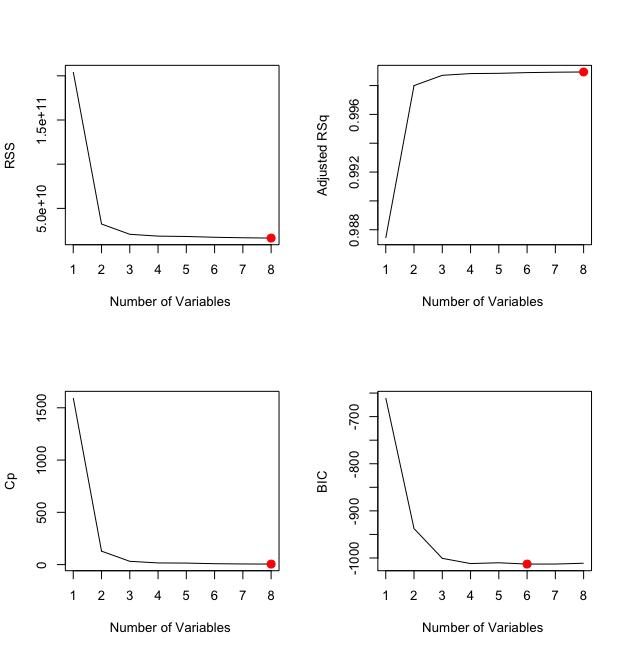
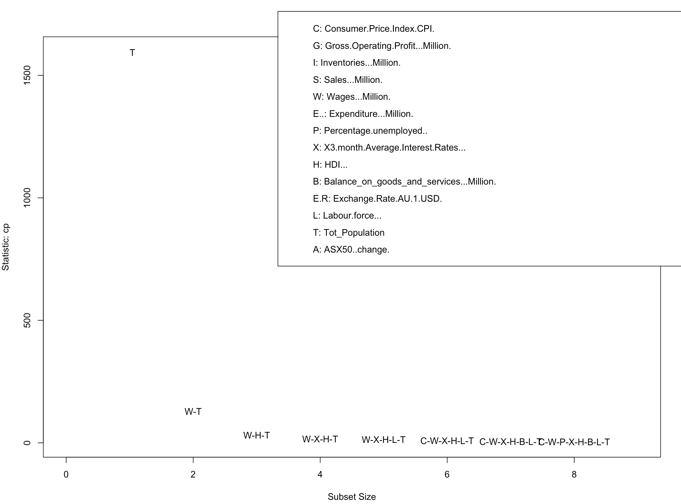
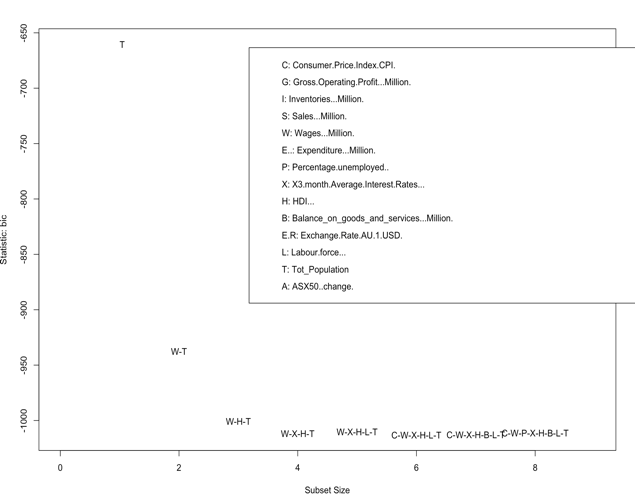
**Multivariate linear regression:**

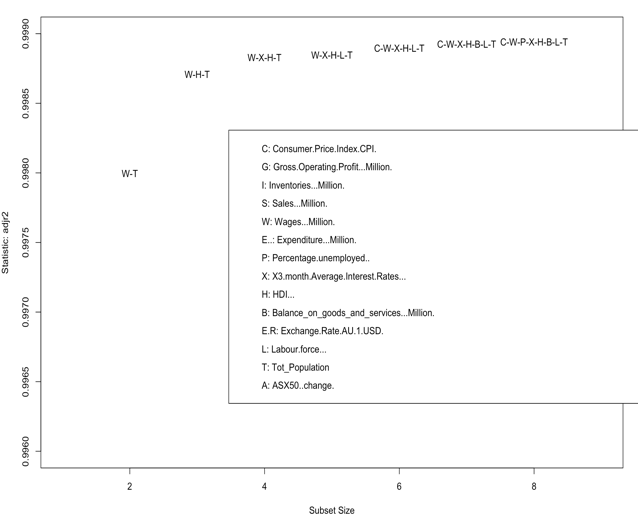
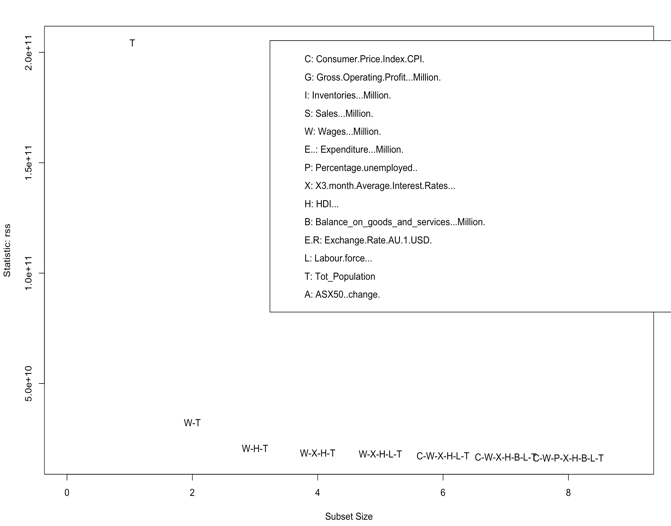
After examining the correlation plot during the exploratory data analysis, it became evident that most of the predictors were correlated with each other leading to multicollinearity. Effects like this could cause coefficient estimates to be erroneous and show significance with target variable even when there is no relationship. Avoiding such a situation, it was decided to drop some of the variables while modelling. To pick the most optimal predictors best subset selection was used. Best subset selection uses various subset of predictors to arrive at the best model based on Mallow’s Cp, Bayesian Information Criterion(BIC), Adjusted-RSquare, Residual sum of squares(RSS).

The best model for each of the criterion is picked and plotted in the figure below. Majority of them have selected model with 8 predictors to be the best model.



As we have the problem of collinearity and our aim was to pick lesser predictors without losing prediction accuracy consequent sub-optimal predictor combinations are examined. The below plots show the best predictor combination in each category. The model with four predictors is chosen as the measures are negligibly different from the best model while reducing the predictor count.



Each alphabet of the 4-variable model denotes:

**W – Wages**

**X – Interest rates**

**H- HDI**

**T- Total population**

There is no prediction feature for best subset regression. So, the selected predictors were used once again in the linear model. Variance inflation factor(VIF) was measured and the values for the predictors had reduced compared to the previous models. A check of the assumptions of linear model for the data was done.

**Assumptions of linear model:**

**Linearity –** There exists a linear relationship between X and Y.

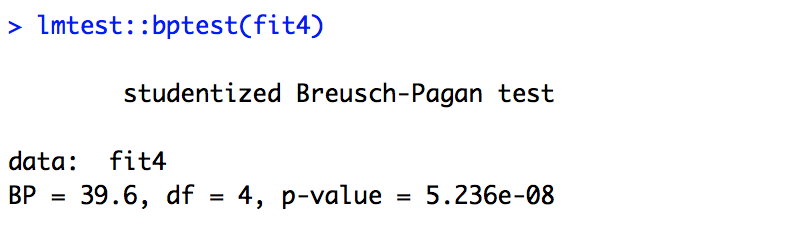
**Homoscedasticity-** The variance of the residuals is same for all values of X.

**Independence –**Residuals are independent to each other.

**Normality –** The values of X and Y are normally distributed.

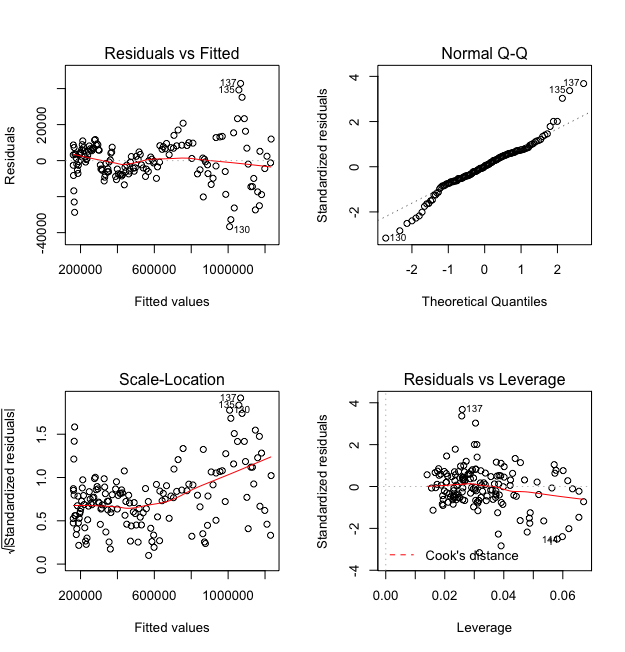
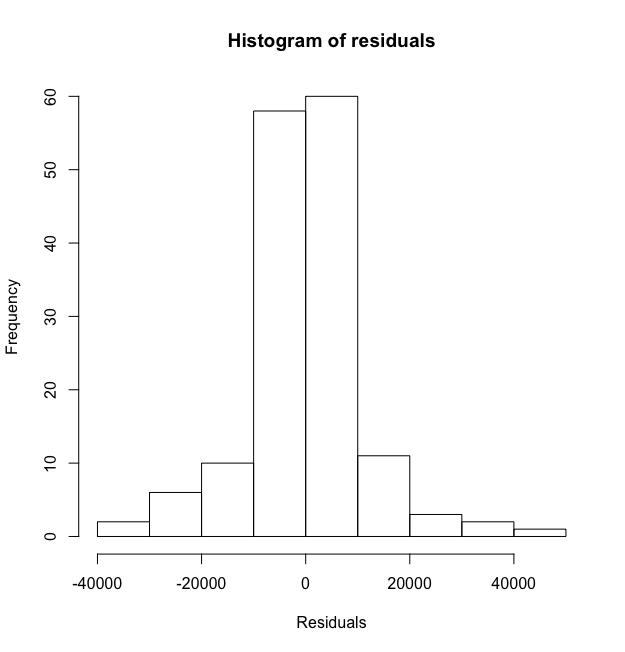
The left-side plot on the first row is a check for linear relationship between X and Y and homoscedasticity. The red straight line show there exists a linear relationship. But the residuals are not evenly spread, and the variance of the residuals get broader towards the end. This give intuition that there is heteroscedasticity in data and one of the model assumption in violated.

To verify the intuition **Breush Pagan Test is run on the fitted model. A value as low as 5.236e-08 was observed which is very less than the significance level of 0.05. So we reject the null hypothesis that the variance of the residuals are constant and infer heteroscedasticity is present in data.**



The right-side plot on the first row is a check for normal distribution of residuals. Residual of each observation is plotted along a 45degree QQ-line. Most of the residuals are along the line except for a few in the tails. To visualise better, a histogram is plotted with the residuals. The bell-shaped curve of the histogram shows the residuals are normally distributed satisfying the assumption. We can see both the plots have identified observations 135 and 137 as outliers.

The first plot in the bottom row shows scaled linear relationship between residuals and fitted values and the second plot has picked observation 137 to be influential leverage point that affect the slope.

Heteroscedasticity in the data can be rectified using Box-Cox transformation. But since each observation in the data is about the economic indicators of Australia they depend on the previous observations and do not satisfy the Independence assumption.

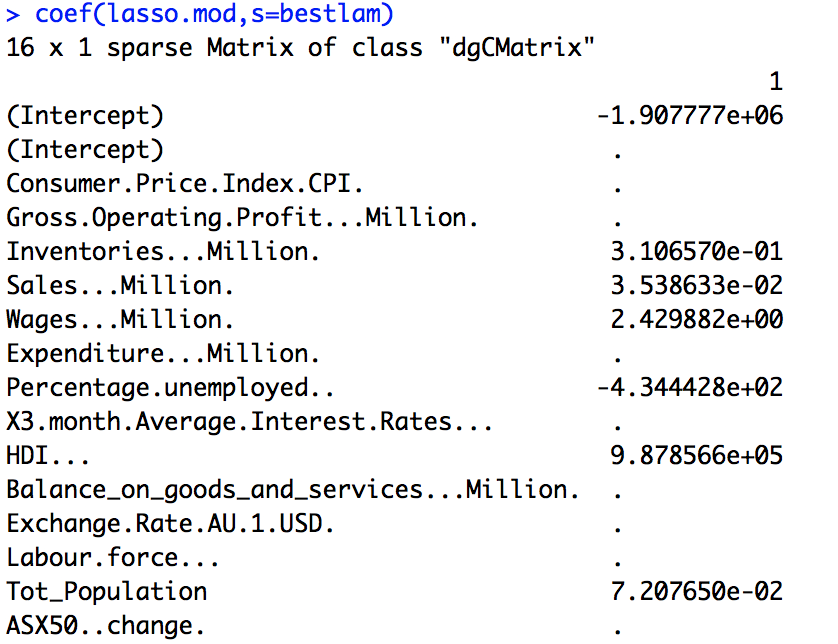
Having said that we try to run regularization models on the dataset.

**Regularization techniques:**

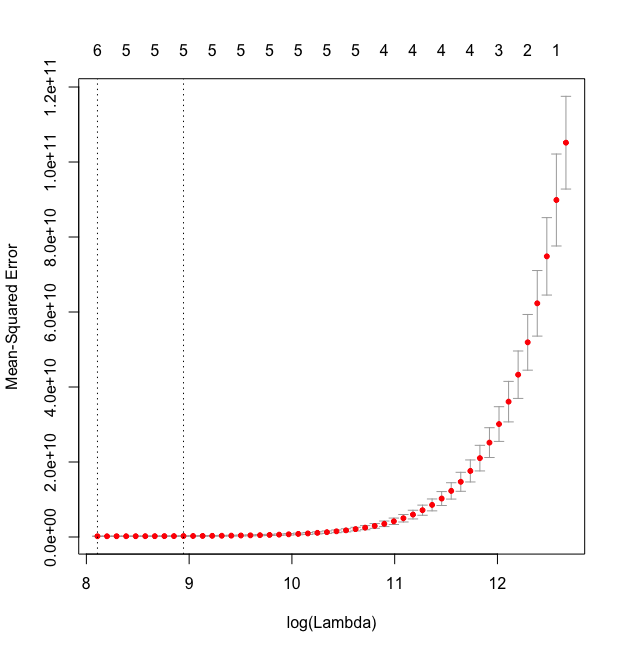
**Lasso regression:**

Lasso regression is a regularization technique that shrinks the coefficient estimates of some of the predictors to zero thereby reducing the number of dimensions in the model. They handle multicollinearity well and do not make any prior assumptions about the residuals in the data. Lasso considers the joint distribution of Y on X and does not assume Y to be generated from a linear model. So, it considered to be a robust model for this setting.

Lasso regression was run on the data and the best lambda (regularization parameter) chose the following model.



The model gave more weightage to Human Development Index(HDI), Total population and percentage unemployed while aggressively shrinking and eliminating other predictors. The plot below shows how lasso has reduced the predictor count as lambda increases. The two vertical lines shows the best lambda and lambda value with 1 standard error.



A prediction was done with the best lambda model and a **Mean Absolute Error (MAE)** of **9991.27** was estimated on the test set. Considering the values in millions scale for GDP the error value seemed very low. Lasso did a good job on the dataset.

**Ridge regression:**

Ridge regression is another type of regularization technique which uses L2 Norm to shrink the coefficients of the predictors near to zero but does not eliminate any predictor in the process.

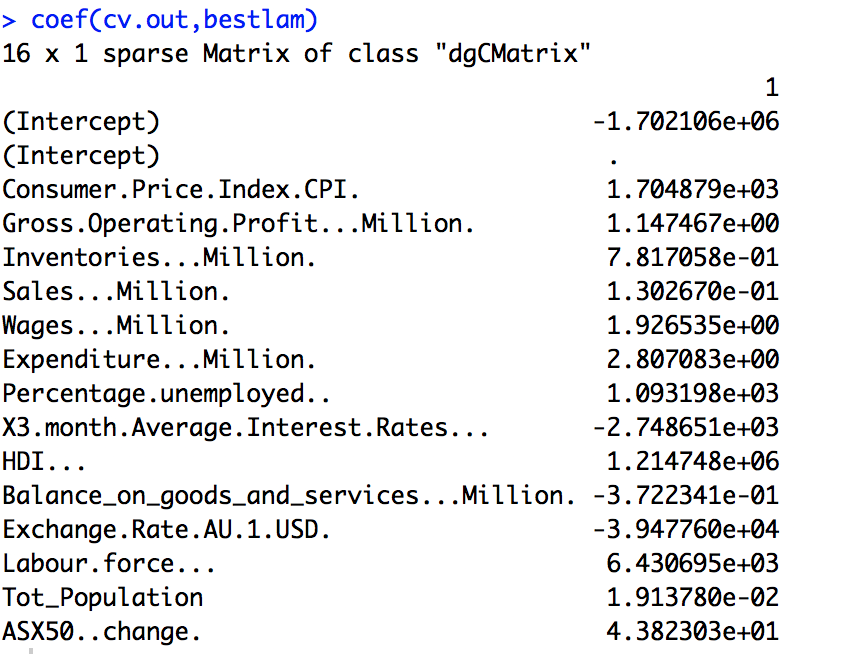
**Assumptions of ridge regression:**

**Linearity –** There exists a linear relationship between X and Y.

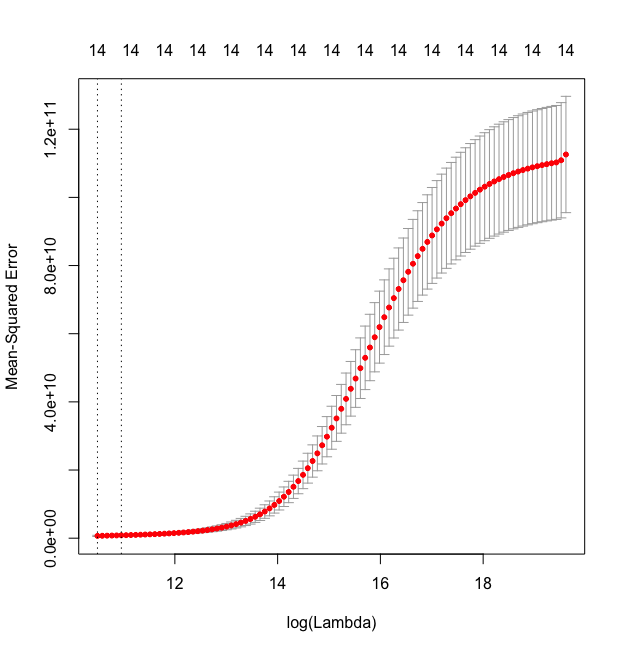
**Homoscedasticity-** The variance of the residuals is same for all values of X.

**Independence –**Residuals are independent to each other.

Ridge regression does not make assumption that the residuals are normally distributed. Various research shows that ridge is good at handling collinearity in data similar to Lasso (CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION). Subsequently data was modelled with ridge regression.



The plot below shows how ridge works. The X axis of the plot shows that the number of predictors has not reduced after training the model with the data.



The trained model was used to predict on the test dataset and MAE value calculated was **18240.82**. The error estimate from the ridge regression is a bit higher than measure for lasso regression.

**Feature extraction models:**

**Principal Component regression (PCR):**

PCR is a feature extraction model and helps in reducing the feature space to principal components that are orthogonal to each other. PCR usually performs better in collinear setting of predictors and hence selected.

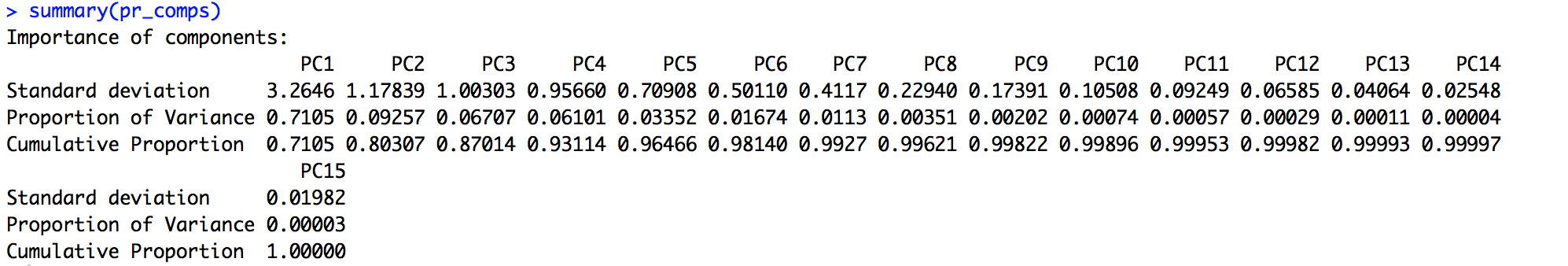
**Assumptions of PCR:**

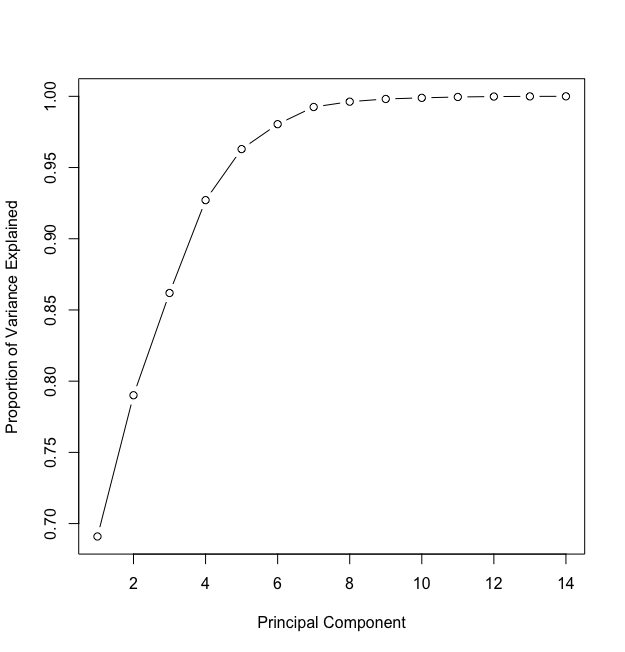
**Linearity –** There exists a linear relationship between X and Y.

**Homoscedasticity-** The variance of the residuals is same for all values of X.

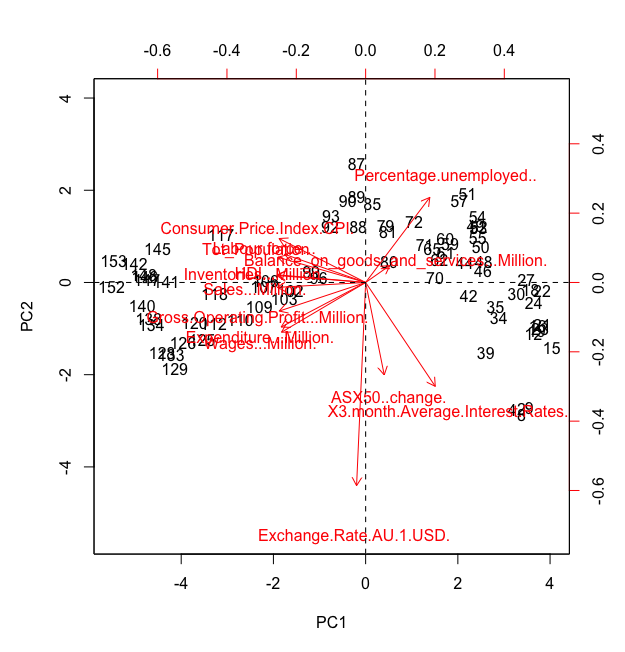
**Independence –**Residuals are independent to each other.

The predictors for principal components are picked based on the proportion of variance explained. Usually the first few principal components explain most of the variance in data.





In our dataset the first 6 principal components explained about 98% of variance in data. Each principal component takes data from different dimension and hence are not affected by multicollinearity.



The predictors taken into consideration for the first two principal components are shown in the bi-plot above. Predictors Inventories and Sales explain most of the variance for PC1. Percentage unemployed and Exchange rates explain the most of variance in PC2.

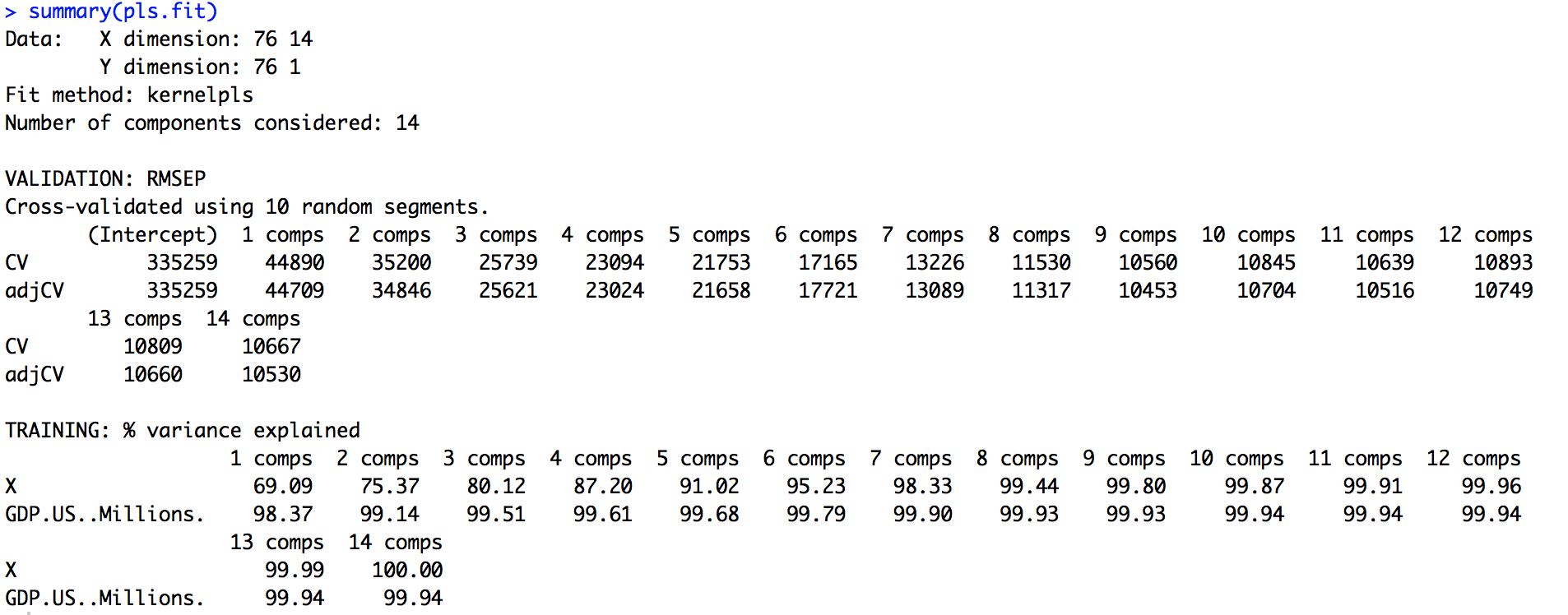
A prediction was done with the first 6 principal components and a MAE value of **62064.09**

was calculated which is very much higher than regularization techniques.

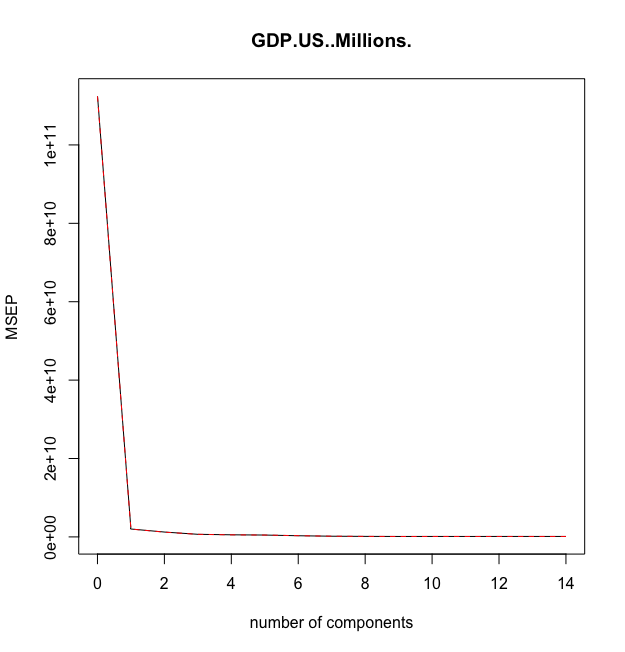
**Partial least squares (PLS):**

PLS is another type of feature extraction technique. The main difference between PCR and PLS is PLS considers the dependent variable and hence would be able to arrive at a higher proportion of variance explained within a few components. The assumptions of PCR apply to PLS as well.

PLS was run on the GDP dataset and the following components were returned. As stated 99.5% of variance is explained within the first 3 components. The cross validated error estimates slowly reduce as more components are added.



The plot shows reduction in the Mean Squared Error(MSE) as PLS components increase. A very big drop in the MSE is observed when the first component is added. The first 4 components were chosen and used for modelling.



MAE estimate for PLS is calculated to be **21418.03.** It is lesser than the value calculated by PCR but still high compared to regularization techniques.

Reference:

*CONFRONTING MULTICOLLINEARITY IN ECOLOGICAL MULTIPLE REGRESSION - Graham - 2003 - Ecology - Wiley Online Library* 2003,viewed 28 May 2018, <<https://esajournals.onlinelibrary.wiley.com/doi/full/10.1890/02-3114>>.