Predicting the happiness index using the gender statistics database

Introduction to Data Science - Assignment 3 - Gergely Máriás

1 Introduction

This research involves an analysis of the relationship between various variables from the Gender Statistics database and the World Happiness Index. The primary objective is to identify country characteristics that are associated with the happiness score and offer insights to address gender inequality and enhance overall well-being. The research question under consideration is:

What are the key determinants of a country's happiness score, and how do they relate to gender equality metrics?

2 Data Overview

Gender Statistics Database: It was collected and maintained by the World Bank, which includes gender-related data on demographics, education, health, economic opportunities, public life, and decision-making across 115 countries. **World Happiness Report:** It collects survey data from over 134 countries to construct the World Happiness Index, representing how individuals rate their own lives based on a set of indicators. This dataset only contains the name of each country and its so-called "Happiness score".

3 Methodes

Principal Component Analysis (PCA) simplifies complex data by creating new coordinate systems to capture important patterns and reduce dimensionality, with methods like Kaiser's rule, cumulative variance explained, and the Scree diagram assisting in the determination of the optimal number of components. Additionally, statistical techniques like **permutation tests** are employed to evaluate observed values, typically within a 95% confidence interval, while **bootstrapping** provides an approach for estimating statistical uncertainty, conducting hypothesis testing, and validating models, particularly in cases with non-standard assumptions or small sample sizes. Furthermore, **Principal Component Regression** (PCR) can be employed after PCA to build a predictive model by using the reduced-dimensional components to improve prediction accuracy and mitigate overfitting in complex datasets.

LASSO (Least Absolute Shrinkage and Selection Operator) regression is a linear regression technique that incorporates a penalty term on the absolute values of the regression coefficients, encouraging a sparse model selection by forcing some coefficients to be exactly zero. This feature makes it effective for variable selection and preventing overfitting. **LOOCV** (Leave-One-Out Cross-Validation) is a validation method, particularly useful when the sample size is small, as it assesses model performance by iteratively training on all but one data point and testing on the omitted point, offering a robust evaluation of LASSO regression for predictive accuracy and variable selection. After LOOCV Lambda min and lambda 1se are values that help determine the level of regularization in LASSO. Lambda min is the smallest lambda that simplifies the model by setting some coefficients to zero. Lambda 1 se strikes a balance between model complexity and predictive power, offering a more interpretable model while maintaining reasonable performance.

4 Results

I analyzed the Gender Statistics Database, initially comprising over 1000 features across 115 countries. However, after merging the Happiness score dataset as the dependent variable, my analysis can continue only on 56 countries. To prepare for LASSO and potential PCR predictions, feature reduction was essential. This process involved removing unsuitable binary variables, addressing missing data (including two countries with substantial missing data), and eliminating redundant features with "total" values. Furthermore, I retained a maximum of 7 missing values per variable, replacing them with respective mean values. This led to an analysis involving 54 countries and 54 features, including Country name, Happiness score, and Country Code.

Given the small sample size, Leave-One-Out Cross-Validation was employed to determine the optimal lambda for the model. Figure 1 illustrates the MSE values for various lambda values, with dashed lines indicating lambda min and lambda 1 se. It is important to note that selecting the model with the lowest RMSE and highest R-squared value can be challenging due to the limited sample size.

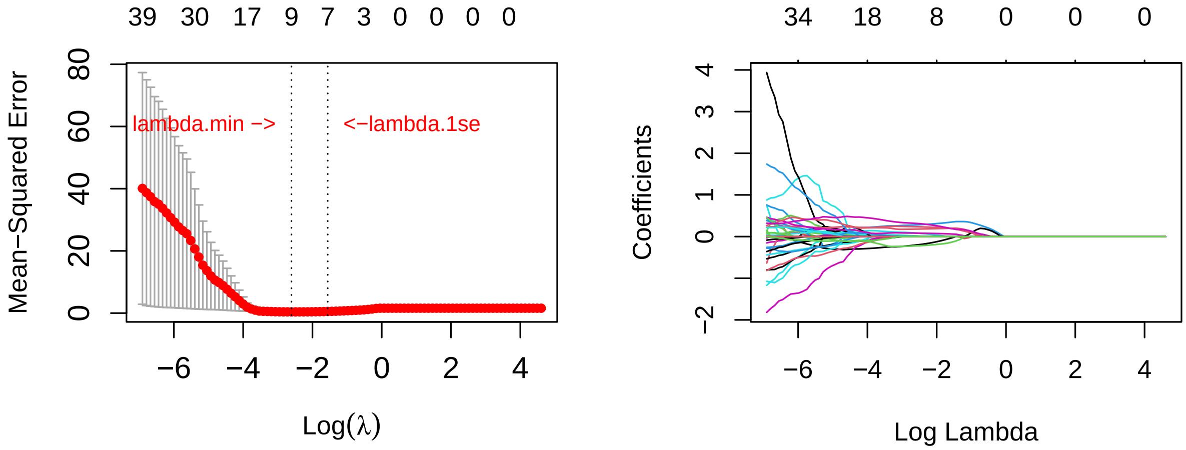

Figure 1: Tuning LASSO hyperparameter with Cross-Validation and visualizing regularization paths

When I looked at the results in Table 1 on the left side, it was pretty clear that the lambda values around 1 standard error (lambda 1 SE) led to lower Root Mean Square Error (RMSE) and higher R-squared values. So, I decided to go with the lambda that gave the smallest prediction error. After that, I delved into a detailed analysis of the model results. As I ran the LASSO regression, I could see which coefficients were still in play, and you can find those listed in Table 1 on the right side.It is important to note that these results are associated with higher Happiness scores first, and they indicate a stronger presence of gender equality second.

NY.GNP.PCAP.PP.CD: GNI per Capita (PPP): A positive coefficient shows that higher income and economic development are linked to greater gender equality, indicating that increased resources and opportunities in high-GNI countries promote gender empowerment. **SG.GEN.PARL.ZS:** Proportion of Seats Held by Women in National Parliaments: A positive coefficient suggests that countries with more women in national parliaments tend to have improved gender equality, highlighting the role of political representation in advancing gender equality goals. **SG.LAW.INDX.AS:** Women's Legal Rights and Access to Assets (Assets Indicator Score): A positive coefficient indicates that countries with strong legal protections and economic opportunities for women tend to have better gender equality, emphasizing the significance of legal frameworks and economic opportunities in promoting gender equality. **SG.LAW.INDX.PR:** Women's Legal Rights Related to Parenthood (Parenthood Indicator Score): A positive coefficient implies that countries with favorable legal rights related to parenthood for women tend to exhibit better gender equality, indicating that supportive legal rights for working mothers and fathers contribute to gender equality. *Interpretation was written only regarding the coefficients, which are aligned with the context of my research question.*

Table 1: Selected LASSO coefficients and their values

	MSE	RMSE	R.squared
LASSO lamda.min	0.51	0.72	0.544
LASSO lamda.1se	0.36	0.60	0.641

Coefficient name	Coefficient value
SP.POP.DPND	-0.029
NY.GNP.PCAP.PP.CD	0.348
SG.GEN.PARL.ZS	0.191
SG.AGE.RTRE.FL.MA	0.061
SL.UEM.TOTL.MA.ZS	-0.132
SG.LAW.INDX.AS	0.200
SG.LAW.INDX.PR	0.192

Moving to Principal Component Analysis (PCA), it is important to note that PCA is an unsupervised learning method. In PCA, I am interested in the independent variables, often called features, and I do not need a dependent variable. My primary goal here is to simplify the dataset by focusing on the most important components. I aim to have these components explain more than 70% of the variation in the data. To ensure that I am on the right track, I am looking for eigenvalues greater than 1. The result of a permutation test further supports my approach, as it is within the confidence interval, suggesting that this PC's contribution to explaining the variance in the data is not significantly different from what would be expected by random chance. In Table 2, I can see that the fifth Principal Component (PC5) aligns well with my criteria. For precise component values and confidence intervals for both permutation test and bootstrap, I refer to Appendix Table 5.

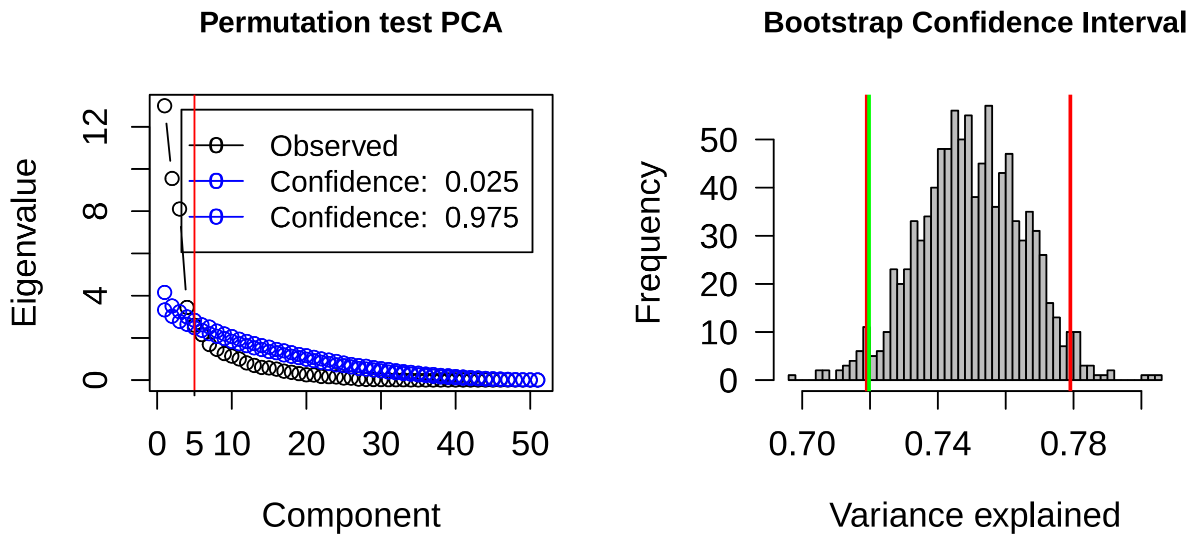

Figure 2: Permutation test and Bootstrap for PCA

The bootstrap histogram (Figure 2, right side), based on the first 5 Principal Components (PCs), provides insights into the explained variance, with an original value of 72% and confidence intervals ranging from 71.9% to 77.9%. Following a normal distribution assumption, the histogram's shape, center, and spread suggest the bootstrapped explained variances align well with the expected distribution. An approximately symmetric, bell-shaped curve is observed with the peak near the original value, indicating consistency. Minimal variability supports the reliability of the initial explained variance estimate for these 5 PCs.

Following the selection of the first 5 Principal Components (PCs), a relabeling process was undertaken based on the top contributors within each component. The top contributors were determined with reference to

Appendix Table 6, which provides insight into the primary contributing variables. Notably, for the first and second components, the focus was narrowed to the top 8 contributors, as they collectively account for more than 44% of the explained variance, as indicated in Table 2. For the third PC, attention was directed toward the top 6 contributors, while the fourth and fifth PCs were renamed by considering only the top 3 contributors. Utilizing the provided dictionary in Excel, the following labels were assigned: 1PC as "Economic Metrics," 2PC as "Demographic Metrics," 3PC as "Labor Market Metrics," 4PC as "Women's Economic Opportunities," and 5PC as "Economic and Labor Market Dynamics."

Table 2: Explained variance, Cumulative variance and Eigenvalues for 5 Principal Components

	PC1	PC2	PC3	PC4	PC5
% of explained variance	0.255	0.187	0.159	0.068	0.051
Cumulative % of explained variance	0.255	0.442	0.601	0.669	0.720
Eigenvalues	13.001	9.552	8.108	3.444	2.600

Next, I applied Principal Component Regression (PCR) analysis to the same training dataset used in the previous LASSO analysis to facilitate a thorough comparison. Subsequently, I utilized the PCR model to make predictions on the same test dataset. I obtained consistent evaluation metrics such as Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and R-squared again.

Table 3 provides a comprehensive overview of the performance indicators of the three models. Considering the detailed analysis performed, especially due to the relatively small sample size, it is reasonable to say that the lambda 1 SE Lasso is the optimal choice for predicting the happiness point in the given three countries, because it has the smallest Mean Squared Error and the highest R-Squared value. Additionally with small sample size lambda 1 SE can bring a more robust result than Lambda min.

Table 3: Model comparison: PCR and LASSO performance metrics

Model	MSE	RMSE	R-squared
PCR	0.60	0.77	0.408
LASSO lambda.min	0.51	0.72	0.544
LASSO lamda.1se	0.36	0.60	0.641

The predictions for the Happiness Scores of the three countries are provided in the Appendix, Table 4.

5 Conclusion

The LASSO and PCA models reveal the vital role of specific variables in predicting Happiness Scores, including Age dependency ratio, GNI per capita, Proportion of seats held by women in national parliaments, Retirement age for males, Male unemployment rate, Women's Legal Rights related to assets, and Women's Legal Rights related to parenthood. These variables align with economic, demographic, labor market, women's economic opportunities, and economic and labor market dynamics in the PCA model, underscoring their impact on a country's happiness index. In summary, these findings provide insights into enhancing overall well-being and addressing gender equality, with LASSO supporting the importance of gender equity in boosting happiness. Thus, governments should prioritize gender equality efforts, recognizing their intrinsic connection to increased happiness. This research offers valuable guidance for policymakers, emphasizing the need for targeted interventions to promote both societal happiness and gender equality simultaneously.

6 Appendix

Table 6: Top contributors by each 5 Principal Components

PC1_Contributors	PC2_Contributors	PC3_Contributors
SG.LAW.INDX	SP.POP.0014.MA.IN	SL.TLF.CACT.MA.ZS
SL.EMP.1524.SP.FE.ZS	SP.POP.0014.FE.IN	SL.EMP.TOTL.SP.MA.ZS
SL.TLF.ACTI.1524.FE.ZS	SP.POP.1564.FE.IN	SL.EMP.1524.SP.MA.ZS
SL.EMP.TOTL.SP.FE.ZS	SP.POP.1564.MA.IN	SP.URB.TOTL.IN.ZS
SL.TLF.CACT.FM.ZS	SL.TLF.TOTL.MA.IN	SP.RUR.TOTL.ZS
SL.TLF.TOTL.FE.ZS	SL.TLF.TOTL.FE.IN	
NY.GNP.PCAP.CD	SP.POP.65UP.MA.IN	
SL.TLF.CACT.FE.ZS	SP.POP.65UP.FE.IN	

PC4_Contributors	PC5_Contributors
SG.LAW.INDX.MR	SL.UEM.TOTL.MA.ZS
SG.LAW.INDX.AS	SL.UEM.1524.FM.ZS
SL.TLF.TOTL.FE.ZS	SL.TLF.CACT.FM.ZS

Table 4: LASSO regression predictions for Happiness Scores

Country Name	Happiness score
Costa Rica	5.343
Croatia	5.880
Peru	5.358

Table 5: Confidence intervals for Principal Component 5

	Permutation test	Bootstrap
Lower Interval	2.480	0.719
Upper Interval	2.833	0.779
PC5	2.600	0.720

Data Manipluation and Cleaning

```r
#Import from Happiness_score.csv
Happiness <- load(Happiness_score)
colnames(Happiness)[1] <-"CountryName"
compare <- merge(Gender, Happiness, by="CountryName", all.y = T, all.x = T)
#I found 4 countires which had different names,
#therefore I renamed and then merged the datasets
Happiness[60,1]<- "Kyrgyz Republic"
Happiness[98,1]<- "Iran, Islamic Rep."
Happiness[118,1]<- "Egypt, Arab Rep."
Happiness[130,1]<- "Congo, Rep."

Gender<- merge(Happiness,Gender, by="CountryName")

#Replace NAs with column's means
Gender <- data.frame(lapply(Gender, function(x) {
  if(is.numeric(x)) {
    ifelse(is.na(x), mean(x, na.rm = TRUE), x)
  } else {
    x
  }
}))
```

Code

```r
##### L A S S O ######
#Splitting to a training and a test set by 80%-20%
set.seed(29)
sample_size = floor(0.8*nrow(LASSO))
picked = sample(seq_len(nrow(LASSO)),size = sample_size)
train_data_lasso = LASSO[picked,]
test_data_lasso = LASSO[-picked,]
#Building optimal LASSO model
#Leave-One-Out cross validation to find the best lambda which means number of folds are 54
lambdas_lasso <- 10^seq(2, -3, length.out = 100)
cv_lasso <- cv.glmnet(train_x_lasso,
                      train_y_lasso,
                      alpha = 1,
                      lambda = lambdas_lasso,
                      nfolds = 54,
                      intercept = T,
                      grouped = F)
#After comapring the lambdas I selected to go with lambda= lambda1se
#I obtained the coefficient's names from the model which are not zero
coefficients<-lasso_l1se$beta
non_zero<- apply(coefficients,1,function(x) x != 0)
coeffs2<- data.frame(subset(non_zero, non_zero != 0))
coeffnames<-rownames(coeffs2)
coeffs<- data.frame(round(subset(coefficients, coefficients[1:51] != 0),3))
coeffs<- cbind(coeffnames,coeffs)
colnames(coeffs)<- c("Coefficient name", "Coefficient value")
##### P C A ######
#After computin pca I did permutation test, and bootstrap
#Permutation test
```

```r
source("C:/Users/gergo/Desktop/permtestPCA.R")
par(cex.main = 1, cex.lab = 1.2, cex.axis = 1.2, cex.legend = 0.5)
permrange <- permtestPCA(PCA[,4:ncol(PCA)])
abline(v = 5, col = "red")
axis(side = 1, at = c(5), labels = c("5"), las = 0.5, tck = -0.015)
#Can I keep the fifth PC?
in_interval<-data.frame(c(interval= permrange[5,], pca_result$sdev[5]^2))
rownames(in_interval)<- c("Lower Interval", "Upper Interval", "PC5")
colnames(in_interval)<- "Value"
kable(round(in_interval, 5),caption = "TABLE")
#YES
#Conduct the bootstrap
my_boot_pca <- function(x, ind){
  res <- princomp(x[ind, ], cor = TRUE)
  return(res$sdev^2)
}
#Run bootstrap
set.seed(29)
fitboot  <- boot(data = PCA[,4:ncol(PCA)], statistic = my_boot_pca, R = 1000)
boot_ev <- fitboot$t
variance_explained <- rowSums(boot_ev[,1:5])/rowSums(boot_ev)
bootstrap <- hist(variance_explained, xlab = "Variance explained", las = 1, col = "grey",
                  main = "Bootstrap Confidence Interval", breaks = 40,
                  cex.axis = 1.2,cex.lab=1.2)
perc.alpha <- quantile(variance_explained, c(0.025, 1 - 0.025) )
abline(v = perc.alpha, col = "red", lwd = 2)
abline(v = sum(pca_result$sdev[1:5]^2)/sum(pca_result$sdev^2), col = "green",lwd=2)
#Splitting to a training and a test set with the same seed for the right comparison
set.seed(29)
sample_size = floor(0.8*nrow(PCA))
#Perform PCR with 5 PC
pcr_model<- pcr(data= train_data, Happiness_score ~ 'all features', validation= "CV",
                scale= TRUE)
pcr_pred <- predict(pcr_model, newdata = test_data, ncomp = 5)
#I bind all models metrics into one dataframe
decision_table<- rbind(comparison_table,comparison_table_lasso)
rownames(decision_table)[1]<- "PCR"
rownames(decision_table)[2]<- "LASSO lambda.min"
rownames(decision_table)[3]<- "LASSO lamda.1se"
#Selected the LASSo lambda.1se model and predict on the given sample
lasso_pred<-predict(lasso_l1se, newx = x_pred_sample)
colnames(lasso_pred)<- "Happiness_score"
Prediction <- cbind(Prediction[, 1:1], lasso_pred, Prediction[, (1 + 1):ncol(Prediction)])
colnames(Prediction)[1]<- "CountryName"
Prediction<-Prediction[,1:2]
kable(Prediction, caption = "TABLE" ,col.names = c("Country Name","Happiness score"))
```

7 References

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani. 2013. An Introduction to Statistical Learning. Vol. 112. Springer.