Idividual Assignment

FEM11149 - Introduction to Data Science

Alexandros Stavropoulos - 643414

2022-10-26

Introduction

In this project i was challenged to investigate what is the relationship between the variables used to measure the Sustainable Development Goals (SDGs) and the World Happiness Indicator and finally to predict the World Happiness Indicator, across different countries around the world for a specific year. In order to make my analysis and predictions i used a sample of data from The World Bank data repository,which contains relevant indicators drawn from the World Development Indicators, reorganized according to the goals and targets of the Sustainable Development Goals (SDGs) ("The Global Movement for Our Children's Future-World Top 20 Project" 2022). In combination with The World Happiness Report (WHP) witch collects survey data from over 150 countries in order to see how people evaluate their own live, the World Happiness Index is measured. The most important task of this assignment was to create a model that could predict given some data for a specific country the World Happiness Index .("World Happiness Report," n.d.)

Data description

For this assignment two data sets were used. Both data sets are based on The World Happiness Report that investigates the World Happiness Index through 150 countries. . The first data set contains a sub sample of variables for the year of 2015 and it will be used for an initial analysis and the second data set contains different variables for the same amount of countries for the year 2015. After cleaning these two datasets, my final datasets consist of 112 observations and 23 variables the first own and the second one of 50 observations and 86 variables. Each row of my final data sets refers to the score obtained for World Happiness Index nd a certain indicator by a specific country in 2015.

Methods

In this Analysis two predictive analysis methods where used Lasso Regression and Principal Component Regression (PCR). Also a technique for analyzing large data sets which contain a high number of dimensions called Principal Component Analysis (PCA) was used. Lastly two statistical hypothesis tests were included in the analysis, one being the Permutation test and the other is Bootstrapping.

A regression analysis method based on principal component analysis (PCA) is principal component regression (PCR). In a standard linear regression model, PCR is especially employed to estimate the unknown regression coefficients.Principal component analysis (PCA) is a common method for analyzing huge data sets with a high number of dimensions or features pre observation. Formally, PCA is a statistic method for lowering a dataset's dimensionality. To do this, the data are transformed linearly into a new coordinate system, where (most) of the variance in the data can be expressed with fewer dimensions than the initial data.(Jian Yang 2004) The Permutation test is useful in this situation since it helps determine how many PCs should be present. The permutation test aids in identifying PCs that contribute to some systematic variance in the data and those that are merely a source of noise and shouldn't be considered. Bootstrapping is a different statistical test. The usefulness of this test allowed me to confirm the precision of the PCs that were selected.(Jian Yang 2004)

Lasso Regression was presented as a strategy that, unlike ordinary least squares (OLS), exchanges bias for variance in order to improve the mean square error. To achieve that it uses penalization to the estimators. The penalisation in lasso regression shrinks the estimators exactly to 0. The amount of penalisation is controlled by the parameter (λ) than can be chosen by cross-validation In order to obtain the Lasso estimates we have to minimise the $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}\beta_j x_{ij})^2 + \sum_{j=1}^{p}|\beta_j| = RSS + \lambda\sum_{j=1}^{p}|\beta_j|$

The root-mean-square deviation (RMSD) or root-mean-square error (RMSE). is the metric i used to compare models.It is computed as $RMSD(\tilde{\theta}) = \sqrt{MSE(\tilde{\theta})} = \sqrt{E(((\tilde{\theta}) - \theta)^2)}$ where $(\tilde{\theta})$ is an estimator with respect to an estimated parameter and is it defined by mean square error given by $MSE = \sum((y_i - y)^2)/n$ where y_i is the observed value, y is the corresponding predicted values and n is the number of observations. (Gareth James 2021)

Results

Starting of with the analysis, PCA is used to determine how many components to be included in a reduced model.Upon Kaiser's rule, only the first three component should be retained since only theese component are greater than one (Figure 1). In order to make sure that the right amount of Principal components were selected a Permutation test was used. Similar to the Kaiser's rule, the Permutation Test indicated that three Principal Components should be retained in order for the total variance of the data to be explained more that 80%. (Figure 1)

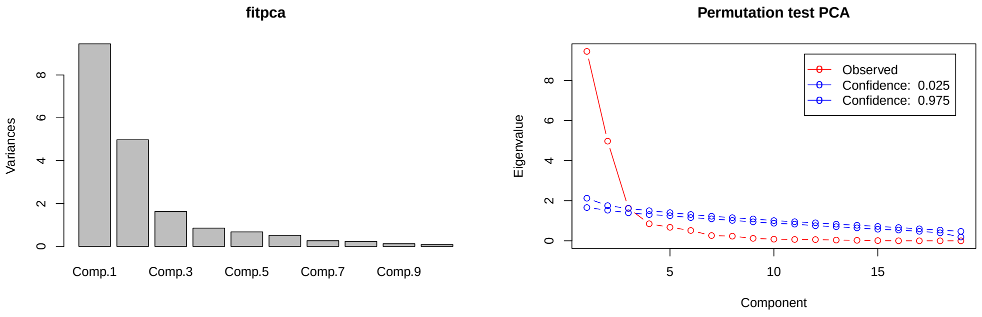

Figure 1: Principal Components Fit and Permutation Test plot

Figure 2 indicates that almost every variable is somehow explained by the first Principal Component. The 1st PC could be explained as general factors that formulate the economic social growth of a country and has an effect on the happiness of its people. The 2nd PC could be interpreted as strict economic factors that affect the happiness of people as it explains better variables related to Gross. Domestic Product and manufacturing. The 3rd PC takes a more environmental approach as it explains more variables related to clean fuels and carbon dioxide emissions.

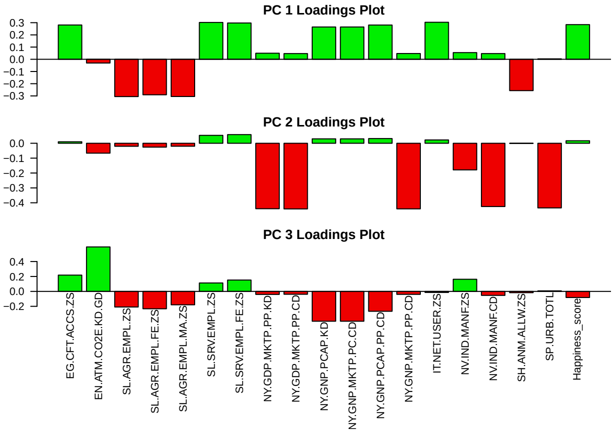

Figure 2: Plot of the three first Principal Components

The 1st PC explains the 49.7% of the total variance, with the 2nd PC , the 75.9% oft he cumulative variance can be explained and with the addition of the 3rd PC the 84.4% of total variance can be explained . As the Bootstrap was executed, the null Hypothesis that the variance explained by the selected components explains at least 70% of the total variance. As we can see from the Figure 3 we can't reject our null Hypothesis.

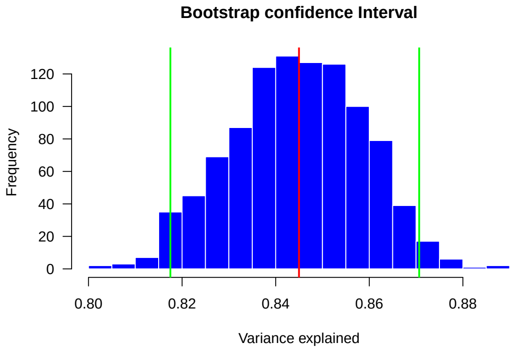

Figure 3: Variance explained from the 1st Principal Component

In order to construct a prediction model we spitted our first sample into train and test in a way to have 15 observations in test data. A Principal Components Regression was called on train data and a Lasso Regression was called on a merge dataset between our two initial data sets. According to our findings the Lasso Regression model performed slightly better as it had smaller RMSE but not significantly better than PCR models.

To continue , a new predictions data set was used that contains 5 observations by different county for year 2019 in order to predict the values of Happiness Index. According to Table 1 the predictive values of both models were pretty close but we can see how its model treats differently every country based on their economic strength. For example for strong economies like The Netherlands and the Switzerland the

Table 1: Predicted values for both models

Country_Name	Year	Happines_Lassomodel	Happines_PCR_model
Brazil	2019	6.02	6.05
Netherlands	2019	6.68	7.08
Switzerland	2019	7.26	7.5
Ukraine	2019	5.32	5.54
Zimbabwe	2019	4.88	4.26

PCR models predicts higher values than the ones of Lasso and on the other hand on poor economies like Zimbabwe PCR model predicts lower values than Lasso model. We can say that the Lasso model takes a more moderate approach when predicting than when PCR does.

Conclusions and Discussion

Taking everything into consideration, the SDGs variables are related to the World Hapiness Index and are suitable for explaining it . Both the PCR model and the Lasso model are appropriate for the prediction part with Lasso in this particular data set having a slight edge in performance compare to PCR model but it is not significantly better. As a takeaway we can say that both models can give us accurate predictions and depending on the data set one model can have better performance over the other

Appendix

Code

```
#DATA PREPARATION
##Read datasets
happinnes <- read.csv(
"C:/Users/alexa/OneDrive/E   /Idividual DS/643414as_Happiness_2015.csv")
happinnes_expanded <- read.csv(
"C:/Users/alexa/OneDrive/E   /Idividual DS/643414as_Happiness_2015_expanded.csv")
#Omit X variable
happinnes<- happinnes[-1]
happinnes_expanded<- happinnes_expanded[-1]
#Omit NAs
sum(is.na(happinnes))
happinnes<- na.omit(happinnes)
happinnes_expanded<- na.omit(happinnes_expanded)
library(caret)
library(caretEnsemble)
library(psych)
# Fit Principal Component Analysis to determine how many PC we will use
fitpca <- princomp(happinnes[,5:23], cor = TRUE, scores = T)
screeplot(fitpca)
# Perform  a PCA Permutation test
 permtestPCA <- function (X, nTests = 100, alpha = 0.05, center.data = TRUE,
                         scale.data = TRUE, ...){
  n <- nrow(X)
  m <- ncol(X)
  X <- scale(X, center = center.data, scale = scale.data)
  if (scale.data) {a <- 1/(n - 1)} else {a <- 1}

  res.X      <- prcomp(X)
  eigs.X     <- res.X$sdev^2
  eigs.Xperm <- matrix(0, m, nTests)
  Xperm      <- matrix(0, n, m)
  Xperm[, 1] <- X[, 1];
  for (i in 1:nTests){
    for (j in 2:m) {
      ind <- sort(runif(n), index.return = TRUE)$ix
      Xperm[, j] <- X[ind, j]
    }
    res.Xperm  <- prcomp(Xperm)
    eigs.Xperm[, i] <- res.Xperm$sdev^2
  }

  perc.alpha <- matrix(0, m, 2)
  for (s in 1:m){
    perc.alpha[s,] <- quantile(eigs.Xperm[s,], c(alpha/2, 1 - alpha/2) )
  }
  plot(1:m, eigs.X, type = "b", col = "red", main = "Permutation test PCA",
       xlab = "Component", ylab = "Eigenvalue", ...)
  lines(1:m, perc.alpha[, 1], type = "b", col="blue")
  lines(1:m, perc.alpha[, 2], type = "b", col="blue")

  string1 <- paste("Confidence: ",formatC(alpha/2, digits=3, width=5,
                                          format="f"))
  string2 <- paste("Confidence: ",formatC(1-alpha/2, digits=3, width=5,
                                          format="f"))
  legend("topright", inset=.05, c("Observed", string1, string2),
         lty = c(1, 1, 1), col = c("red", "blue", "blue"),
         pch = c("o", "o", "o"))
  return(perc.alpha)
}
perm_range <- permtestPCA(happinnes[,5:23])
#Plot loadings of the three first PC
# Change colour of bar plot
c.pc1 <- ifelse(fitpca$loadings[,1] > 0, yes="green2", no="red2")
c.pc2 <- ifelse(fitpca$loadings[,2] > 0, "green2", "red2")
```

```
c.pc3 <- ifelse(fitpca$loadings[,3] > 0, "green2", "red2")
# Get position for variable names
n.pc1 <- ifelse(fitpca$loadings[,1] > 0, -0.01, fitpca$loadings[,1]-0.01)
n.pc2 <- ifelse(fitpca$loadings[,2] > 0, -0.01, fitpca$loadings[,2]-0.01)
n.pc3 <- ifelse(fitpca$loadings[,3] > 0, -0.01, fitpca$loadings[,3]-0.01)
# Plot
layout(matrix(1:3, ncol=1)) # Set up layout
par(mar=c(1,3,2,1), oma=c(7.5,0,0,0)) # Set up margins
# Plot PC 1
b1 <- barplot(fitpca$loadings[,1], main="PC 1 Loadings Plot", col=c.pc1, las=2,
              axisnames=FALSE)
abline(h=0)
# Plot PC 2
b2 <- barplot(fitpca$loadings[,2], main="PC 2 Loadings Plot", col=c.pc2, las=2,
              axisnames=FALSE)
b3 <- barplot(fitpca$loadings[,3], main="PC 3 Loadings Plot", col=c.pc3, las=2,
              axisnames=FALSE)
abline(h=0)
# Add variable names
text(x=b2, y=ifelse(fitpca$loadings[,3] > 0, -0.01,fitpca$loadings[,3]-0.01),
     labels=names(fitpca$loadings[,3]), adj=1, srt=90, xpd=NA)


#Call for bootstrap test
library("boot")
my_boot_pca <- function(x, ind){
  res <- princomp(x[ind, ], cor = TRUE)
  return(res$sdev^2)
}
fit.boot  <- boot(data = happinnes[,5:23], statistic = my_boot_pca, R = 1000)
eigs.boot <- fit.boot$t
head(eigs.boot)
fitpca$sdev^2
par(mfrow =c(1,1), mar = c(5, 4, 4, 1) + 0.1)
#total variance explained by 3 PC
var.expl <- rowSums(eigs.boot[,1:3])/rowSums(eigs.boot)

#Make a histogram and add confidence intervals
hist(var.expl,xlab = "Variance explained", las = 1, col = "blue",
     main = "Bootstrap confidence Interval", breaks = 20, border = "white")
perc.alpha <- quantile(var.expl, c(0.025, 1 - 0.025) )
abline(v = perc.alpha, col = "green", lwd = 2)
abline(v = sum(fitpca$sdev[1:3]^2)/sum(fitpca$sdev^2), col = "red", lwd = 2)
```

```r
# Split our train and set data
library(caret)
library(caretEnsemble)
chooseCRANmirror(graphics = FALSE, ind = 10)
if (!require("pacman")) install.packages("pacman")
pacman::p_load(pls)
set.seed(100)
sample <- sample.int(n = nrow(happinnes), size = floor(0.86607142858 *nrow(happinnes)), replace = F)
trainData <- happinnes[sample, ]
testData <- happinnes[-sample, ]
x = trainData[,c(5:23)]
y = trainData$Happiness_score


# Call Pcr regression on train data
set.seed(100)
pcr_model <- pcr(data = trainData[,c(5:23)],Happiness_score ~ EG.CFT.ACCS.ZS +
                   EN.ATM.CO2E.KD.GD + SL.AGR.EMPL.ZS + SL.AGR.EMPL.FE.ZS +
                   SL.AGR.EMPL.MA.ZS +
                   SL.SRV.EMPL.ZS + SL.SRV.EMPL.FE.ZS + NY.GDP.MKTP.PP.KD +
                   NY.GDP.MKTP.PP.CD + NY.GNP.PCAP.KD +
                   NY.GNP.MKTP.PC.CD + NY.GNP.PCAP.PP.CD + NY.GNP.MKTP.PP.CD +
                   IT.NET.USER.ZS + NV.IND.MANF.ZS +
                   NV.IND.MANF.CD + SH.ANM.ALLW.ZS + SP.URB.TOTL,
                 validation = "CV",
                 scale = TRUE)

#merge happiness and happinnes expanded data set
data<- merge(happinnes,happinnes_expanded, by = "Happiness_score" )
#Call for lasso regression on the merged data set
y <- as.vector(data$Happiness_score)
x <- model.matrix(Happiness_score ~ EG.CFT.ACCS.ZS +
                    EN.ATM.CO2E.KD.GD + SL.AGR.EMPL.ZS + SL.AGR.EMPL.FE.ZS +
                    SL.AGR.EMPL.MA.ZS +
                    SL.SRV.EMPL.ZS + SL.SRV.EMPL.FE.ZS + NY.GDP.MKTP.PP.KD +
                    NY.GDP.MKTP.PP.CD + NY.GNP.PCAP.KD +
                    NY.GNP.MKTP.PC.CD + NY.GNP.PCAP.PP.CD + NY.GNP.MKTP.PP.CD +
                    IT.NET.USER.ZS + NV.IND.MANF.ZS +
                    NV.IND.MANF.CD + SH.ANM.ALLW.ZS + SP.URB.TOTL,
                  data = data)
x <- x[,-1]
set.seed(100)
```

```
fitcontrol <- trainControl(method = "repeatedcv", number = 10, repeats = 5, verboseIter = TRUE)
lasso <- train(Happiness_score ~ EG.CFT.ACCS.ZS +
                 EN.ATM.CO2E.KD.GD + SL.AGR.EMPL.ZS + SL.AGR.EMPL.FE.ZS +
                 SL.AGR.EMPL.MA.ZS +
                 SL.SRV.EMPL.ZS + SL.SRV.EMPL.FE.ZS + NY.GDP.MKTP.PP.KD +
                 NY.GDP.MKTP.PP.CD + NY.GNP.PCAP.KD +
                 NY.GNP.MKTP.PC.CD + NY.GNP.PCAP.PP.CD + NY.GNP.MKTP.PP.CD +
                 IT.NET.USER.ZS + NV.IND.MANF.ZS +
                 NV.IND.MANF.CD + SH.ANM.ALLW.ZS + SP.URB.TOTL,
                data,
               method = 'glmnet',
               tuneGrid = expand.grid(alpha = 1,
                                      lambda = seq(0.0001, 1, length =10)),
               trControl = fitcontrol)
#comparing models on test data
#Lasso
predictionlassotest <- predict(lasso, testData)
sqrt(mean((testData$Happiness_score-predictionlassotest)^2))
#PCR
predictPCRtest <- predict(pcr_model, testData, ncomp = 3)
sqrt(mean((testData$Happiness_score-predictPCRtest)^2))
#load new prediction data set
happiness_2019 <- read.csv(
  "C:/Users/alexa/OneDrive/E   /Idividual DS/Happiness_2019.csv")
happiness_2019<- happiness_2019[-1]

predictionlassohappiness <- predict(lasso, happiness_2019)
predictPCRhappiness <- predict(pcr_model, happiness_2019, ncomp = 3)
predictionlassohappiness <- round(predictionlassohappiness, digits = 2)
predictPCRhappiness <- round(predictPCRhappiness, digits = 2)
predicts <- cbind("Country_Name" = happiness_2019$Country.Name  ,
"Year" = happiness_2019$Time ,"Happines_Lassomodel" =
predictionlassohappiness ,"Happines_PCR_model" =
predictPCRhappiness)
sort(predicts)

library(kableExtra)
knitr::kable(predicts,
             caption = "Predicted values for both models") %>%
  kable_styling(font_size = 10,full_width = T)


options(tinytex.verbose = TRUE)
```

Bibliography

Gareth James, Trevor Hastie, Daniela Witten. 2021. *An Introduction to Statistical Learning with Applications in r Second Edition*. Springer.

Jian Yang, Senior Member, David Zhang. 2004. *Two-Dimensional PCA: A New Approach to Appearance-Based Face Representation and Recognition*. Ieee Transactions on pattern analysis; machine intelligence.

"The Global Movement for Our Children's Future- World Top 20 Project." 2022. *Educate Every Child on the Planet: The World Top 20 Project*. https://worldtop20.org/global-movement?gclid=Cj0KCQjwteOaBhDuARIsADBqReid4azyC0H3LZh8x60OsHwgSE7fAE23UUgY0XsGQEXp9JYehJDkdT8aAk_ZEALw_wcB.

"World Happiness Report." n.d. *TWHR | The World Happiness Report*. https://worldhappiness.report/.