GDP FORECASTING MODELS USING DIMENSIONALITY REDUCTION TECHNIQUES

STUDY ORIENTED PROJECT

Submitted by:

KATHAN KASHIPAREKH 2014B3A7792G

GUIDE: Dr. Anoop S Kumar

Completed in partial fulfilment of the course

ECON F366 Study Project

ACKNOWLEDGEMENTS

I would like to thank my Study Oriented Project guide and mentor Dr. Anoop S Kumar for helping me throughout this project. His insights and outlook on various macro-economic and micro-economics concepts were of invaluable importance in understanding the topic, data and the results as well. His mentorship has not only allowed me to look at such concepts differently but also provided me with sufficient knowledge in this area.

INDEX

ABSTRACT

This paper aims at analysing the forecasting accuracy of a GDP forecasting model based on a large number of macro-economic variables that tend to have a major impact on the GDP of India. Principal Component Analysis has been used to reduce the dimensionality of the dataset following which four most important such principal components are extracted and used to fit a linear regression model for forecasting GDP using OLS regression. It is shown that the assumption of linearity in data is quite strong and in order to make accurate forecasts we need to go for more intricate models that capture the non-linearites in the data.

Keywords: GDP Forecasting, Principal Component Analysis, OLS Regression

Introduction:

In order to assess the performance of a country, the de facto parameter is the Gross Domestic Product of a nation. The Gross Domestic Product (GDP) is a numerical measure for the value of all goods and services produced within the borders of a nation. This sole indicator helps economists and policy makers alike in order to gauge how a nation is progressing and in which areas should the new policies be implemented, if any. Every nation is unique in its own sense and has some areas where it is pre-dominantly strong, for example, the Middle Eastern countries are the major supplier of oil to the world market and earn majority of their revenue from oil exports to the world. Similarly, while India is known for its precious stones, parts of Africa are known for their coffee and tea exports. Thus in order to calculate a nations GDP, a vast multitude of factors need to be considered. That being said, accurate GDP forecasting is one the major priorities of a nation in order to implement better and more importantly, implement the right policies. Variety of factors ranging from the forex market, money market, total production of various sectors, inflation, exchange rates etc affect the GDP and need to be accounted for. That being said, the major aim of this report is to

1) Create a GDP forecasting model based on simple OLS regression.

2) Use dimensionality reduction techniques like the Principal Component Analysis (PCA) in order to concisely represent all the variables and make OLS estimates significant.

Literature Review:

The Government of India uses outputs and production of a large number of sectors in order to reach a somewhat correct measure of the Indian GDP. Collecting adequate and correct data is always challenging and thus it is difficult to reach a proper value for the Indian GDP.

The widely used methods for GDP predictions have their groundwork based upon the famous work done by James Stock and Mark Watson (2002) which primarily focuses on forecasting one series using multiple series which explain the dependent series. They cite the most frequent problem found in such analysis: number of candidate series (N) are very large as compared to the number of time series observations (T). In order to overcome this problem they suggest first to model the covariability in the time series into latent factors which are smaller in number than the observations (T). Then these latent factors are used to forecast the predictor series using linear regression.

Calista Cheung and Frederick Demers (2007) from the Bank of Canada used Factor Models to forecast Canadian GDP growth and Core Inflation. They compared various static and dynamic factor models to forecast the Canadian GDP. The results are compared with an AR(p) model. Their results show that factor models improve the forecast accuracy relative to the standard benchmark models for upto 8 quarters ahead. They used out of sample forecasting to evaluate their models which is a more robust technique.

Franciso Dias, Maximiano Pinheiro and Antonio Rua used factor models to forecast Portuguese GDP. They primarily focus on the Diffusion Index model as proposed by Stock and Watson. Their results match with the results of Calista and Frederick that factor models significantly outperform the univariate autoregressive models for forecasting.

Chrsistian Schumacher (2005) used static principal component analysis and dynamic principal component analysis using frequency domain models to forecast German GDP using quarterly data. He also compares these models with subspace algorithms for state space models. Like the above reviews, even this paper concludes that factor models result in a smaller error rate as compared to autoregressive models. However, the dynamic factor models don’t result in significant improvement over the static PCA.

Graem Chamberlain (2007) compares forecasting after using PCA and a simple linear regression model to forecast GDP. The conclusions point towards the fact that using the first 4 principal components that explain ~ 100% of data leads to better results than the simple OLS estimates. He also compares his results with the quarterly forecasts of the Office of National Statistics (ONS) which are revised as per the new data received. The aim of the paper is to develop a method which would reduce the necessity for regular updates on the forecasts. The data has been collected from two major sources: Business and Consumer surveys and Monetary and Financial Data.

Nikolaos Sdrakas and Cedric Viguie (2003) have employed a VAR model to forecast GDP of the euro area. Even they used PCA to reduce dimension of the data. Consumer confidence indicators and retail trade confidence indicators have been included to make the forecast more robust. A simple auto-regressive model was also employed. However the VAR model boosted by the PCA out-performed that auto-regressive model.

Christian Dreger Christian Schumacher (2002) employed four techniques for GDP forecasting. Apart from the standard VAR and dynamic and static factor models they also accounted for the forecasts based on the ifo business climate indicatior which is a business cycle indicator obtained from surveys. It has also been included in the factor models to make them more robust. The results show that the dynamic factor models outperform all other models at all lags of step ahead predictions. Followed by this was the ifo model which performed pretty well as well. However in terms of error rate improvement, factor models don't provide much gains.

Data Sources:

In order to forecast GDP, large amount of variables need to be taken care of since a country earns in money from a variety of sources. 31 macro-economic aggregates have been considered in order to predict the GDP. Details about each of the time series' taken is mentioned in the appendix. Data has been collected from RBI Handbook of Indian economy and EPWR Indian Time Series. Annual data starting from 1985 till 2012 has been collected. Where annual data was not available, adequate averaging has been done to obtain the best estimate. All values are either in % or in Indian Rupees. GDP (% growth) has been chosen as the dependent variable in the OLS regression that follows the Principal Component Analysis. The final data matrix is a 28*31 dimension matrix with 28 observations and 31 variables.

Methodology:

As can be seen from the data sources, the number of variables are more than the number of observations. Simple Linear Regression thus can't be performed as OLS estimates are not BLUE estimates. Thus, in order to capture all the information presented by all the variables the idea of dimensionality reduction has been used. Principal Component Analysis (PCA) has been used to reduce the dimension of the variables without losing the information captured by them. To begin with observations for 31 macro-economic variables have been taken for 28 years starting from 1985 to 2012. Due to the aforementioned issue, OLS regression can't be directly applied to the dataset. PCA has been used to reduce the number of explanatory variables. From the results of PCA, it can be seen that the first 5 principal components explain around 98% of the total variation in the 31 macro-economic variables and thus can be used as a proxy for all the 31 of them. Following this, the 5 principal components are extracted and their coefficients used to create 5 new variables using which Linear Regression using OLS estimates has been run on GDP (% growth). Coefficients of the Linear Regression are identified along with the standard analysis of the coefficient of determination and the F-statistic.

Results:

Appendix A2 results of Principal Component Analysis shows that around 98.5% of variance in the 31 macro-economic variables is explained by the first 4 principal components which are each a linear combination of the 31 macro-economic time series. Since these 4 components explain a majority of the data set, they have been taken as the explanatory variables to perform the OLS Regression. In order to do so, the coefficients obtained from the PCA have been multiplied with the original 31 variable series to get a reduced dimension of 4 for the explanatory variables. The Linear Regression results shows that all the coefficients are statistically significant at 5% level of significance and a p-value of 0.006397 indicated a statistically significant model as a whole. However the R^2 value of 0.4499 indicates that 50% of the variation in GDP is explained by the chosen 4 principal components. This can be attributed to two factors.

(i) The assumption of linearity in data is quite strong and it is very unlikely that there exists a linear relation between GDP and the macro-economic data and can be modelled more accurately with a non-linear relationship.

(ii) GDP forecasting is a very complex process and is affected by a vast multitude of parameters. Thus more number of explanatory variables are needed that accurately model the output of a nation.

Conclusions:

The work's main aim is to present a simple GDP forecasting model. GDP is one of the most important indicators of a country's overall progress and is affected by a variety of parameters. All macro-economic data relevant to the Indian economy have been chosen as the explanatory variables in the model. Due to being large is number PCA has been applied to reduce their dimension in order to run an OLS regression. The coefficients of the linear regression are all significant indicating all parameters are important in explaining the variation in GDP. However the R^2 value is not quite high indication not much influence of the variables on GDP variance. This probably arises due to the dependence of GDP on a large number of parameters and it is difficult to not only quantify some of them but also gather such grass root level data.

It should be noted that this simple model assumes a linear relation between GDP and the macro-economic variables. This assumption is highly unlikely to occur in real life. Also, this is just an entry level primary work and in no way should be considered an accurate forecast of the GDP. Ample amount of work can be done if relevant data and prior knowledge regarding dependency of various macro-economic aggregates on GDP is known.

References:

[1] Amisano, Gianni, and John Geweke. "Prediction Using Several Macroeconomic Models". *Review of Economics and Statistics* (2017): n. pag. Web.

[2] Cheung, Calista, and Frédérick Demers. "Evaluating Forecasts From Factor Models For Canadian GDP Growth And Core Inflation". (2007): n. pag. Print.

[3] Dias, Francisco, Maximiano Pinheiro, and António Rua. "Forecasting Portuguese GDP With Factor Models: Pre- And Post-Crisis Evidence". *Economic Modelling* 44 (2015): 266-272. Web.

[4] Schumacher, Christian, and Jörg Breitung. "Real-Time Forecasting Of GDP Based On A Large Factor Model With Monthly And Quarterly Data". *SSRN Electronic Journal* n. pag. Web.

[5] Schumacher, Christian. "Forecasting German GDP Using Alternative Factor Models Based On Large Datasets". *Journal of Forecasting* 26.4 (2007): 271-302. Web.

[6] Stock, James H, and Mark W Watson. "Forecasting Using Principal Components From A Large Number Of Predictors". *Journal of the American Statistical Association* 97.460 (2002): 1167-1179. Web.

[7] Chamberlin, Graeme. "Forcasting GDP Using External Data Sources". *Economic & Labour Market Review* 1.8 (2007): 18-23. Web.

[8] Schumacher, Christian, and Christian Dreger. "Estimating Large-Scale Factor Models For Economic Activity Ingermany: Do They Outperform Simpler Models?". *SSRN Electronic Journal* n. pag. Web.

[9] Sdrakas, Nikolaos, and Cedric Viguie. "VAR Modelling Of The Euro Area GDP On The Basis Of Principal Component Analysis". (2003): n. pag. Print.

Appendix:

A1: Details about the explanatory macro-economic variables

1. Foreign Accounts: Total Exports, Total Imports, Net Trade Balance, Balance of Payments (BOP), Foreign Exchange Reserves.

2. National Accounts: Net national disposable Income, Indirect Taxes, Gross Fiscal Deficit, Gross Final Deficit, Consumption of Fixed Capital, Capital formation, Gross Domestic Savings, Gross Total Debt, Private Final Consumption Expenditure, Government Final Consumption Expenditure, Total Bank Credit.

3. Fixed Income Securities, Money and Call Market: 182 and 91 day Indian treasury bills, Sensex closing values, INR-USD Exchange rate.

4. Inflation and Production Indices: Index of Industrial Production, Wholesale Price Index (WPI), Consumer Price Index (CPI).

Others: Total receipts, Total Expenditure.

A2: Complete R results

- **Summary of Principal Component Analysis:**

-
  ```
  ## Importance of components:
  ##                           PC1     PC2    PC3     PC4     PC5     PC6
  ## Standard deviation     4.8910 1.48232 1.2363 0.78558 0.55125 0.46666
  ## Proportion of Variance 0.8249 0.07577 0.0527 0.02128 0.01048 0.00751
  ## Cumulative Proportion  0.8249 0.90066 0.9534 0.97465 0.98513 0.99264
  ```

```
##                            PC7    PC8     PC9    PC10    PC11    PC12
## Standard deviation     0.37256 0.1783 0.14703 0.09570 0.06045 0.05341
## Proportion of Variance 0.00479 0.0011 0.00075 0.00032 0.00013 0.00010
## Cumulative Proportion  0.99742 0.9985 0.99926 0.99958 0.99971 0.99980
##                           PC13    PC14    PC15    PC16    PC17    PC18
## Standard deviation     0.05034 0.03882 0.02583 0.02058 0.01702 0.01081
## Proportion of Variance 0.00009 0.00005 0.00002 0.00001 0.00001 0.00000
## Cumulative Proportion  0.99989 0.99994 0.99997 0.99998 0.99999 0.99999
##                            PC19     PC20     PC21     PC22    PC23
## Standard deviation     0.008586 0.006191 0.004966 0.002934 0.00225
## Proportion of Variance 0.000000 0.000000 0.000000 0.000000 0.00000
## Cumulative Proportion  1.000000 1.000000 1.000000 1.000000 1.00000
##                            PC24     PC25      PC26      PC27      PC28
## Standard deviation     0.001523 0.000967 0.0004255 0.0001309 4.563e-16
## Proportion of Variance 0.000000 0.000000 0.0000000 0.0000000 0.000e+00
## Cumulative Proportion  1.000000 1.000000 1.0000000 1.0000000 1.000e+00
```

- **Linear Regression results:**

Linear Regression Equation:

GDP= $\beta_0 + \beta_1$*PC1+ β_2*PC2+ β_3*PC3+ β_4*PC4

Co-efficient	**Estimate (p-value)**
β_0	4.665 (6.47 e-05)
β_1	-2.603e-05 (0.0438)
β_2	2.723e-04 (0.0528)
β_3	1.574e-04 (0.0422)
β_4	-2.301e-04 (0.0522)

```
## Residual standard error: 1.903 on 20 degrees of freedom
## Multiple R-squared:  0.3909, Adjusted R-squared:  0.2691
## F-statistic: 3.209 on 4 and 20 DF,  p-value: 0.03448
```

Forecasting Results:

```
##           Forecast    Actual
## 2010-11 11.812677 10.259963
## 2011-12  5.995784  6.638364
## 2012-13  3.850475  5.618563
```

A3: Full R code used for the analysis:

~~~
---
title: "3-2 SOP"
author: "Kathan"
date: "5 April 2017"
output: word_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```


```{r,echo=FALSE}
data1<-read.csv('Small Dataset.csv')
GDP<-data.frame(data1[,2])
PCA_data<-data1[,3:31]
pca<-prcomp(PCA_data,scale=TRUE)
summary(pca)
coeffs<-pca$rotation
coeffs<-coeffs[,c(1,2,3,4,5)]
~~~

~~~
features<-as.matrix(PCA_data)%*%as.matrix(coeffs)
features<-data.frame(features)
final_data<-data.frame(GDP,features)
colnames(final_data)<-c("GDP","PC1","PC2","PC3","PC4","PC5")
linear_reg<-lm(GDP~PC1+PC2+PC3+PC4,data=final_data[1:25,])
summary(linear_reg)
forecast1=predict(linear_reg,final_data[26,])
forecast2=predict(linear_reg,final_data[27,])
forecast3=predict(linear_reg,final_data[28,])
a=rbind(forecast1,forecast2,forecast3)
res=cbind(a,final_data[26:28,1])
results=data.frame(res)
colnames(results)<-c("Forecast","Actual")
rownames(results)<-c("2010-11","2011-12","2012-13")
results
```
~~~