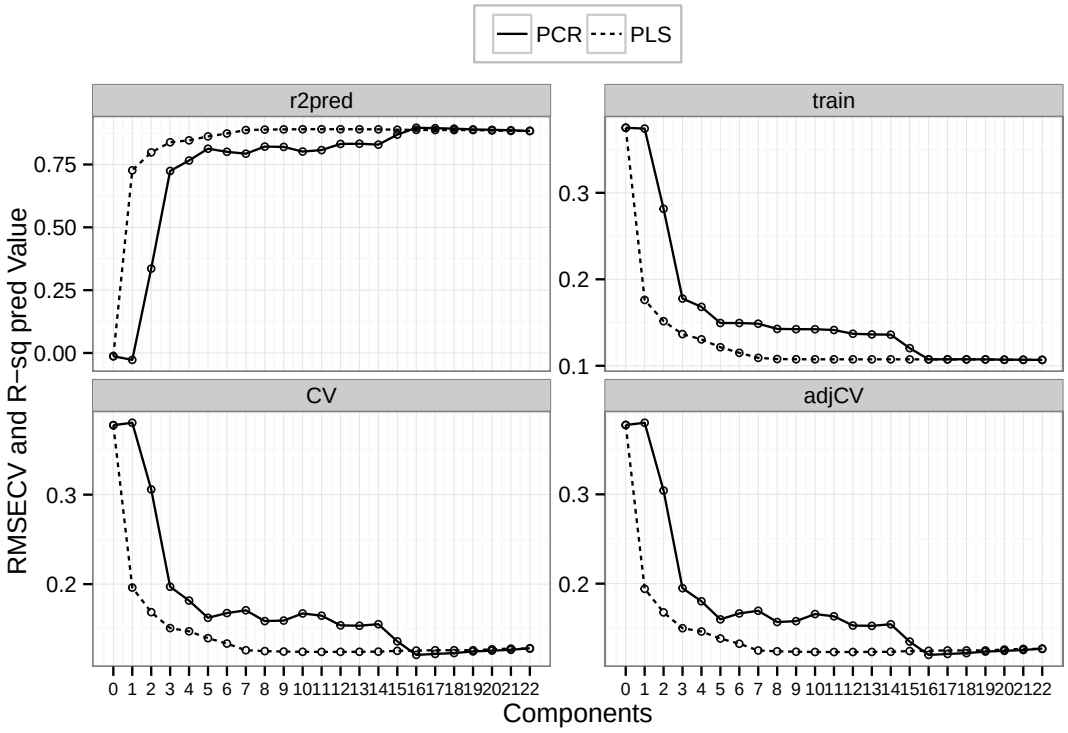


Fig 4.10: RMSEP plot for PCR and PLS model with and without cross-validation. Cross-validation is done with 12 observation in each consecutive segments within training dataset.

cross-validation, it is expected to have best prediction from the PLS model with 9 components.

4.8 Predection on test Dataset

After getting some idea about the prediction ability of a model from cross-validation procedure, it is time to observe its performance in the case of test dataset. Exchange Rate from Jan 2013 to Nov 2014 are predicted using the training dataset which includes the financial and commodity variables from Jan 2000 to Dec 2012. For the prediction, a multiple linear regression model, its

subsets selected from various selection criteria, a PLS model with 6, 7, 8, 9 components, a PCR model with 15, 16, 17 components and a ridge regression model with paramter $\lambda = 0.005$ are applied. A prediction is also made on the calibration set and the results for both predictions - Training set and Test set are plotted on figure-4.13.

The plot shows that the predictions from all the models are very close to the true value. From the RMSEP and R2pred value at the top left corner of each panel, subset model selected from minimum BIC criteria and model selected from Backward elimination procedure with F-value based criteria are predicted the test observations more closely as they have minimum RMSEP and maximum R2pred. However, in the case of in-sample prediction on the training dataset, linear model has least RMSEP, but since it is suffered from multi-collinearity problem, PLS model with 8 components can be an alternative.

4.9 Comparison of Models

Models can be compared on the basis of their predictability and goodness of fit. As discussed in chapter-3, the goodness of fit of a model can be accessed from (a) variation the model has described, (b) distribution of residuals and (c) information criteria and the predictability of the model can be compared from (a) RMSEP and (b) R^2 predicted for calibration set and testset.

4.9.1 Goodness of fit

All the linear models (full and subset) have explained almost 90 percent of variation in response which is seen in R^2 and R^2 adjusted presented in table-4.7. Further,

the models are significant as their p-value is very close to zero. However, subset model chosen from maximum R^2 adjusted and minimum AIC have minimum AIC and BIC value. Both the models have selected same set of variables so both of them are equivalent. In addition, these models have the least model standard deviation (sigma). Hence, these models can be selected as best model among all the linear models.

Table 4.7: Summary statistic and information criteria for model comparison

Model	AIC	BIC	R.Sq	R.Sq.Adj	Sigma	F.value	P.value
linear	-207.1781	-133.9816	0.9190	0.9056	0.1157	68.5936	0.0000
cp.model	-230.3234	-205.9245	0.9143	0.9108	0.1124	264.8486	0.0000
r2.model	-227.9952	-191.3969	0.9173	0.9116	0.1119	160.9059	0.0000
aicMdl	-227.9952	-191.3969	0.9173	0.9116	0.1119	160.9059	0.0000
bicMdl	-229.2344	-213.9852	0.9103	0.9085	0.1139	514.1058	0.0000
forward	-229.2344	-213.9852	0.9103	0.9085	0.1139	514.1058	0.0000
backward	-230.3234	-205.9245	0.9143	0.9108	0.1124	264.8486	0.0000

Further, the residues obtained from these selected regression models are nearly Normal and random which can be seen from the diagnostic plots in **??** but still there are some outliers due to the globel financial crisis discussed in chapter-**??**. Despite having outliers in these models, the outliers are not very influencial as their cook's distace is still less than a unit.

4.9.2 Predictability

The main concert of this thesis is about the predictability of a model. The predictability of a model is measured using RMSEP and R^2 predicted. A model exhibit different nature in the case of prediction in training dataset, during cross-

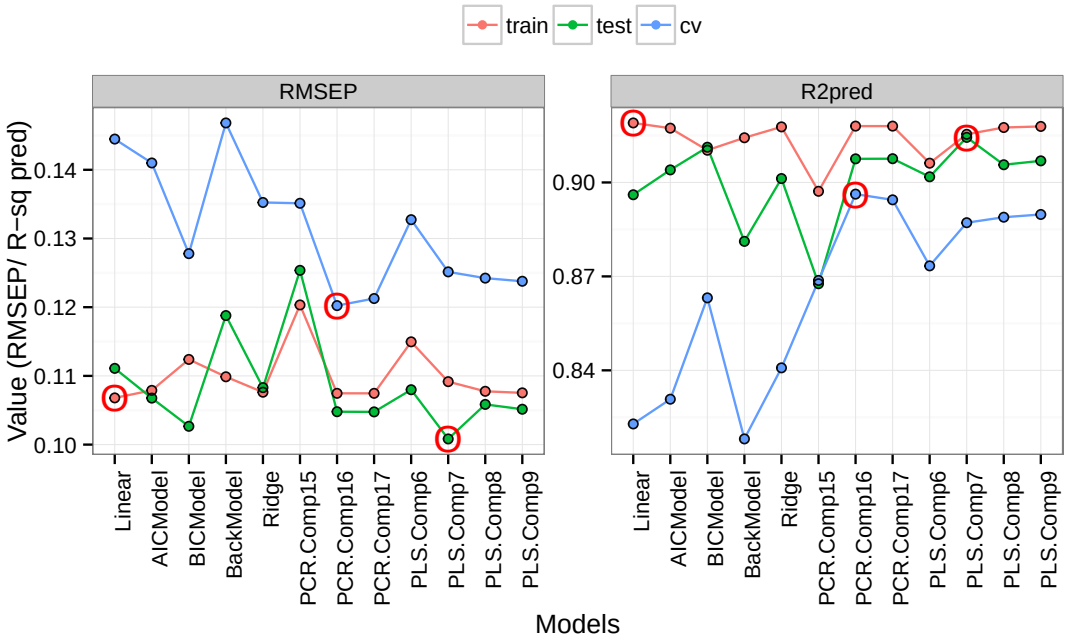


Fig 4.11: Comparision of Model on the ground of calibration model, cross-validation models and prediction model on the basis of RMSEP and R^2 predicted

validation and when implementing it to predict the test dataset. The plot in fig-4.11 shows this discripancies.

For all the candidate models considered as best, RMSEP and R^2 predicted are tabulated for training dataset, during cross-validation and for test dataset. It is observed that Linear Model has generated least prediction error when predicting the samples on training dataset. During cross-validation, PLS model with 9 components perform best by giving least RMSEP (0.1237746). The main concert of this this thesis is the prediction of test dataset. The Subset of linear model selected from minimum BIC criteria has least prediction error (RMSEP=0.1026631) and maximum R^2 predicted (0.8631472).

Table 4.9: Validation result containing RMSEP and R2pred for training set, cross-validation set and test set

Model	Training		Cross Validation		Test	
	RMSEP	R2pred	RMSEP	R2pred	RMSEP	R2pred
Linear	0.1068	0.9190	0.1445	0.8229	0.1111	0.8961
AICModel	0.1079	0.9173	0.1410	0.8308	0.1068	0.9040
BICModel	0.1124	0.9103	0.1278	0.8631	0.1027	0.9112
BackModel	0.1099	0.9143	0.1468	0.8181	0.1188	0.8812
Ridge	0.1076	0.9177	0.1352	0.8408	0.1083	0.9012
PCR.Comp15	0.1203	0.8972	0.1351	0.8687	0.1254	0.8677
PCR.Comp16	0.1075	0.9180	0.1202	0.8963	0.1048	0.9075
PCR.Comp17	0.1075	0.9180	0.1213	0.8945	0.1048	0.9076
PLS.Comp6	0.1150	0.9062	0.1327	0.8734	0.1080	0.9018
PLS.Comp7	0.1092	0.9154	0.1251	0.8871	0.1008	0.9144
PLS.Comp8	0.1078	0.9175	0.1242	0.8889	0.1058	0.9057
PLS.Comp9	0.1075	0.9179	0.1238	0.8898	0.1051	0.9069

4.10 Coefficients Estimates

The estimated coefficients of a linear model are larger in magnitude than the Ridge, PCR and PLS models. The first lagged response has very high (1.0907) positive coefficient and has large influence on the model. The plot in figure-4.12 shows that Import of old ship has larger coefficients than other import and export variables. On Dec 2008, a large sum of money is used to import elderly ships in Norway (2.10) which has an impact on its effect on the exchange rate models.

In addition, the PLS (9 Comp) and PCR (16 Comp) model have identified Oil spot price, Key interest rate, CPI and its lagged value as influential variable apart from the two lagged response variables. Some of the variables having higher coefficients obtained from these two models are presented in table -4.10.

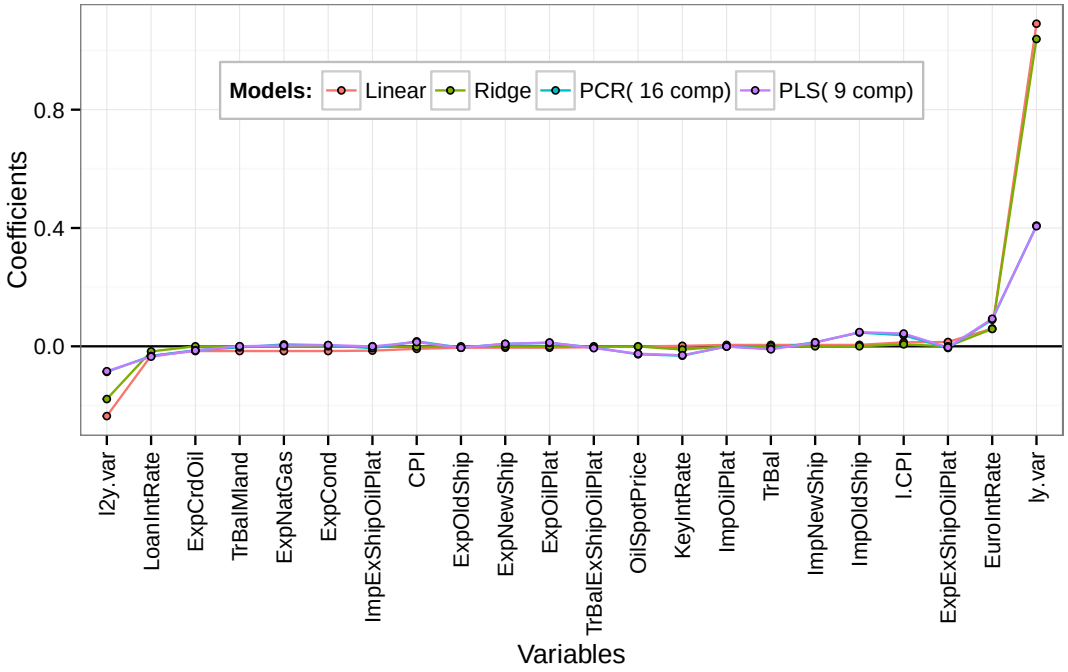


Fig 4.12: Comparision plot for coefficients estimates of predictor variables. The variables are sorted according to their estimates from linear model.

Table 4.10: Top four (both positive and negative) Coefficient Estimate of PLS and PCR model

vars	pcr	pls
l2y.var	-0.0863	-0.0843
LoanIntRate	-0.0328	-0.0350
KeyIntRate	-0.0317	-0.0302
OilSpotPrice	-0.0270	-0.0251
ly.var	0.4081	0.4060
EuroIntRate	0.0902	0.0932
ImpOldShip	0.0464	0.0479
l.CPI	0.0374	0.0431

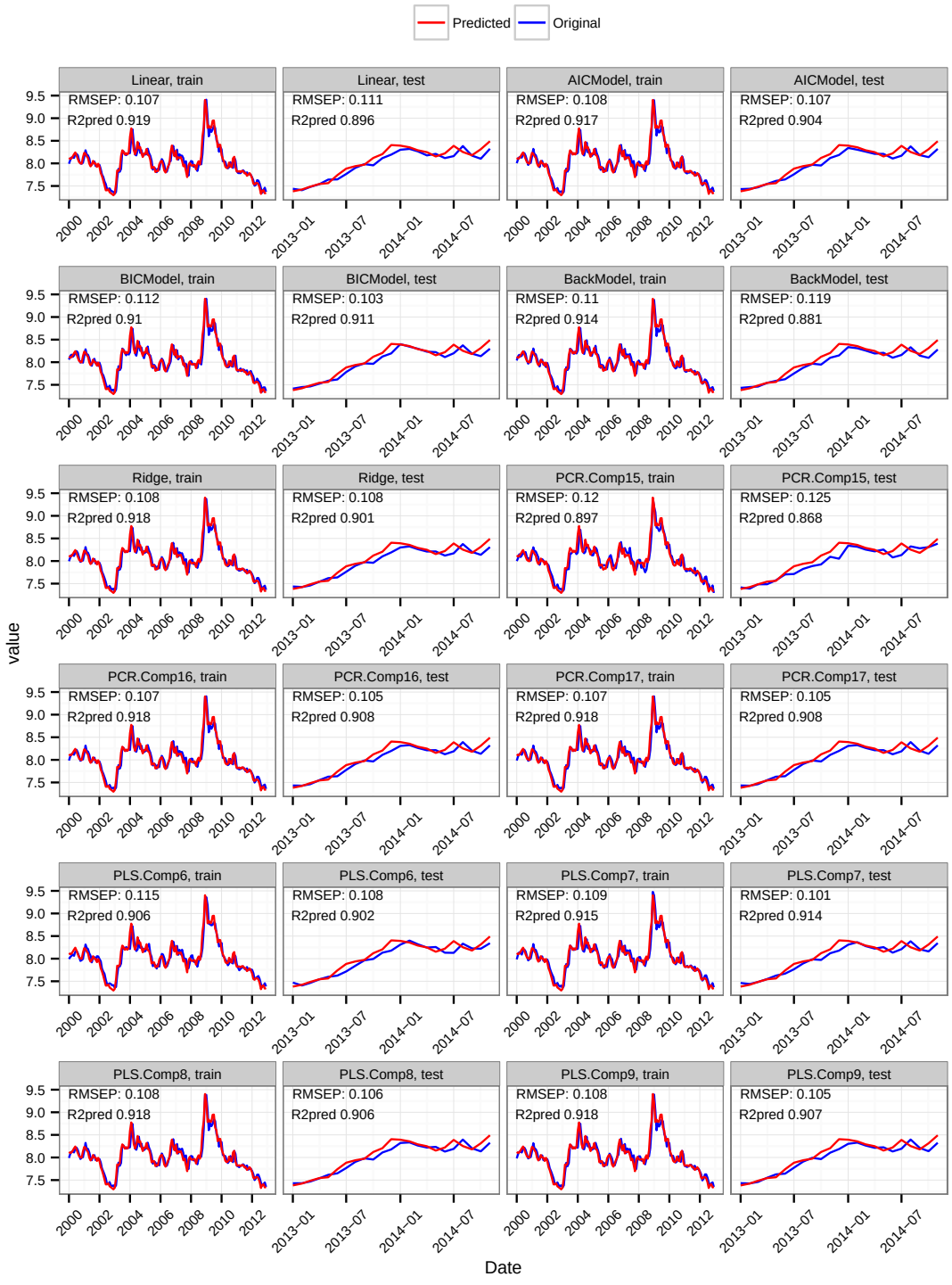


Fig 4.13: Prediction made on trained and test dataset using different models

Chapter 5

Discussions and Conclusion

It is always a priliminary idea to use basic liner model. A full linear model does not results always on selecting important and significant variables. An attempts of choosing a subset of linear model with variable selection technique results on selecting a model with minimum Mallows Cp.

The plot in appendix - C.1 contains four plot. The first one is the fitted value vs square root of standardize residuals. In the plot the crisis period have higher fitted values and have greater residues. the second plot elaborate the problem a step forward. The plot clearly shows that the distortion on the normality are due the observation of the crisis period. The third plot of cook's distance shows the most of the outlier observation are from the crisis period which have larger influence. Their influence is shown in the fourth plot of Leverage vs standardized residuals.

Although have some influntial outliers, the observations are still within the limit. The value of most influencing outlier is from Dec 2008 which is a crutial

time point of the recent great recession (*The Financial Market in Norway 2008: Risk outlook* 2009).

The loading plot (appendix-C.2) for PLS model shows that component one constitute the effect of lagged value of response which generate high positive values in loadings of first components. Some of the export related variables which has positive contribution on second components has negative contribution on first components. The second components has high negative influence of interest rate variable while this component has positive contribution of the oil spot price. Since there is more than 77 percent of contribution of first component, it shows that the lagged value of response has huge contribution on explaining the variation present on Exchange rate. In addition, the effect of interest rate , Oil price and export related variables are gathered by the second components.

Additionally, score plots (appendix-C.3) for the first three components of partial least square regression revels the fact that the second componets which contains 36.96 percent of $\mathbf{X}$ variation has accumulate the effect of crisis period. Most of the positive large scores of second components are from the crisis period.

Although the models have made close predictions, on the basis of comparision criteria set before, `cp.model`(Model selected with minimum Mallow's Cp criteria) and `backward`(Model selected with criteria of F-test using backward elimination procedure) models can be selected as better than other linear models. These models have minimum AIC

Bibliography

[AFC14] D.R. Appleyard, A.J.J. Field, and S.L. Cobb. *International Economics*. The McGraw-Hill series economics. McGraw-Hill Education, 2014. ISBN: 9781259010576. URL: https://books.google.no/books?id=kUFTMAEACAAJ.

[Aka74] Hirotugu Akaike. "A new look at the statistical model identification". In: *Automatic Control, IEEE Transactions on* 19.6 (1974), pp. 716–723.

[AM09] Thomas Lumley using Fortran code by Alan Miller. *leaps: regression subset selection*. R package version 2.9. 2009. URL: http://CRAN.R-project.org/package=leaps.

[Aug12] Baptiste Auguie. *gridExtra: functions in Grid graphics*. R package version 0.9.1. 2012. URL: http://CRAN.R-project.org/package=gridExtra.

[CDI12] Erika Cule and Maria De Iorio. "A semi-automatic method to guide the choice of ridge parameter in ridge regression". In: *arXiv preprint arXiv:1205.0686* (2012).

[Cul14] Erika Cule. *ridge: Ridge Regression with automatic selection of the penalty parameter*. R package version 2.1-3. 2014. URL: http://CRAN.R-project.org/package=ridge.

[D'A86] R.B. D'Agostino. *Goodness-of-Fit-Techniques*. Statistics: A Series of Textbooks and Monographs. Taylor & Francis, 1986. ISBN: 9780824774875. URL: https://www.google.no/books?id=1BSEaGVBj5QC.

[Dah14] David B. Dahl. *xtable: Export tables to LaTeX or HTML*. R package version 1.7-4. 2014. URL: http://CRAN.R-project.org/package=xtable.

[Eco] *Norway The rich cousin*. Feb. 2013. URL: http://www.economist.com/news/special-report/21570842-oil-makes-norway-different-rest-region-only-up-point-rich.

[Eur] *The euro*. 2015. URL: http://ec.europa.eu/economy_finance/euro/index_en.htm.

[Fin] *The Financial Market in Norway 2008: Risk outlook*. Report. Kredittilsynet, 2009.

[FK91] Hsing Fang and K Kern Kwong. "Forecasting Foreign Exchange Rates". In: *The Journal of Business* 92 (1991).

[FRR12] Domenico Ferraro, Kenneth S Rogoff, and Barbara Rossi. *Can oil prices forecast exchange rates?* Tech. rep. National Bureau of Economic Research, 2012.

[FW11] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Second. Thousand Oaks CA: Sage, 2011. URL: http://socserv.socsci.mcmaster.ca/jfox/Books/Companion.

[FW74] George M Furnival and Robert W Wilson. "Regressions by leaps and bounds". In: *Technometrics* 16.4 (1974), pp. 499–511.

[Fxi] *Introduction to the Forex Market*. eng. URL: http://www.forex.com/uk/intro-forex-market.html.

[Gar94] Paul H Garthwaite. "An interpretation of partial least squares". In: *Journal of the American Statistical Association* 89.425 (1994), pp. 122–127.

[GK86] Paul Geladi and Bruce R Kowalski. "Partial least-squares regression: a tutorial". In: *Analytica chimica acta* 185 (1986), pp. 1–17.

[HK70] Arthur E Hoerl and Robert W Kennard. "Ridge regression: Biased estimation for nonorthogonal problems". In: *Technometrics* 12.1 (1970), pp. 55–67.

[Hu07] Shuhua Hu. "Akaike information criterion". In: *Center for Research in Scientific Computation* (2007).

[Int] *Effect of Interest Rate Changes*. Tech. rep. Norges Bank, 2004. URL: http://www.norges-bank.no/en/Monetary-policy/Effect-of-interest-rate-changes/.

[Jol82] Ian T Jolliffe. "A note on the use of principal components in regression". In: *Applied Statistics* (1982), pp. 300–303.

[JW07] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Pearson Education International. Pearson Prentice Hall, 2007. ISBN: 9780135143506. URL: http://books.google.no/books?id=SJZnPwAACAAJ.

[KO06] P.R. Krugman and M. Obstfeld. *International Economics: Theory and Policy - 9th Edition*. Business & Investing, Professional &; Technical, Science. Pearson Education, Limited, 2006. ISBN: 9780321461834. URL: https://books.google.no/books?id=ojU_BgAAQBAJ.

[LCG03] James M Lattin, J Douglas Carroll, and Paul E Green. *Analyzing multivariate data*. Thomson Brooks/Cole Pacific Grove, CA, 2003.

[LL04] Thomas Lumley and Maintainer Thomas Lumley. "The leaps package". In: *R Project for Statistical Computing, Vienna, Austria (Available from: cran. r-project. org/doc/packages/leaps. pdf)* (2004).

[LS14] Kristian Hovde Liland and Solve Sæbø. *mixlm: Mixed Model ANOVA and Statistics for Education*. R package version 1.0.7. 2014. URL: http://CRAN.R-project.org/package=mixlm.

[Mad12] Jeff Madura. *International financial management*. Cengage Learning, 2012.

[Mal73] Colin L Mallows. "Some comments on C p". In: *Technometrics* 15.4 (1973), pp. 661–675.

[Mas98] D.L. Massart. *Handbook of Chemometrics and Qualimetrics*. Data handling in science and technology pt. 1. Elsevier, 1998. ISBN: 9780444897244. URL: http://books.google.no/books?id=0u7vAAAAMAAJ.

[MN92] H. Martens and T. Naes. *Multivariate Calibration*. Wiley, 1992. ISBN: 9780471930471. URL: http://books.google.no/books?id=6lVcUeVDg9IC.

[MS75] Donald W Marquardt and Ronald D Snee. "Ridge regression in practice". In: *The American Statistician* 29.1 (1975), pp. 3–20.

[MWL13] Bjørn-Helge Mevik, Ron Wehrens, and Kristian Hovde Liland. *pls: Partial Least Squares and Principal Component regression*. R package version 2.4-3. 2013. URL: http://CRAN.R-project.org/package=pls.

[Nok] *Norwegian Kroner*. 2014/12. URL: http://www.oanda.com/currency/iso-currency-codes/NOK.

[Nora] *Brief History Of Norges Bank*. 2014-11. URL: http://www.norges-bank.no/en/about/History/Norges-Banks-history/.

[Norb] *FAQ: Monetary Policy, Inflation and Interest Rates*. 2007. URL: http://www.norges-bank.no/en/faq/monetary-policy/.

[Oâ07] RobertM. Oâbrien. “A Caution Regarding Rules of Thumb for Variance Inflation Factors”. English. In: *Quality & Quantity* 41.5 (2007), pp. 673–690. ISSN: 0033-5177. DOI: 10.1007/s11135-006-9018-6. URL: http://dx.doi.org/10.1007/s11135-006-9018-6.

[Seb08] George AF Seber. *A matrix handbook for statisticians*. Vol. 15. John Wiley & Sons, 2008.

[Sur] Steve M. Suranovic. *Balance of Payments Deficits and Surpluses*. URL: http://internationalecon.com/Finance/Fch80/F80-8.php.

[Tay90] Richard Taylor. “Interpretation of the correlation coefficient: a basic review”. In: *Journal of diagnostic medical sonography* 6.1 (1990), pp. 35–39.

[VR02] W. N. Venables and B. D. Ripley. *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer, 2002. URL: http://www.stats.ox.ac.uk/pub/MASS4.

[War+14] Gregory R. Warnes et al. *gdata: Various R programming tools for data manipulation*. R package version 2.13.3. 2014. URL: http://CRAN.R-project.org/package=gdata.

[Wei05] Sanford Weisberg. *Applied linear regression*. Vol. 528. John Wiley & Sons, 2005.

[WF14] Hadley Wickham and Romain Francois. *dplyr: A Grammar of Data Manipulation*. R package version 0.3.0.2. 2014. URL: http://CRAN.R-project.org/package=dplyr.

[Wic07] Hadley Wickham. “Reshaping Data with the reshape Package”. In: *Journal of Statistical Software* 21.12 (2007), pp. 1–20. URL: http://www.jstatsoft.org/v21/i12/.

[Wic09] Hadley Wickham. *ggplot2: elegant graphics for data analysis*. Springer New York, 2009. ISBN: 978-0-387-98140-6. URL: http://had.co.nz/ggplot2/book.

[Wic11] Hadley Wickham. “The Split-Apply-Combine Strategy for Data Analysis”. In: *Journal of Statistical Software* 40.1 (2011), pp. 1–29. URL: http://www.jstatsoft.org/v40/i01/.

[Wic14] Hadley Wickham. *scales: Scale functions for graphics*. R package version 0.2.4. 2014. URL: http://CRAN.R-project.org/package=scales.

[Woo12] Jeffrey Wooldridge. *Introductory econometrics: A modern approach*. Cengage Learning, 2012.

[WSE01] Svante Wold, Michael Sjöström, and Lennart Eriksson. "PLS-regression: a basic tool of chemometrics". In: *Chemometrics and intelligent laboratory systems* 58.2 (2001), pp. 109–130.

[Xie13] Yihui Xie. *Dynamic Documents with R and knitr*. ISBN 978-1482203530. Boca Raton, Florida: Chapman and Hall/CRC, 2013. URL: `http://yihui.name/knitr/`.

[YG02] Ozgür Yeniay and Atill Goktas. "A comparison of partial least squares regression with other prediction methods". In: *Hacettepe Journal of Mathematics and Statistics* 31.99 (2002), p. 111.

[ZG05] Achim Zeileis and Gabor Grothendieck. "zoo: S3 Infrastructure for Regular and Irregular Time Series". In: *Journal of Statistical Software* 14.6 (2005), pp. 1–27. URL: `http://www.jstatsoft.org/v14/i06/`.

[R C14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2014. URL: `http://www.R-project.org/`.

Appendix A

Data Description

The variables used in this paper are listed in following table along with the code used for them.

Code	Description
Date	Date
PerEURO	Exchange Rate of NOK per Euro
PerUSD	Exchange Rate of NOK per USD
KeyIntRate	Key policy rate (Percent)
LoanIntRate	Overnight Lending Rate (Nominal)
EuroIntRate	Money market interest rates of Euro area (EA11-2000, EA12-2006, EA13-2007, EA15-2008, EA16-2010, EA17-2013, EA18)
CPI	Consumer Price Index (1998=100)
OilSpotPrice	Europe Brent Spot Price FOB (NOK per Barrel)
ImpOldShip	Imports of elderly ships (NOK million)
ImpNewShip	Imports of new ships (NOK million)
ImpOilPlat	Imports of oil platforms (NOK million)
ImpExShipOilPlat	Imports excl. ships and oil platforms (NOK million)
ExpCrdOil	Exports of crude oil (NOK million)
ExpNatGas	Exports of natural gas (NOK million)
ExpCond	Exports of condensates (NOK million)
ExpOldShip	Exports of elderly ships (NOK million)
ExpNewShip	Exports of new ships (NOK million)
ExpOilPlat	Exports of oil platforms (NOK million)
ExpExShipOilPlat	Exports excl. ships and oil platforms (NOK million)
TrBal	Trade balance (Total exports - total imports) (NOK million)

Code	Description
TrBalExShipOilPlat	Trade balance (Exports - imports, both excl. ships and oil platforms) (NOK million)
TrBalMland	Trade balance (Mainland exports - imports excl. ships and oil platforms) (NOK million)
ly.var	First Lag Exchange Rate of NOK per Euro
l2y.var	Second Lag Exchange Rate of NOK per Euro
l.CPI	First Lag of Consumer Price Index
ExcChange	Change status of Exchange Rate (Increase, Decrease and Unchange)
Testrain	Test and Train seperation of data
season	Seasons

Appendix B

R packages used

Name	Version	Title
`MASS`(Venables and Ripley, 2002)	7.3-35	Support Functions and Datasets for Venables and Ripley's MASS
`car`(Fox and Weisberg, 2011)	2.0-22	Companion to Applied Regression
`pls`(Mevik, Wehrens, and Liland, 2013)	2.4-3	Partial Least Squares and Principal Component regression
`xtable`(Dahl, 2014)	1.7-4	Export tables to LaTeX or HTML
`grid`(Auguie, 2012)	3.1.2	The Grid Graphics Package
`gridExtra`(Auguie, 2012)	0.9.1	functions in Grid graphics
`knitr`(Xie, 2013)	1.8	A General-Purpose Package for Dynamic Report Generation in R
`leaps`(Alan Miller, 2009)	2.9	regression subset selection
`zoo`(Zeileis and Grothendieck, 2005)	1.7-11	S3 Infrastructure for Regular and Irregular Time Series (Z's ordered observations)
`gdata`(Warnes et al., 2014)	2.13.3	Various R programming tools for data manipulation
`ridge`(Cule, 2014)	2.1-3	Ridge Regression with automatic selection of the penalty parameter

Name	Version	Title
`plyr`(Wickham, 2011)	1.8.1	Tools for splitting, applying and combining data
`dplyr`(Wickham and Francois, 2014)	0.3.0.2	A Grammar of Data Manipulation
`ggplot2`(Wickham, 2009)	1.0.0	An implementation of the Grammar of Graphics
`reshape2`(Wickham, 2007)	1.4	Flexibly reshape data: a reboot of the reshape package.
`scales`(Wickham, 2014)	0.2.4	Scale functions for graphics.
`mixlm`(Liland and Sæbø, 2014)	1.0.7	Mixed Model ANOVA and Statistics for Education
`graphics`(R Core Team, 2014)	3.1.2	The R Graphics Package
`grDevices`(R Core Team, 2014)	3.1.2	The R Graphics Devices and Support for Colours and Fonts
`utils`(R Core Team, 2014)	3.1.2	The R Utils Package
`datasets`(R Core Team, 2014)	3.1.2	The R Datasets Package
`methods`(R Core Team, 2014)	3.1.2	Formal Methods and Classes
`base`(R Core Team, 2014)	3.1.2	The R Base Package

Appendix C

Some Relevent Plots

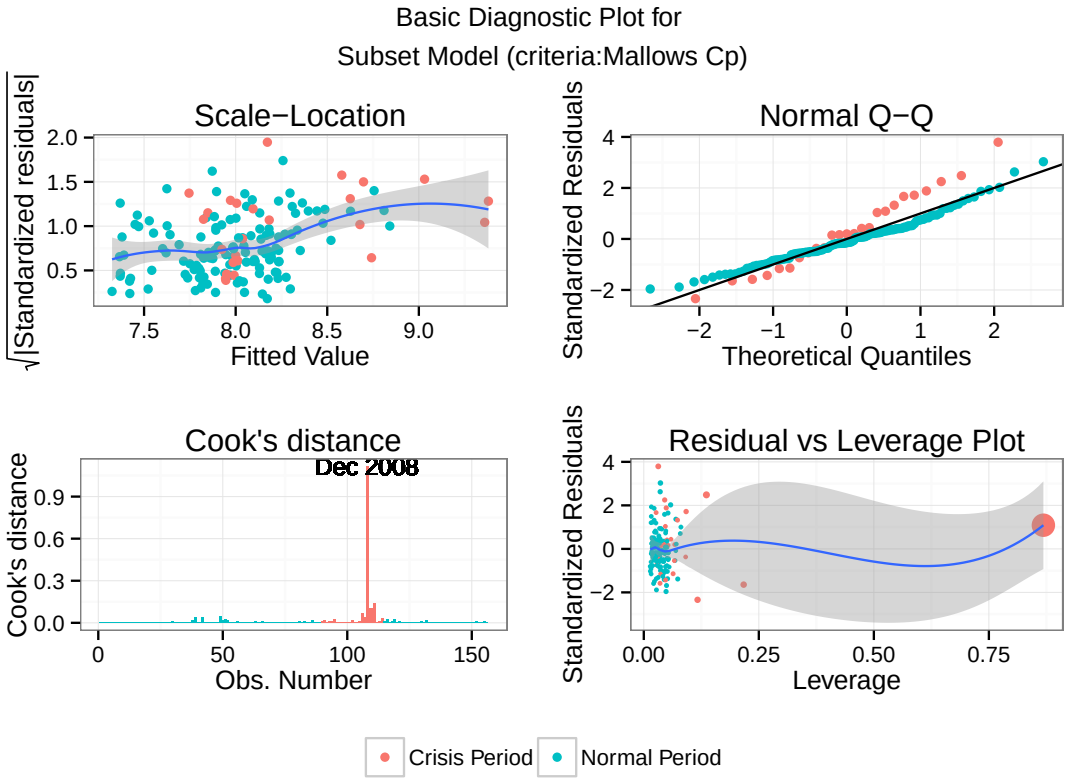


Fig C.1: Diagnostic plot for the subset of linear model selected from minimum C_p criteria. The red bubble represents the two years of crisis period from june 2007 till june 2009. The size of a bubbles in the plot of leverage vs standardized residuals on bottom right corner represents the cooks' distance.

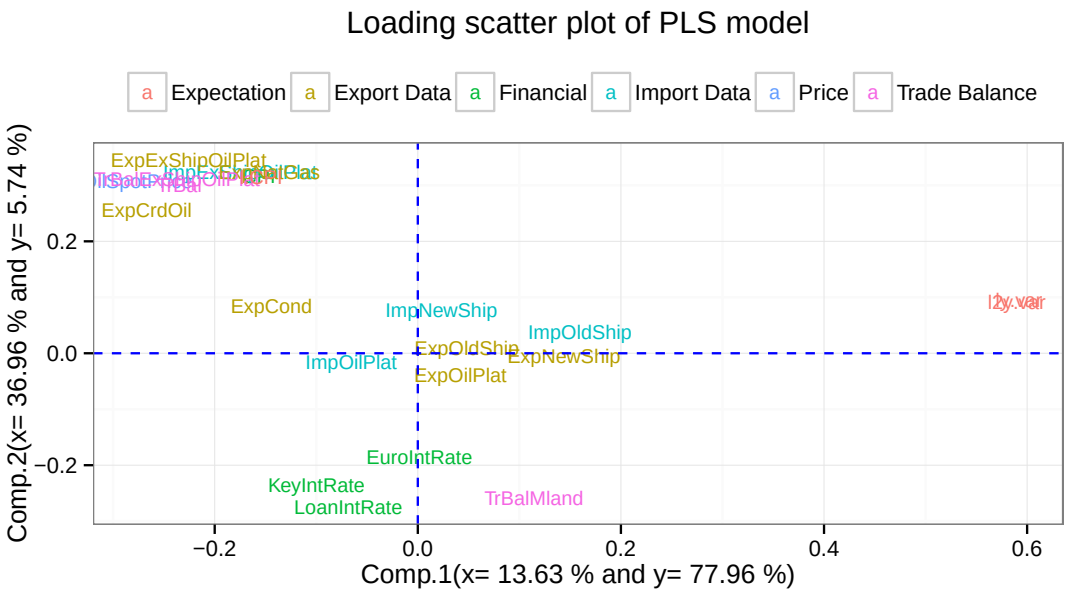


Fig C.2: Scatter loading plot of PLS with its first and second components. Labels are colored according to their domain of fields.

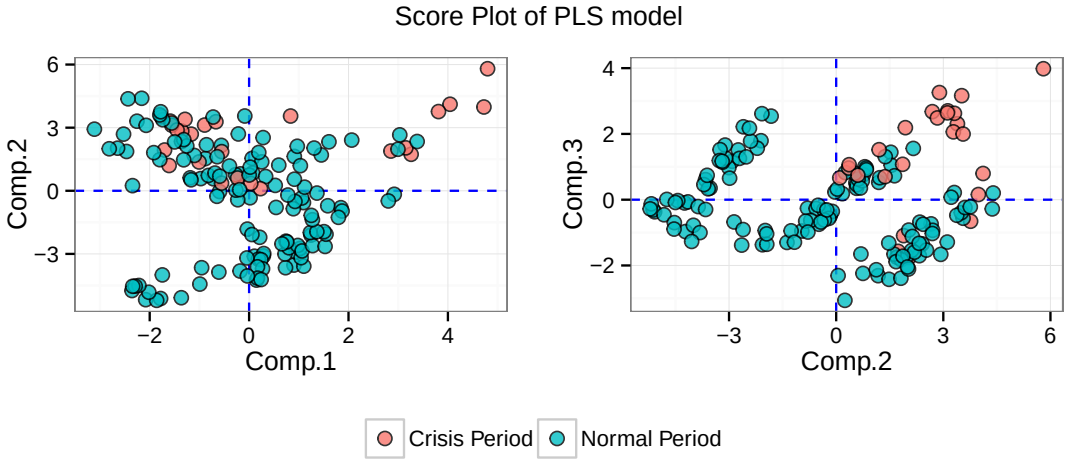
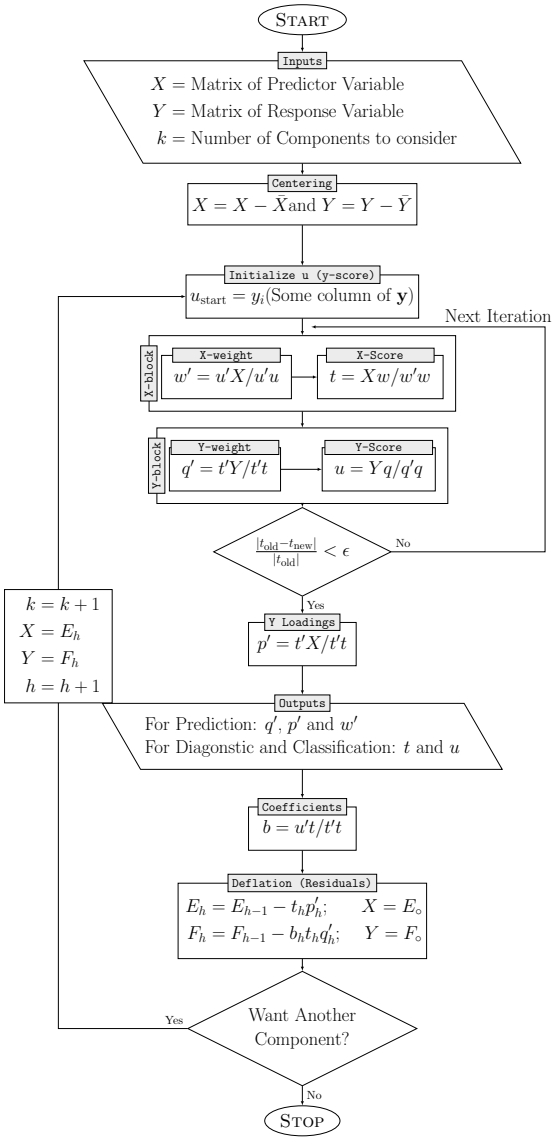


Fig C.3: Scoreplot of first three component of PLS regression. The red bubbles represents the crisis period.

Appendix D

Flowchart of NIPLS Algorithm



Source: Wold, Sjöström, and Eriksson, 2001; Geladi and Kowalski, 1986

Fig D.1: Flowchart of NIPLS algorithm to perform Partial Least Square Regression

Appendix E

Codes in Use

```
1
## ----frontMatter, child="frontMatter.Rnw"--------------------------------
3

## ----LoadingPkgs, echo=FALSE, message=FALSE, warning=FALSE, results='hide
    '----
req.package<-c("MASS", "car", "pls", "xtable", "grid", "gridExtra", "knitr", "
    leaps", "zoo", "gdata","ridge", "plyr", "dplyr", "ggplot2", "reshape2", "
    scales","mixlm")
lapply(req.package, require, character.only=TRUE, quietly = T, warn.conflicts =
     F)

9



11
## ----setup, include=FALSE, cache=FALSE, echo=TRUE------------------------
opts_chunk$set(fig.path='Include', fig.align='center')
render_listings()
setwd('~/Dropbox/UMB/Thesis/MSThesis/')
Sys.setenv(TEXINPUTS=getwd(),
           BIBINPUTS=getwd(),
           BSTINPUTS=getwd())
#data.path<-path.expand(file.path(dirname(dirname(getwd())), "Datasets", "
    CompleteDataSet.xlsx"))
data.path<-path.expand(file.path(dirname(getwd()), "Datasets", "CompleteDataSet
    .xlsx"))
21

23 ## ----readFun, child="Include/functions.Rnw"------------------------------

25
## ----functions, echo=FALSE, cache=FALSE, warning=FALSE-------------------
27
```

```
## Setting up Crisis Period
cp.cat<-function(dateVec){
cp.col<-ifelse(dateVec<cperiod[1] | dateVec>cperiod[2],
               "Normal Period",
               "Crisis Period")
return(cp.col)
}

## Timeseries plot
plotTS<-function(dataSet, dateVarColIdx, nc){
  plt<-ggplot(melt(dataSet, dateVarColIdx), aes(Date, (value/100)))
  plt<-plt+geom_line()
  plt<-plt+facet_wrap(~variable,
                      ncol=nc,
                      scale="free_y")
  plt<-plt+theme_bw()
  plt<-plt+theme(text=element_text(size=12))
  plt<-plt+labs(x="Date (Monthly)", y="Value (NOK hundreds)")
  return(plt)
}

## Plotting Model Coefficients with their state of significance
test.plot<-function(model, alpha=0.05){
  .e<-environment()
  coef.matrix<-data.frame(summary(model)$coef)
  names(coef.matrix)<-c("Estimate", "StdError", "t.value", "p.value")
  idx<-order(row.names(coef.matrix))
  cp<-ggplot(coef.matrix[idx,], aes(x=row.names(coef.matrix[idx,]), y=t.value),
     environment = .e)
  cp<-cp+geom_bar(stat="identity", position = "identity",
                  fill=ifelse(coef.matrix[idx,"p.value"]<alpha, "coral3", "
    cornflowerblue"))
  cp<-cp+geom_text(aes(y=ifelse(coef.matrix[idx, "t.value"]>0,t.value+0.7, t.
    value-0.7),
                       label=round(coef.matrix[idx,"Estimate"], 2)), angle=45,
    size=5)
  cp<-cp+theme_bw()+labs(x="", y="T-Value")
  cp<-cp+theme(axis.text.x=element_text(angle=90, hjust=1))
  cp<-cp+theme(text=element_text(size=20))
  cp<-cp+scale_fill_manual("Status", values=c("firebrick2", "dodgerblue3"),
                           labels=c("Significant", "Non-Significant"))
  cp<-cp+geom_hline(yintercept=c(-1,1)*qt(alpha/2, df = abs(diff(dim(model$
    model[,-1])))), lower.tail = F),
                    color="red", linetype="dashed")
  cp<-cp+theme(legend.title=element_blank(),
               legend.position=c(0.8, 0.2))
  cp<-cp+geom_hline(yintercept=0, color="black", size=.2)
  return(cp)
}
```

```
## Fitting Linear Model
fit.model<-function(Model, yVar, xVars, dataSet, scaling=TRUE){
  model<-match.fun(Model)
  formula<-as.formula(paste(yVar, paste(xVars, collapse="+"), sep="~"))
  if(scaling){
      model<-model(formula, data=dataSet, scale=TRUE)
  }else{
      model<-model(formula, data=dataSet)
  }
  return(list(formula=formula, model=model, dataset=dataSet))
}


## Diagnostic Plot using GGPlot
diagPlot<-function(model, cp.color){
  p1<-ggplot(model, aes(.fitted, .resid))+geom_point(aes_string(color=cp.color)
    )
  p1<-p1+stat_smooth(method="loess")
  p1<-p1+geom_hline(yintercept=0, col="red", linetype="dashed")
  p1<-p1+xlab("Fitted values")+ylab("Residuals")
  p1<-p1+ggtitle("Residual vs Fitted Plot")+theme_bw()

  ## qline slope and intercept
  qline<-ldply(data.frame(res=stdres(mdl.ft$linear$model)), function(x){
      slope = (quantile(x,p=.75)-quantile(x,.25))/(qnorm(.75)-qnorm(.25))
      intercept = quantile(x,.25) - slope*qnorm(.25)
      data.frame(slope, intercept)})

  p2<-ggplot(model, aes(sample=.stdresid))+stat_qq(aes_string(color=cp.color))
  p2<-p2+geom_abline(data = qline, aes(slope, intercept))+xlab("Theoretical
    Quantiles")+ylab("Standardized Residuals")
  p2<-p2+ggtitle("Normal Q-Q")+theme_bw()

  p3<-ggplot(model, aes(.fitted, sqrt(abs(.stdresid))))+geom_point(na.rm=TRUE,
    aes_string(color=cp.color))
  p3<-p3+stat_smooth(method="loess", na.rm = TRUE)+xlab("Fitted Value")
  p3<-p3+ylab(expression(sqrt("|Standardized residuals|")))
  p3<-p3+ggtitle("Scale-Location")+theme_bw()

  p4<-ggplot(model, aes(seq_along(.cooksd), .cooksd))+geom_bar(stat="identity",
     position="identity", aes_string(fill=cp.color))
  p4<-p4+xlab("Obs. Number")+ylab("Cook's distance")
  p4<-p4+geom_text(aes(x=which.max(.cooksd),
                       y = max(.cooksd),
                       label=format(baseTable[which.max(.cooksd), "Date"], "%b %Y")
    ),
                       size=4)
  p4<-p4+ggtitle("Cook's distance")+theme_bw()
```

```
  p5<-ggplot(model, aes(.hat, .stdresid))
  p5<-p5+geom_point(aes_string(color=cp.color, size=".cooksd"), na.rm=TRUE)
  p5<-p5+stat_smooth(method="loess", na.rm=TRUE)
  p5<-p5+xlab("Leverage")+ylab("Standardized Residuals")
  p5<-p5+ggtitle("Residual vs Leverage Plot")
  p5<-p5+scale_size_continuous("Cook's Distance", range=c(1,5))
  p5<-p5+theme_bw()+theme(legend.position="bottom")

  p6<-ggplot(model, aes(.hat, .cooksd))+geom_point(na.rm=TRUE, aes_string(color
    =cp.color))+stat_smooth(method="loess", na.rm=TRUE)
  p6<-p6+xlab("Leverage hii")+ylab("Cook's Distance")
  p6<-p6+ggtitle("Cook's dist vs Leverage hii/(1-hii)")
  p6<-p6+geom_abline(slope=seq(0,3,0.5), color="gray", linetype="dashed")
  p6<-p6+theme_bw()

  return(list(rvfPlot=p1, qqPlot=p2, sclLocPlot=p3, cdPlot=p4, rvlevPlot=p5,
    cvlPlot=p6))
}

## Generate summary plot from a fitted model to annotate other plot
sumryBlock<-function(model){
  return(paste("R-Sq = ",signif(summary(model)$r.squared, 3),
               "\nAdj R-Sq =",signif(summary(model)$adj.r.squared, 3),
               "\nSigma =",signif(summary(model)$sigma, 3),
               "\nF =",signif(as.vector(summary(model)$fstatistic[1]), 4),
               paste("(",paste(as.vector(summary(mdl.ft$cp.model$model)$f[2:3])
    , collapse=','),")", sep="")
         ))
}

model.sumry<-function(model, call=TRUE, coefMat=TRUE, sumry=TRUE){
    if(!"lm"%in%class(model)){
        stop("Model should be of class 'lm'.\n")
    }
    else{
        s<-summary(model)$sigma
        df<-summary(model)$df
        r.sq<-summary(model)$r.squared
        adj.r.sq<-summary(model)$adj.r.squared
        f<-summary(model)$fstatistic[1]
        f.df.num<-summary(model)$fstatistic[2]
        f.df.den<-summary(model)$fstatistic[3]
        if(call){
            print(summary(model)$call)
            cat("\n")
        }
        if(coefMat){
            printCoefmat(summary(model)$coef, digits = 3)
```

```
        }
        if(sumry){
            data.frame(Sigma=summary(model)$sigma,
                       R.Sq=summary(model)$r.squared,
                       R.Sq.adj=summary(model)$adj.r.squared,
                       F.value=summary(model)$fstatistic[1],
                       df=paste(summary(model)$fstatistic[2:3], collapse=","),
                       p.value=pf(summary(model)$fstatistic[1],
                                  summary(model)$fstatistic[2],
                                  summary(model)$fstatistic[3],
                                  lower.tail = FALSE))
        }
    }
}

vifPlot<-function(model){
    if("lm"%nin%class(model)){
        stop("Model should be of class 'lm'.")
    }else{
        coef<-names(vif(model))
        vif<-as.vector(vif(model))
        mdl.label<-ifelse(label(model)=="", deparse(substitute(model)), label(model))
        vifMat<-data.frame(coef, vif)
        p<-ggplot(vifMat, aes(coef, vif))
        p<-p+geom_bar(stat="identity", color="black", fill=NA)+theme_bw()
        p<-p+ggtitle(label = paste("Variance Inflation Function plot\nModel:", mdl.label))
        if(length(coef)>5){
            p<-p+theme(axis.text.x=element_text(hjust=1, angle=90))
        }
        return(p)
    }
}

addline_format <- function(x,...){
    gsub('\\s','\n',x)
}


## Function to perform cross-validation splitting into 12 consecutive segments on Linear model and its subsets
makeFormula<-function(x.var, y.var){
    formula<-paste(y.var, paste(x.var, collapse="+"), sep="~")
    return(formula)
}
mdl.cv<-function(dataSet, x.var, y.var, model="lm", step=FALSE, criteria=NULL, split=12, lmd=NULL){
    segment<-split(1:nrow(dataSet), ceiling(1:nrow(dataSet)/split))
```

```
207     formula=makeFormula(x.var, y.var)
        mdl<-list()
209     predVec<-rep(NA, nrow(dataSet))
        errVec<-rep(NA, nrow(dataSet))
211
        for(i in seq_along(segment)){
213         dataset<-dataSet[-segment[[i]],]
            testset<-dataSet[segment[[i]],]
215         if(step & model=="lm"){
                if(!criteria %in% c("AIC", "BIC", "Cp", "R2adj", "forward", "
    backward")){
217                 stop("Please! enter the correct criteria")
                }else{
219                 require(leaps)
                    if(criteria=="Cp"){
221                     ## Model selected by Mallows Cp Criteria
                        cp.leaps<-leaps(x=dataset[,x.var],
223                                         y=dataset[,y.var],
                                         method="Cp", nbest = 1, names = x.var)
225                     # Model fitting
                        cp.which<-names(which(cp.leaps$which[which.min(cp.leaps$Cp)
    ,]))
227                     formula<-makeFormula(cp.which, y.var)
                        mdl[[i]]<-lm(formula, data=dataset)
229                 }else if(criteria=="R2adj"){
                        ## Model selected by R2adj Criteria
231                     r2adj.leaps<-leaps(x=dataset[,x.var],
                                           y=dataset[,y.var],
233                                        method="adjr2", nbest = 1, names=x.var)
                       # Model fitting
235                     r2.which<-names(which(r2adj.leaps$which[which.max(r2adj.
    leaps$adjr2),]))
                        formula<-makeFormula(r2.which, y.var)
237                     mdl[[i]]<-lm(formula, data=dataset)
                    }else if(criteria=="AIC" | criteria=="BIC"){
239                     lmBstSetSmry <- summary(regsubsets(dataset[,x.var],
                                                           dataset[,y.var],
241                                                        nbest = 1, nvmax =
    length(x.var)))
                        nvars<-apply(lmBstSetSmry$which, 1, sum)
243                     bic.vec<-lmBstSetSmry$bic
                        aic.vec<-bic.vec-nvars*log(sum(train))+nvars
245
                        ## Fitting selected linear model
247                     aic.which<-names(which(lmBstSetSmry$which[which.min(aic.vec
    ),]))[-1]
                        bic.which<-names(which(lmBstSetSmry$which[which.min(bic.vec
    ),]))[-1]
249                     if(criteria=="AIC"){
```

```
                        formula<-makeFormula(aic.which, y.var)
251                     mdl[[i]]<-lm(formula, data=dataset)
                    }else if(criteria=="BIC"){
253                     formula<-makeFormula(bic.which, y.var)
                        mdl[[i]]<-lm(formula, data=dataset)
255                 }
                }else if(criteria=="forward"){
257                     require(mixlm)
                        fm.log<-capture.output({
259                     mdl[[i]]<- forward(do.call(lm, list(formula, dataset)),
     alpha = 0.05, full = FALSE)
                    })
261             }else if(criteria=="backward"){
                        require(mixlm)
263                     fm.log<-capture.output({
                        mdl[[i]]<- backward(do.call(lm, list(formula, dataset))
    , alpha = 0.05, full = FALSE)
265                 })
                }
267         }
        }else if(step & model!='lm'){
269         stop("Stepwise can only be performed using Linear Model, Please
    input 'lm' in the model.")
        }else if(model=='lm'){
271         mdl[[i]]<-lm(formula, dataset)
        }else if(model=='ridge'){
273         require(ridge)
            mdl[[i]]<- linearRidge(formula, dataset, lambda = lmd)
275     }else{
            stop("Model can take 'lm' or 'ridge' value.")
277     }
        predVec[segment[[i]]]<-predict(mdl[[i]], newdata=testset[,x.var])
279     errVec[segment[[i]]]<-testset[,y.var]-predVec[segment[[i]]]
    }
281 rmse.cv<-sqrt(1/nrow(dataSet)*sum(errVec^2))
    r2pred<-1-sum(errVec^2)/sum((predVec-mean(dataSet[,y.var]))^2)
283 invisible(list(Model=mdl, Predicted=predVec, Error=errVec, rmsep=rmse.cv,
    r2pred=r2pred))
}
285

## Grid Arrange with common Legend
287 grid_arrange_shared_legend <- function(plotList, ncol=2, main=NULL, ...) {
    plots <- plotList
289 g <- ggplotGrob(plots[[1]] +
                    theme(legend.position="bottom",
291                       legend.title=element_blank()))$grobs
    legend <- g[[which(sapply(g, function(x) x$name) == "guide-box")]]
293 lheight <- sum(legend$height)
    plt.lst<-lapply(plots, function(x){
```

```
        x + theme(legend.position="none")
    })
    plt.lst$ncol<-ncol
    plt.lst$main<-main
    grid.arrange(
        do.call(arrangeGrob, plt.lst),
        legend,
        ncol = 1,
        heights = unit.c(unit(1, "npc") - lheight, lheight))
}



## ----data-prep, child="Include/DataPreperation.Rnw"----------------------

## ----dataSetup, echo=FALSE, message=FALSE, warning=FALSE, results='hide'----
baseTable<-read.xls(data.path, sheet = "FinalData")
baseTable[,1]<-as.Date(baseTable[,1], format="%d/%m/%Y")
baseTable[,"Testrain"]<-as.logical(baseTable[,"Testrain"])
# baseTable1<-baseTable

## Log Transform some variable using log1p() Function
## baseTable[, "ImpOldShip"]<-log1p(baseTable[, "ImpOldShip"])
# baseTable[, "ExpOilPlat"]<-log1p(baseTable[, "ExpOilPlat"])
# baseTable[, "ExpExShipOilPlat"]<-log1p(baseTable[, "ExpExShipOilPlat"])


## Label Variables in baseTable
labelTable<-read.xls(data.path, sheet = "FinalCodeBook", stringsAsFactors=FALSE
    )
for(i in 1:ncol(baseTable)){
    Hmisc::label(baseTable[,i])<-labelTable[i,2]
    class(baseTable[,i])<-rev(class(baseTable[,i]))
}

# Variable Declaration
y.var<-grep("PerEURO", names(baseTable), value=TRUE)
fin.var<-grep("^CPI|Int", names(baseTable), value=TRUE)
price.var<-grep("^Oil", names(baseTable), value=TRUE)
import.var<-grep("^Imp", names(baseTable), value=TRUE)
export.var<-grep("^Exp", names(baseTable), value=TRUE)
tradeBal.var<-grep("^Tr", names(baseTable), value=TRUE)
expct.var<-grep("^l", names(baseTable), value=TRUE)
y2.var<-grep("ExcCh", names(baseTable), value=TRUE)
season<-grep("season", names(baseTable), value=TRUE)
train<-grep("Testrain", names(baseTable), value=TRUE)
```

```
343
x.var<-c(fin.var, price.var, import.var, export.var, tradeBal.var, expct.var)
345 # baseTable$Testrain<-baseTable$Date<"2013-01-01"
train<-baseTable[,"Testrain"]
347
balTot<-balTot<-read.xls(file.path(dirname(data.path), "Balance of Payment
    Quarterly Data.xlsx"), sheet = "BalTot")
349 balTot<-balTot[-nrow(balTot),]
balTot$Date<-as.yearqtr(gsub("K", "Q", balTot$Date))
351
## Crisis Period
353 cperiod<-c("2007-06-01", "2009-06-01") ## Three Years of crisis Period


355



357
## ----AbvSymb-include, child="Include/AbvSymb.Rnw", eval=TRUE-------------
359

361 ## ----getSymb, echo=FALSE, warning=FALSE, message=FALSE, results='asis'----
Abv<-read.xls(file.path(dirname(data.path), "Symbols and Abbrivation.xlsx"),
    sheet = 1)
363 Symb<-read.xls(file.path(dirname(data.path), "Symbols and Abbrivation.xlsx"),
    sheet = 2)


365
## ----AbvPrint, echo=FALSE, results='asis'--------------------------------
367 AbvTbl<-xtable(Abv, caption = "Abbreviations and their full forms used in this
    Thesis", align = 'llX')
print(AbvTbl,
369       include.rownames = F,
      tabular.environment = "tabularx",
371       width = "\\textwidth", floating=FALSE,
      booktabs = TRUE, add.to.row = list(pos = list(0),
373                                         command = "\\hline \\endhead "),
      sanitize.text.function = function(x){x},
375       caption.placement = "top",
      table.placement = 'htbp')
377

379 ## ----symbPrint, echo=FALSE, results='asis'-------------------------------
SymbTbl<-xtable(Symb, caption = "Symbols and their meaning used in this Thesis"
    , align='llX')
381 print(SymbTbl,
      include.rownames = F,
383       tabular.environment = "tabularx",
      width = "\\textwidth", floating=FALSE,
385       booktabs = TRUE, add.to.row = list(pos = list(0),
                                         command = "\\hline \\endhead "),
```

```
387       sanitize.text.function = function(x){x},
          caption.placement = "top",
389       table.placement = 'htbp')

391

393
## ----chapter1-include, child="Include/Chapter-1.Rnw", eval=TRUE----------
395

397

399 ## ----chapter2-include, child="Include/Chapter-2.Rnw", eval=TRUE----------

401

403

405

407

409

411
## ----tsPlotExp, echo=FALSE, fig.height=5, fig.cap="Time Series plot of major
    exports of Norway", warning=FALSE, error=FALSE----
413 plotTS(baseTable[,c("Date", ls(baseTable, pattern = "Exp"))], 1, nc=2)

415

417

419

421

423
## ----chapter3-include, child="Include/Chapter-3.Rnw", eval=TRUE----------
425

427 ## ----mdlFitCriteriaPlot, child="mdlFitCriteriaPlot.Rnw"------------------

429

431

433
## ----chapter4-include, child="Include/Chapter-4.Rnw", eval=TRUE----------
```

```
435
437 ## ----sumryTablSetup, echo=FALSE, results='hide'-------------------------
    sumryTabl<-t(sapply(baseTable[,c(y.var, x.var)],
439                         function(x){c(min=min(x),
                                          median=median(x),
441                                       max=max(x),
                                          mean=mean(x),
443                                       stdev=sd(x))}))
    sumryXtable<-xtable(sumryTabl)
445
    ## Repeat Table Header Row for longtable ########
447 addtorow          <- list()
    addtorow$pos      <- list()
449 addtorow$pos[[1]] <- c(0)
    addtorow$command  <- c(paste("\\hline \n",
451                              "\\endhead \n",
                                 "\\hline \n",
453                              "{\\footnotesize Continued on next page} \n",
                                 "\\endfoot \n",
455                              "\\endlastfoot \n",sep=""))
    ## ------------------------ #########
457
    caption(sumryXtable)<-"Summary Report of all the variables used in this report"
459 label(sumryXtable)<-"tbl:sumryTabl"


461


463


465


467 ## ----commons, child="Include/Commons.Rnw"--------------------------------


469
    ## ----modelFitting, echo=FALSE, results='hide'----------------------------
471 pls.options(plsralg="oscorespls")
    mdl<-c("lm", "pcr", "plsr", "linearRidge")
473 mdl.ft<-lapply(seq_along(mdl),
                  function(x){
475                   do.call(fit.model, list(
                           mdl[x],
477                        y.var,
                           x.var,
479                        baseTable[train,],
                           scaling=c(mdl %in% c("plsr", "pcr"))[x]
481                   ))
                      })
483 names(mdl.ft)<-c("linear", "PCR", "PLS", "ridge")
```

```
## ------------------------------------------------------------------|
## Model selected by Mallows Cp Criteria
cp.leaps<-leaps(x=mdl.ft$linear$dataset[,x.var],
                y=mdl.ft$linear$dataset[,y.var],
                method="Cp", nbest = 1, names = x.var)

# Prepare for plot
cpdf<-data.frame(p=cp.leaps$size, cp=cp.leaps$Cp)

# Model fitting
cp.which<-names(which(cp.leaps$which[which.min(cp.leaps$Cp),]))
mdl.ft$cp.model<-do.call(fit.model, list("lm", y.var, cp.which, baseTable[train
    ,], scaling = FALSE))

## ------------------------------------------------------------------|
## Model selected by R-sq Adjusted Criteria
r2adj.leaps<-leaps(x=mdl.ft$linear$dataset[,x.var],
                y=mdl.ft$linear$dataset[,y.var],
                method="adjr2", nbest = 1, names=x.var)
# Prepare for plot
r2df<-data.frame(p=r2adj.leaps$size, r2adj=r2adj.leaps$adjr2)

# Model fitting
r2.which<-names(which(r2adj.leaps$which[which.max(r2adj.leaps$adjr2),]))
mdl.ft$r2.model<-do.call(fit.model, list("lm", y.var, r2.which, baseTable[train
    ,], scaling=FALSE))

## ------------------------------------------------------------------|
## Model selected by AIC and BIC criteria
lmBstSetSmry <- summary(regsubsets(mdl.ft$linear$dataset[,x.var],
                                   mdl.ft$linear$dataset[,y.var],
                                   nbest = 1, nvmax = length(x.var)))
nvars<-apply(lmBstSetSmry$which, 1, sum)
bic.vec<-lmBstSetSmry$bic
aic.vec<-bic.vec-nvars*log(sum(train))+nvars
infoMat<-data.frame(p=nvars, aic=aic.vec, bic=bic.vec)

## Fitting selected linear model
aic.which<-names(which(lmBstSetSmry$which[which.min(aic.vec),]))[-1]
bic.which<-names(which(lmBstSetSmry$which[which.min(bic.vec),]))[-1]

mdl.ft$aicMdl<- do.call(fit.model, list("lm", y.var, aic.which, dataSet =
    baseTable[train,], scaling = F))
mdl.ft$bicMdl<- do.call(fit.model, list("lm", y.var, bic.which, dataSet =
    baseTable[train,], scaling = F))

## ------------------------------------------------------------------|
## Forward Selection Model (criteria: level of significance)
```

```
fw.model.log <- capture.output(fw.model<-forward(lm(formula = mdl.ft$linear$
    formula, data=mdl.ft$linear$data), alpha = 0.1, full = FALSE))
mdl.ft$forward<-list(formula=mdl.ft$linear$formula, model=fw.model, data=mdl.ft
    $linear$data)

## Backward Elimination Model (criteria: level of significance)
bw.model.log<-capture.output(bw.model<-backward(lm(formula = mdl.ft$linear$
    formula, data=mdl.ft$linear$data), alpha = 0.1, full = FALSE, hierarchy =
    TRUE))
mdl.ft$backward<-list(formula=mdl.ft$linear$formula, model=bw.model, data=mdl.
    ft$linear$data)

## ------------------------------------------------------------------|
## Labeling the models
mdl.labels<-c("Linear Model", "Principal Component Regression", "Partial Least
    Square Regression", "Ridge Regression", "Subset Model (criteria:Mallows Cp)
    ", "Subset Model (criteria:R-sq adjusted)", "Model selected (criteria:AIC)"
    ,"Model selected (criteria:BIC)", "Forward Selection Model(criteria:F-test)
    ", "Backward Elimination Model (criteria: F-test)")
mdl.prnt.lab<-c("Linear Model", "Principal Component \\\\ Regression", "Partial
     Least Square \\\\ Regression", "Ridge Regression", "Subset Model \\\\ (
    criteria:Mallows Cp)", "Subset Model \\\\ (criteria:R-sq adjusted)", "Model
     selected \\\\ (criteria:AIC)","Model selected (criteria:BIC)", "Forward
    Selection Model \\\\ (criteria:F-test)", "Backward Elimination Model \\\\ (
    criteria: F-test)")

for(i in 1:length(mdl.ft)){
    # Label the model
    Hmisc::label(mdl.ft[[i]][[2]])<-mdl.labels[i]
    # Reverse the class
    class(mdl.ft[[i]][[2]])<-rev(class(mdl.ft[[i]][[2]]))
}

## ------------------------------------------------------------------|
## Principal Component Analysis
pc.a<-princomp(baseTable[, x.var], cor = TRUE, )

## ------------------------------------------------------------------|
## Setting up Ridge Parameter lambda
lmd.seq<-seq(0,0.01,0.0005)
tuningRidge<-ldply(lmd.seq, function(x){
    rdg.rmsep<-mdl.cv(baseTable[train,], x.var, y.var,
                      model="ridge", split=12, lmd = x)$rmsep
    rdg.r2pred<-mdl.cv(baseTable[train,], x.var, y.var,
                       model="ridge", split=12, lmd = x)$r2pred
    data.frame(lmd=x, rmsep=rdg.rmsep, r2pred=rdg.r2pred)
})
tuningRidge<-data.frame(tuningRidge)
lmd<-lmd.seq[which.min(tuningRidge$rmsep)]
```

```
## ------------------------------------------------------------------|
## Updating Linear Ridge model with new paramter lmd
mdl.ft$ridge$model<-linearRidge(mdl.ft$ridge$formula,
                                data=mdl.ft$ridge$dataset,
                                lambda = lmd)

## Color for crisis period
cperiod.col<-cp.cat(cperiod)



## ----chapter4a-include, child="Include/Chapter-4a.Rnw", eval=TRUE--------

## ----sigCoef, echo=FALSE, warning=FALSE, results='hide'------------------
coefMat<-as.data.frame(summary(mdl.ft$linear$model)$coefficients)
sigVarIdx<-which(coefMat$`Pr(>|t|)`<=0.05)









## ----chapter4b-include, child="Include/Chapter-4b.Rnw", eval=TRUE--------

## ----pcaSumrySetup, echo=FALSE, results='hide'---------------------------
stdev<-pc.a$sdev
varprop<-pc.a$sdev^2/sum(pc.a$sdev^2)
pcaSumry<-data.frame(cbind( `Comp`=1:length(varprop),
                            `Std.Dev`=stdev,
                            `Var.Prop`=varprop,
                            `Cum.Var.Prop`=cumsum(varprop)))
pcaSumry$Comp<-1:nrow(pcaSumry)
pcaSumry1<-xtable(cbind(pcaSumry[1:7,], pcaSumry[8:14,]), digits = 3)
caption(pcaSumry1)<- "Dispersion of data explained by principal components"
label(pcaSumry1)<- "tbl:pcaSumry"
```

```
613 align(pcaSumry1)<- "lrrrr|rrrr"

615

617
## ----pcrSumrySetup, echo=FALSE, results='hide'--------------------------
619 pcr.expVar.x<-cumsum(explvar(mdl.ft$PCR$model))
pcr.expVar.y<-apply(fitted(mdl.ft$PCR$model), 3, var)/var(mdl.ft$PCR$dataset[,y
    .var])*100
621 pcrSumry<-data.frame(Comp=1:length(pcr.expVar.x),
                       X=pcr.expVar.x,
623                    PerEURO=pcr.expVar.y,
                       row.names = NULL)
625

627

629

631 ## ----chapter4c-include, child="Include/Chapter-4c.Rnw", eval=TRUE--------

633
## ----plsSumry, echo=FALSE, results='hide'-------------------------------
635 pls.expVar.x<- cumsum(explvar(mdl.ft$PLS$model))
pls.expVar.y<-apply(fitted(mdl.ft$PLS$model), 3, var)/var(mdl.ft$PCR$dataset[,y
    .var])*100
637 plsSumry<-data.frame(Comp=1:length(pls.expVar.x), X=pls.expVar.x, PerEURO=pls.
    expVar.y, row.names = NULL)

639

641 ## ----PLSnPCRcomp, echo=FALSE, results='hide'-----------------------------
PLSnPCRcomp<-melt(list(`PCR Model`=list(`Predictor Variable`=pcr.expVar.x,
643                                     `Response Variable`=pcr.expVar.y),
                       `PLS Model`=list(`Predictor Variable`=pls.expVar.x,
645                                     `Response Variable`=pls.expVar.y)))
names(PLSnPCRcomp)<-c("Variance Explained", "type", "model")
647 PLSnPCRcomp$Components<-factor(1:length(pcr.expVar.x), levels = 1:length(pcr.
    expVar.x))

649

651

653
## ----chapter4d-include, child="Include/Chapter-4d.Rnw", eval=TRUE--------
655

657 ## ----rmsepPLSnPCR, echo=FALSE---------------------------------------------
```

```
## Fitting PCR and PLS using Cross-validation
pcr.cv<-pcr(mdl.ft$PCR$formula, data=mdl.ft$PCR$dataset,
            scale=TRUE, validation="CV", segments=12,
            segments.type="consecutive")
pls.cv<-plsr(mdl.ft$PCR$formula, data=mdl.ft$PCR$dataset,
            scale=TRUE, validation="CV", segments=12,
            segments.type="consecutive")
## RMSEP using Cross-validation
rmsep.pcr<-data.frame(comp=RMSEP(pcr.cv)$comps,
                      r2pred=as.vector(R2(pcr.cv)$val),
                      t(sapply(RMSEP(pcr.cv)$comps,
                               function(x){RMSEP(pcr.cv)$val[,,x+1]})))
rmsep.pls<-data.frame(comp=RMSEP(pls.cv)$comps,
                      r2pred=as.vector(R2(pls.cv)$val),
                      t(sapply(RMSEP(pls.cv)$comps,
                               function(x){RMSEP(pls.cv)$val[,,x+1]})))
rmsep.mat<-melt(list(PCR=rmsep.pcr, PLS=rmsep.pls), 1)

## ----cvStat, echo=FALSE-----------------------------------------------
pcr.sc<-15:17
pls.sc<-6:9

lm.cv<-mdl.cv(baseTable[train,], x.var, y.var)
aic.cv<-mdl.cv(baseTable[train,], x.var, y.var, step = TRUE, criteria = "AIC",
    split = 12)
bic.cv<-mdl.cv(baseTable[train,], x.var, y.var, step = TRUE, criteria = "BIC",
    split = 12)
backward.cv<-mdl.cv(baseTable[train,], x.var, y.var, step = TRUE, criteria = "
    backward", split = 12)
ridge.cv<-mdl.cv(baseTable[train,], x.var, y.var, step=FALSE, split=12, model =
     "ridge", lmd = lmd)

rmse.cv<-data.frame(RMSEP=c(Linear=lm.cv$rmsep,
            AICModel=aic.cv$rmsep,
            BICModel=bic.cv$rmsep,
            BackModel=backward.cv$rmsep,
            Ridge=ridge.cv$rmsep,
            PCR=rmsep.pcr[rmsep.pcr$comp%in%pcr.sc, "adjCV"],
            PLS=rmsep.pls[rmsep.pls$comp%in%pls.sc, "adjCV"]))
r2pred.cv<-data.frame(R2pred=c(Linear=lm.cv$r2pred,
            AICModel=aic.cv$r2pred,
            BICModel=bic.cv$r2pred,
            BackModel=backward.cv$r2pred,
            Ridge=ridge.cv$r2pred,
            PCR=rmsep.pcr[rmsep.pcr$comp%in%pcr.sc, "r2pred"],
            PLS=rmsep.pls[rmsep.pls$comp%in%pls.sc, "r2pred"]))
cvStat<-data.frame(rmse.cv, r2pred.cv)
rownames(cvStat)[grep("PCR", rownames(cvStat))]<-paste("PCR.Comp", pcr.sc, sep=
    "")
```

```
rownames(cvStat)[grep("PLS", rownames(cvStat))]<-paste("PLS.Comp", pls.sc, sep=
    "")

pls.min.comp<-as.numeric(summarize(cvStat[grep("PLS", rownames(cvStat)), ], pls
    .sc[which.min(RMSEP)]))
pcr.min.comp<-as.numeric(summarize(cvStat[grep("PCR", rownames(cvStat)), ], pcr
    .sc[which.min(RMSEP)]))



## ----predMat, echo=FALSE-----------------------------------------------
lm.pred<-predict(mdl.ft$linear$model,
                      newdata = baseTable[!train, x.var])
pcr.pred<-list()
pls.pred<-list()
pcr.pred<-lapply(pcr.sc, function(x){as.vector(predict(mdl.ft$PCR$model,
                                    newdata = baseTable[!train, x.var],
                                    ncomp = x))})
pls.pred<-lapply(pls.sc, function(x){as.vector(predict(mdl.ft$PLS$model,
                                    newdata=baseTable[!train, x.var],
                                    ncomp=x))})
names(pcr.pred)<-paste("Comp",pcr.sc, sep="")
names(pls.pred)<-paste("Comp",pls.sc, sep="")

ridge.pred<-predict(mdl.ft$ridge$model,
                        newdata = baseTable[!train, x.var])
cp.model.pred<-predict(mdl.ft$cp.model$model,
                       newdata=baseTable[!train, x.var])
aicMdl.pred<-predict(mdl.ft$aicMdl$model,
                       newdata=baseTable[!train, x.var])
bicMdl.pred<-predict(mdl.ft$bicMdl$model,
                       newdata=baseTable[!train, x.var])
backward.pred<-predict(mdl.ft$backward$model,
                       newdata=baseTable[!train, x.var])
## Predicting Testset
predMat.test<-data.frame(Date=baseTable[!train, "Date"],
                     TrueValue=baseTable[!train, "PerEURO"],
                     Linear=lm.pred,
                     AICModel=aicMdl.pred,
                     BICModel=bicMdl.pred,
                     BackModel=backward.pred,
                     Ridge=ridge.pred,
                     PCR=pcr.pred,
                     PLS=pls.pred)

## Predicting Trainset
predMat.train<-data.frame(Date=baseTable[train, "Date"],
                  TrueValue=baseTable[train, "PerEURO"],
```

```
                Linear=predict(mdl.ft$linear$model),
749             AICModel=predict(mdl.ft$aicMdl$model),
                BICModel=predict(mdl.ft$bicMdl$model),
751             BackModel=predict(mdl.ft$backward$model),
                Ridge=predict(mdl.ft$ridge$model),
753             PCR=predict(mdl.ft$PCR$model, ncomp = pcr.sc),
                PLS=predict(mdl.ft$PLS$model, ncomp = pls.sc))
755
names(predMat.train)[grep("PCR", names(predMat.train))]<-paste("PCR.Comp", pcr.
    sc, sep="")
757 names(predMat.train)[grep("PLS", names(predMat.train))]<-paste("PLS.Comp", pls.
    sc, sep="")

759 predMat<-rbind(train=predMat.train, test=predMat.test)
stkPredMat<-melt(list(train=predMat.train, test=predMat.test), 1:2)
761 stkPredMat$L1<-factor(stkPredMat$L1, levels = c("train", "test"))

763 predMat.rpSumry<-ddply(stkPredMat, .(variable, L1), summarize,
      RMSEP=sqrt(1/length(value)*sum((TrueValue-value)^2)),
765       R2pred=1-(sum((TrueValue-value)^2)/sum((TrueValue-mean(TrueValue))^2)))

767 ## ----testPredErr, echo=FALSE-----------------------------------------------
errMat<-lapply(3:ncol(predMat.test), function(x){rmserr(predMat.test[,2],
    predMat.test[,x])})
769 names(errMat)<-names(predMat.test)[-c(1:2)]
errStkMat<-melt(errMat)
771 errStkMat$L1<-factor(errStkMat$L1, levels = names(errMat))

773
## ----gofSumry, echo=FALSE--------------------------------------------------
775 gofSumry<-ldply(names(mdl.ft)[-c(2:4)], function(x){
    data.frame(Model=x,
777                 AIC=AIC(mdl.ft[[x]][[2]]),
                    BIC=AIC(mdl.ft[[x]][[2]],
779                         k = log(nrow(mdl.ft[[x]][[3]]))),
                    `R-Sq`=summary(mdl.ft[[x]][[2]])$r.squared,
781                 `R-Sq Adj`=summary(mdl.ft[[x]][[2]])$adj.r.squared,
                    `Sigma`=summary(mdl.ft[[x]][[2]])$sigma,
783                 `F-value`=summary(mdl.ft[[x]][[2]])$fstat[1],
                    `P-value`=signif(pf(summary(mdl.ft[[x]][[2]])$fstat[1],
785                              summary(mdl.ft[[x]][[2]])$fstat[2],
                                 summary(mdl.ft[[x]][[2]])$fstat[3],
787                              lower.tail = FALSE), 3))
})
789


791
## ----ValdSumry, echo=FALSE, results='hide'---------------------------------
```

```
ValdSumry<-rbind(predMat.rpSumry, data.frame(variable=rownames(cvStat), L1="cv"
    , cvStat, row.names = NULL))
names(ValdSumry)<-c("Model", "Type", "RMSEP", "R2pred")
vs.cast<-dcast(melt(ValdSumry, 1:2), Model~Type+variable)[, c(1:3,6:7,4:5)]

ValdSumryTabl<-xtable(vs.cast, digits = 4)
caption(ValdSumryTabl)<-"Validation result containing RMSEP and R2pred for
    training set, cross-validation set and test set"
label(ValdSumryTabl)<-"tbl:valdSumry"
align(ValdSumryTabl)<-"lrrrrrrr"
tblHeader<-paste("\\hline Model &
                 \\multicolumn{2}{c}{Training} &
                 \\multicolumn{2}{c}{Cross Validation} &
                 \\multicolumn{2}{c}{Test} \\\\
                 \\cline{2-7} &",
                 paste(rep(c("RMSEP", "R2pred"), 3),
                       collapse=" & "),
                 '\\\\')

## ----ValdSumryPlotSetup, echo=FALSE-----------------------------------
vss<-ddply(ValdSumry, .(Type), summarize,
      Model.rmsep=Model[which.min(RMSEP)],
      Model.r2pred=Model[which.max(R2pred)],
      RMSEP=min(RMSEP),
      R2pred=max(R2pred))
vss1<-filter(melt(vss,1:3), variable=='RMSEP')[,-3]
vss2<-filter(melt(vss,1:3), variable=='R2pred')[,-2]
names(vss1)<-names(vss2)<-c("Type", "Model", "variable", "value")
vss<-rbind(vss1, vss2)



## ----coefMat, echo=FALSE----------------------------------------------
coefMat<-cbind(sapply(c(1,4), function(x){coef(mdl.ft[[x]][[2]])[-1]}),
               coef(mdl.ft$PCR$model, ncomp = pcr.min.comp),
               coef(mdl.ft$PLS$model, ncomp=pls.min.comp))
coefMat<-data.frame(variable=rownames(coefMat), coefMat, row.names = NULL)
names(coefMat)<-c("vars","linear", "ridge", "pcr", "pls")
coefMat$vars<-factor(coefMat$vars, levels = coefMat$vars[order(coefMat$linear)
    ])



## ----coefMatPrint, echo=FALSE, results='asis'-------------------------
coefMatxTable<-xtable(coefMat[c(order(coefMat$pcr, coefMat$pls)[1:4],
```

```
                                order(coefMat$pcr, coefMat$pls,
                                      decreasing = TRUE)[1:4]), c(1,4:5)],
                      digits = 4,
                      caption = c("Top four (both positive and negative)
    Coefficient Estimate of PLS and PCR model", "Coefficient Estimate for PLS
    and PCR model"),
                      label="tbl:coefEstdTbl")
print.xtable(coefMatxTable, include.rownames = FALSE,
             caption.placement = 'top', table.placement = 'htp')


## ----forecast, echo=FALSE, fig.cap="Prediction made on trained and test
    dataset using different models", fig.height=9.5, fig.width="\\textwidth
    "----
ggplot(stkPredMat, aes(Date, value))+
    geom_line(aes(color="red"))+
    facet_wrap(~variable+L1,
               scale="free_x",
               ncol = 4)+
    geom_line(aes(y=TrueValue, color="blue"),
              shape=21)+
    theme_bw()+
    theme(axis.text.x=element_text(angle=45, hjust=0.5, vjust=0.5),
          text=element_text(size=9),
          legend.title=element_blank(),
          legend.position="top")+
    geom_text(data=predMat.rpSumry,
              aes(label=paste("RMSEP:", round(RMSEP, 3),
                              "\nR2pred", round(R2pred, 3))),
              x=-Inf, y=Inf, hjust=-0.1, vjust=1.1, size=2.5)+
    scale_color_manual(values=c("red", "blue"),
                       labels=c("Predicted", "Original"))


## ----chapter5-include, child="Include/Chapter-5.Rnw", eval=TRUE----------


## ----appendixVarUsed, child="Include/Appendix-varUsed.Rnw", eval=TRUE----


## ----dataDescData, echo=FALSE, warning=FALSE, results='hide'-------------
dataDescription<-read.xls(data.path, sheet = 2)


## ----dataDescTable, echo=FALSE, results='asis'--------------------------
dataDescription[,1]<-paste("\\texttt{", dataDescription[,1], "}", sep="")
```

```
names(dataDescription)[1:2]<-c("Code", "Description")
dataDescTab<-xtable(dataDescription[,1:2], align = "llX", caption = "Variable codes and their descriptions used in this paper")
print(dataDescTab, include.rownames = F, tabular.environment = "tabularx", width = "\\textwidth", floating=FALSE, booktabs = TRUE, add.to.row = list(pos = list(0),command = "\\hline \\endhead "), sanitize.text.function = function(x){x})


## ----pkgsUsed, child="Include/Appendix-pkgsUsed.Rnw", eval=TRUE----------

## ----pkgsUsed, echo=FALSE-------------------------------------------------
pkgsDesc<-ldply(c(req.package, "graphics", "grDevices", "utils", "datasets", "methods", "base"), function(x){
    data.frame(
    `Package Name`=packageDescription(x)$Package,
    `Version`=packageDescription(x)$Version,
    `Title`=packageDescription(x)$Title)
})
citeKey<-c('car2011FJnWS','dplyr2014WHFR','gdata2014WG','ggplot22009WH','gridExtra2012AB','knitr2013XY','leaps2009LT','MASS2001WNV','mixlm2014SK','pls2013MBH','plyr2011WH','R2014Rcore','reshape22007WH','scales:2014Wickham','ridge2014CE','xtable2014DD','zoo2005ZAGG')
ckSrtd<-unlist(lapply(paste("^",pkgsDesc$Package.Name, sep=""), function(x){
    grep(x, x = citeKey, value = TRUE)
}))
ckSrtd<-c(ckSrtd,rep('R2014Rcore', 6))
citeCmd<-paste("\\cite{",ckSrtd,"}", sep="")


## ----revPlots, child="Include/Appendix-revPlots.Rnw", eval=TRUE----------


## ----appendixCodeUsed, child="Include/Appendix-PLSflowchart.Rnw", eval=TRUE----
```

```
925

927
## ----appendixCodeUsed, child="Include/Appendix-codeUsed.Rnw", eval=TRUE----
```