Happiness, What proportion of the six factors surveyed attribute to a high score?

Sophie Constant

28/11/2021

GitHub repository and Data

Please see the link below for the github repository associated with this assignment. https://github.com/SConstant/C7801-Assignment-1-Data-Science-for-Global-Agriculture-Food-and-Environment (https://github.com/SConstant/C7801-Assignment-1-Data-Science-for-Global-Agriculture-Food-and-Environment).

Please see the link to the original dataset used for this assignment. https://www.kaggle.com/unsdsn/world-happiness (https://www.kaggle.com/unsdsn/world-happiness)

Background

“The World Happiness Report is a landmark survey of the state of global happiness that ranks 156 countries by how happy their citizens perceive themselves to be.” (Helliwell, J., Layard, R., & Sachs, J. 2019). Its genesis emanates from the UN General assembly, where a meeting was held in 2011 chaired by the Secretary General at the time Ban ki Moon and the prime minister of Bhutan. The following year the first World Happiness Report was released as a function of the United Nations Sustainable development network. The purpose of this report is to analyse the progression of nations, and to pull out trends in seeing how happiness, the sense of well-being people globally experience (or not as the case may be) evolves over years. The Happiness Score is derived as the survey measure of subjective well being from the Gallup world poll of 2019 and based on the Cantril ladder scale. The survey itself includes data from respondents across the years 2005 to 2018 and the majority of which is the aggregated response to the following question: “Please imagine a ladder, with steps numbered from 0 at the bottom to 10 at the top. The top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. On which step of the ladder would you say you personally feel you stand at this time?” (Helliwell, J., Layard, R., & Sachs, J. 2019) There are six factors upon which data is collated and ranked and these are outlined below. GDP per capita, is taken as Purchasing power parity. It is notable that some of the data included (from 2017-2018) were forecasts from the OECD Economic outlook, and the World Bank’s global Economic prospects. Healthy life expectancy is generated from data originating from the World health organisation, Global Health observatory data repository. Social support is an aggregate, the national average of binary responses to the Gallup world poll question: “if you were in trouble, do you have relatives or friends you can count on to help you whenever you need them of not? Freedom to make life choices is a similar aggregate of binary responses to the question: “Are you satisfied or dissatisfied with your freedom to choose what you do with your life?” (Helliwell, J., Layard, R., & Sachs, J. 2019) Generosity, is taken as the residual of regressing the national average of the Gallup World Poll responses to the survey question: “Have you donated money to a charity in the past month?” on GDP per capita. Lastly, much like Social support and freedom to make life choices, Perceptions of corruption are and aggregate of average of binary answers to two GWP questions: “Is corruption widespread throughout the government or not?” (Helliwell, J., Layard, R., & Sachs, J. 2019) and “Is corruption widespread within businesses or not?”. Where there’s no government corruption data available, business corruption has been used instead.

The objective of this report is to analyse the relationship between the overall Score as a dependant variable

attributed to happiness and the values given for the 6 aforementioned quality of life factors as independent variables. While it is likely that the factors by nature of inclusion within the World Happiness Report will have some impact to the overall score, this report is concerned with the proportion at which these quality-of-life factors impact the overall score. Moreover which quality of life factors have the most influence or would be the better predictor of a high Happiness score. This particular document is focussed on the report which was published in 2019, and as such the original source of the dataset found on Kaggle.

Methods

The dependent variable was identified as Score (or overall happiness), with the 6 factors surveyed, GDP per capita, Social Support, Healthy life expectancy, freedom to make life choices, generosity and perceptions of corruption. A note on the data, Score is taken from the Gallup World Poll of 2019, and in this analysis has been taken to be an aggregate of the responses received with the assumption of a transformation from a discrete to continuous scale so enabling confidence in the decisions around the chosen methods of analysis. Following an initial Exploratory data analysis (EDA) on the distribution of the six factors, correlation testing was carried out. This to gain an insight and overview into potential relationships between Score and the 6 quality of life factors but also to venture the possibility of correlation relationships between independent variables themselves (see fig. 1). Four out of the six factors were found to be not Gaussian so proceeded with a Spearman's Rank correlation to ascertain any correlation relationship, and the strength of these between the six factors and the dependant variable Score. A smoothed line was chosen over a regression line due to an acceptance that not all the data points in all the variables were Gaussian. Further EDA was then carried out to assess the assumptions required for a multiple linear regression. It was found that the models for Social support, Freedom to make life choices and Perception of corruption violated the need for Gaussian residuals and homoscedasticity. There was an attempt to log transform these to persuade their residuals to conform to the assumptions but the heteroscedasticity appears worse with Social support and perception of corruption. For the purpose of carrying out a Multiple Linear regression the square of social support and polynomials to the order of 3 were added to the data frame as additional variables. The purpose of this was to offer a more expressive multiple linear regression model while still being able remain in this class of analysis. A linear regression model was carried out to assess the expression of the squared variables where squared Social support displayed a low p-value attached indicating more expression. Following this an ANOVA of the comparison of the linear regression models of the squared and un-squared was then used to ascertain expressiveness. The same was conducted for the polynomial order of 3 (cubed) for the variable Freedom to make life choices. Following this there was a process of model selection, in order to ascertain the model with the greatest performance in terms of expression and attribution of variance within Happiness score via the 6 quality of life factors. Following this Model selection was carried out, to ascertain the most effective models in terms of which independent variables, had the most effect on the dependant variable Score. Firstly carrying out best subset selection identifying the model with the largest R squared, so identifying a model with a low training error with cross validation with Mallows's Cp (Cp), Adjusted R2 and Bayesian information criterion (BIC) then being used to identify a model with a low test error. These models were plotted with a red dot identifying the model with the highest value for Adjusted R2 and the models with the lowest value for Cp and BIC to facilitate a visual approach to the analysis. This was followed by Forwards and Backwards stepwise selection. Finally both a validation set approach and Cross validation were used to look at the test error in its own right thereby increasing confidence in the final selection of a model within this analysis.

Results

For the correlation test Score seems very positively correlated with GDP per capita, Social support, and Healthy life expectancy with coefficients of 0.81, 0.82, and 0.81 respectively. Score also appears very correlated with freedom to make life choices with coefficient of 0.55. Score was weakly correlated with perceptions of corruption with a coefficient of 0.22. Generosity was not correlated with Score, and it is interesting to note that Generosity only correlated with Freedom to make life choices. In terms of correlations between independent variables, GDP per capita, Social support and Life expectancy were all highly correlated.

GDP was slightly correlated with perception of corruption with a coefficient of 0.22 with GDP, social support and life expectancy were all very correlated with freedom to make choices but not as much Perceptions of corruption were correlated with Generosity with a coefficient of 0.29 and freedom to make life choices with a coefficient of 0.40. For the analysis of suitability for inclusion of the polynomials, from the linear regression the squared Social support has a low p-value attached, 0.00000000000000022 indicating significance thereby more expression in terms of letting us know what Social support means for the Score of Happiness. The ANOVA supported this with the squared model holding a significance of 0.00000342 leading to the acceptance of the hypothesis, the model containing both the squared and un-squared is more expressive, than un-squared alone. The observations from the Multiple Linear regression as follows. The un-transformed model for freedom to make life choices yields more significance than the cubed model. The heteroscedasticity in the residuals has the potential for 'pushing' greater significance than present due to a least squares regression assuming a constant variance, leading this to be less convincing. It cannot be ignored that social support squared experienced a high variance inflation factor. Including a quadratic approach that allows for the curves in the data found in social support, freedom to make Life Choices, with the quadratic and polynomials of the aforementioned variables added as new variables. Still however adopting a linear approach to multiple regression. The Multiple R squared is 0.80, so about 80% of the variation in the score can be explained by this model, so the 6 factors, presumably largely the significant factors. Looking at the F statistic 71 (greater than 1) in combination with an very small p-value 2.2e-16 for the overall model, and there is indication of a highly significant model where the slope for Score. The other 6 factors are approaching zero The following convey significance at an alpha of 0.001. The p-value for the model for GDP per capita is 0.0037 indicating significance, The p-value for the model for Healthy life expectancy is 0.0015 indicating significance, he p-value for the model for squared social support is 0.0047 indicating significance. The p-value for the model for perception of corruption is 0.012, conveying significance at an alpha of 0.05. Looking at the intercept, assuming all the factors were 0, the score would be (estimated) as 2.9 It's interesting to see how the Squared social support (transformed due to the heteroscedasticity seen in the un-transformed variable) appears significant at only an alpha of 0.001 when accounting for the curve in its slope. We can see that for GDP per capita, a change of 0.67 (0.67%) would affect the Score. For Healthy Life expectancy we can see that a change of 1.1 (1%) would affect the Score and for Social support, a change of 1.08 (1%) would affect the score of Happiness. When plotting the residuals, these seemed reasonably even, with outliers being consistent at 152, 148 and 133 across the residuals v fitted. When looking at the Variance inflation factors (VIF) Social support has a VIF of 28 and Social support squared has a VIF of 30, and this seems very high indicating a high amount of multicollinearity. From observations gleaned from the multiple linear regression there is confidence in accepting the Hypothesis there is a relationship between the dependant variable (y) Score, and the independent variables (x) GDP per capita, Social support when squared, and Healthy life expectancy. There is a slightly significant relationship between the dependant variable y and perceptions of corruption.

In best subset selection Social support squared features in all the models, with Healthy life expectancy appearing next, and GDP per capita appearing third. The coefficients were examined for models 4, 5 and 6. The independent variables across model four expressed coefficients which were marginally less sensitive than models 5 and 6, however these two were also very similar is sensitivity leading to some doubt. From the plotted models to cross validation statistics (Fig. 4.) it can be seen that model 4 is possibly performing similarly well to model 5. Generosity has very little effect from the perspective of Cp adjusted R squared and BIC. Freedom to make life choices to the order of 2 has very little effect from the perspective of adjusted R squared, Cp, and BIC. Freedom to make life choices to the order of 1 also has little effect from the perspective of Cp. Model 5, has a BIC of 210, which is the highest, but does cross over with the adjusted R squared in terms of including the variables GDP per capita, Healthy life expectancy and Freedom to make life choices 1. The Cp is 8.7, which while isn't the lowest is a lot lower than other values along the same y axis 18, 45 and 98 and includes the variables GDP per capita, Healthy life expectancy, Social support squared and freedom to make life choices. Model 6, in comparison, displays a lower BIC at 200 and includes generosity, whereas the Cp is higher at 10 and includes all the variables. The Adjusted R2 remains 78 losing social support, generosity, Perceptions of corruption and Freedom of life choices squared and cubed. A this point Model 5 was looking promising, with model 4 also showing a promise which could not be ignored. In the forwards stepwise selection the best one variable model starts with Social support squared The best two variable model has Social support

squared, the three variable model has the above with Freedom to make life choices additionally. The four variable model then has GDP per capita additionally (interestingly one of the more significant in the Multiple linear regression) with the five variable model then included social support which was not originally significant. It's interesting that Freedom to make life choices only shows up in the 3 variable model, as it shows up as fairly significant in the Multiple linear regression, though I'm inclined to attribute this to the heteroscedasticity the residuals exhibited inflating this. As such this variable may need to approach this variable with a non-linear regression. All the variables, showed that the coefficients were the same across forwards, and backwards selection with the only differences being seen in subset selection. Finally, the validation set approach and Cross validation techniques yielded a set of minimum error values which when explored in R, led to a model with 4 variables being the one with the least error. Upon taking a visual approach and plotting this (Fig. 5.) aligns with this conveying a smoothed line from the 4 variable model onwards.

Conclusions

In conclusion, there were multiple quality of life factors to which variance and therefore influence to the Score of Happiness could be attributed. Notably in the multiple linear regression GDP per capita while significant was somewhat surprisingly not the greatest purveyor of attributable variance. There are some reservations about Social Support squared with its high variance inflation factor, and a concern about the polynomial to the order of 1 for Freedom to make life choices due to this variable exhibiting a high heteroscedasticity. It would be prudent to isolate this variable and carry out an analysis with a Non-linear regression technique such as Polynomial regression followed by a generalised Additive model. In terms of the coefficients, they were useful in observing the sensitivity of the independent variables to the dependant variable Score. However, by way of direct comparison they would benefit from some kind of scaling especially as this analysis included the use of polynomials, even though there would be a loss in absolute interpretability.
However in this analysis, the model which performed best contained four variables, and these alongside GDP per capita, Healthy life expectancy, Social support when squared and Freedom to make life choices.

```
## HEADER ####
## Who: Sophie Constant
## https://dsgarage.netlify.app/
## What: C7081 - Assignment 1 - What influences happiness?
## Last edited: <DATE TODAY in 2021-12-09 format)

## CONTENTS ####
# 00 Setup
# 01 Multiple Linear Regression
# 02 Model Selection

# 00 Setup ####

# Set your working directory setwd()

setwd("C:/Users/Sophi/OneDrive/Documents/Data Science/Module 1/Assignment 1.2/")

# Load the following packages using library () function

library(openxlsx)
```

```
## Warning: package 'openxlsx' was built under R version 4.1.1
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.1.1
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

```
library(gclus)
```

```
## Warning: package 'gclus' was built under R version 4.1.1
```

```
## Loading required package: cluster
```

```
library(PerformanceAnalytics)
```

```
## Warning: package 'PerformanceAnalytics' was built under R version 4.1.1
```

```
## Loading required package: xts
```

```
## Warning: package 'xts' was built under R version 4.1.1
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 4.1.1
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
##
## Attaching package: 'xts'
```

```
## The following objects are masked from 'package:dplyr':
##
##     first, last
```

```
##
## Attaching package: 'PerformanceAnalytics'
```

```
## The following object is masked from 'package:graphics':
##
##     legend
```

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.1
```

```
##
## Attaching package: 'psych'
```

```
## The following object is masked from 'package:car':
##
##     logit
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.1.2
```

```
library(splines)
library(wesanderson)
```

```
## Warning: package 'wesanderson' was built under R version 4.1.1
```

```
happy_19 <- read.csv("2019.csv")

summary(happy_19)
```

```
##   Overall.rank    Country.or.region      Score       GDP.per.capita
##  Min.   :  1.00   Length:156         Min.   :2.853   Min.   :0.0000
##  1st Qu.: 39.75   Class :character   1st Qu.:4.545   1st Qu.:0.6028
##  Median : 78.50   Mode  :character   Median :5.380   Median :0.9600
##  Mean   : 78.50                      Mean   :5.407   Mean   :0.9051
##  3rd Qu.:117.25                      3rd Qu.:6.184   3rd Qu.:1.2325
##  Max.   :156.00                      Max.   :7.769   Max.   :1.6840
##  Social.support  Healthy.life.expectancy Freedom.to.make.life.choices
##  Min.   :0.000   Min.   :0.0000          Min.   :0.0000
##  1st Qu.:1.056   1st Qu.:0.5477          1st Qu.:0.3080
##  Median :1.272   Median :0.7890          Median :0.4170
##  Mean   :1.209   Mean   :0.7252          Mean   :0.3926
##  3rd Qu.:1.452   3rd Qu.:0.8818          3rd Qu.:0.5072
##  Max.   :1.624   Max.   :1.1410          Max.   :0.6310
##    Generosity     Perceptions.of.corruption
##  Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.1087   1st Qu.:0.0470
##  Median :0.1775   Median :0.0855
##  Mean   :0.1848   Mean   :0.1106
##  3rd Qu.:0.2482   3rd Qu.:0.1412
##  Max.   :0.5660   Max.   :0.4530
```

```
# Scores, and independant variables from 2019

happy_192 <- happy_19[3:9]

summary(happy_192)
```

```
##      Score       GDP.per.capita   Social.support  Healthy.life.expectancy
##  Min.   :2.853   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:4.545   1st Qu.:0.6028   1st Qu.:1.056   1st Qu.:0.5477
##  Median :5.380   Median :0.9600   Median :1.272   Median :0.7890
##  Mean   :5.407   Mean   :0.9051   Mean   :1.209   Mean   :0.7252
##  3rd Qu.:6.184   3rd Qu.:1.2325   3rd Qu.:1.452   3rd Qu.:0.8818
##  Max.   :7.769   Max.   :1.6840   Max.   :1.624   Max.   :1.1410
##  Freedom.to.make.life.choices   Generosity     Perceptions.of.corruption
##  Min.   :0.0000               Min.   :0.0000   Min.   :0.0000
##  1st Qu.:0.3080               1st Qu.:0.1087   1st Qu.:0.0470
##  Median :0.4170               Median :0.1775   Median :0.0855
##  Mean   :0.3926               Mean   :0.1848   Mean   :0.1106
##  3rd Qu.:0.5072               3rd Qu.:0.2482   3rd Qu.:0.1412
##  Max.   :0.6310               Max.   :0.5660   Max.   :0.4530
```

```
# Creation of social support squared

SS_X2 <- happy_192$Social.support^2

summary(SS_X2)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.115   1.617   1.550   2.110   2.637
```

```
# Combining the transformed variables in a new data frame

happy_192_sq_cu <- head(cbind(happy_192, Social.support.sq = SS_X2, Freedom.life.choices = (p
oly(happy_192$Freedom.to.make.life.choices, 3))), 156)

# Removing the Freedom to make life choices variable causingthe multicollinearity flag

happy_192_sq_cu_exFLC <- subset(happy_192_sq_cu, select = -c(Freedom.to.make.life.choices))


# 01 Multiple Linear Regression  ####

happy_192_lm_sq_cu_exFLC <- lm(Score ~ ., data = happy_192_sq_cu_exFLC)

summary(happy_192_sq_cu_exFLC)
```

```
##      Score       GDP.per.capita   Social.support  Healthy.life.expectancy
##  Min.   :2.853   Min.   :0.0000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:4.545   1st Qu.:0.6028   1st Qu.:1.056   1st Qu.:0.5477
##  Median :5.380   Median :0.9600   Median :1.272   Median :0.7890
##  Mean   :5.407   Mean   :0.9051   Mean   :1.209   Mean   :0.7252
##  3rd Qu.:6.184   3rd Qu.:1.2325   3rd Qu.:1.452   3rd Qu.:0.8818
##  Max.   :7.769   Max.   :1.6840   Max.   :1.624   Max.   :1.1410
##    Generosity     Perceptions.of.corruption Social.support.sq
##  Min.   :0.0000   Min.   :0.0000            Min.   :0.000
##  1st Qu.:0.1087   1st Qu.:0.0470            1st Qu.:1.115
##  Median :0.1775   Median :0.0855            Median :1.617
##  Mean   :0.1848   Mean   :0.1106            Mean   :1.550
##  3rd Qu.:0.2482   3rd Qu.:0.1412            3rd Qu.:2.110
##  Max.   :0.5660   Max.   :0.4530            Max.   :2.637
##  Freedom.life.choices.1 Freedom.life.choices.2 Freedom.life.choices.3
##  Min.   :-0.22006       Min.   :-0.07457       Min.   :-0.30097
##  1st Qu.:-0.04741       1st Qu.:-0.06542       1st Qu.:-0.05903
##  Median : 0.01369       Median :-0.02610       Median :-0.01496
##  Mean   : 0.00000       Mean   : 0.00000       Mean   : 0.00000
##  3rd Qu.: 0.06428       3rd Qu.: 0.03575       3rd Qu.: 0.05239
##  Max.   : 0.13365       Max.   : 0.31344       Max.   : 0.27554
```

```
# Variance inflation factors

vif(happy_192_lm_sq_cu_exFLC)
```

```
##            GDP.per.capita            Social.support   Healthy.life.expectancy
##                  4.299047                 29.631799                  3.625958
##               Generosity Perceptions.of.corruption        Social.support.sq
##                  1.296306                  1.673428                 30.926050
##    Freedom.life.choices.1    Freedom.life.choices.2    Freedom.life.choices.3
##                  1.693338                  1.294560                  1.107911
```

```
# Cor function to explore

cor(happy_192_sq_cu_exFLC)
```

```
##                              Score GDP.per.capita Social.support
## Score                    1.00000000     0.79388287    0.777057788
## GDP.per.capita           0.79388287     1.00000000    0.754905727
## Social.support           0.77705779     0.75490573    1.000000000
## Healthy.life.expectancy  0.77988315     0.83546212    0.719009459
## Generosity               0.07582369    -0.07966231   -0.048126454
## Perceptions.of.corruption 0.38561307     0.29891985    0.181899465
## Social.support.sq        0.80657174     0.77137594    0.981175990
## Freedom.life.choices.1   0.56674183     0.37907907    0.447333164
## Freedom.life.choices.2   0.08805772     0.07854603   -0.008681204
## Freedom.life.choices.3   0.14732243     0.05992387    0.165543208
##                          Healthy.life.expectancy  Generosity
## Score                                 0.77988315  0.07582369
## GDP.per.capita                        0.83546212 -0.07966231
## Social.support                        0.71900946 -0.04812645
## Healthy.life.expectancy               1.00000000 -0.02951086
## Generosity                           -0.02951086  1.00000000
## Perceptions.of.corruption             0.29528281  0.32653754
## Social.support.sq                     0.73123345 -0.02019391
## Freedom.life.choices.1                0.39039478  0.26974181
## Freedom.life.choices.2                0.12496339  0.27138237
## Freedom.life.choices.3                0.08962565 -0.01038109
##                          Perceptions.of.corruption Social.support.sq
## Score                                    0.3856131        0.80657174
## GDP.per.capita                           0.2989198        0.77137594
## Social.support                           0.1818995        0.98117599
## Healthy.life.expectancy                  0.2952828        0.73123345
## Generosity                               0.3265375       -0.02019391
## Perceptions.of.corruption                1.0000000        0.22520390
## Social.support.sq                        0.2252039        1.00000000
## Freedom.life.choices.1                   0.4388433        0.45104996
## Freedom.life.choices.2                   0.3464215        0.04027675
## Freedom.life.choices.3                   0.1579564        0.17492329
##                          Freedom.life.choices.1 Freedom.life.choices.2
## Score                             5.667418e-01           8.805772e-02
## GDP.per.capita                    3.790791e-01           7.854603e-02
## Social.support                    4.473332e-01          -8.681204e-03
## Healthy.life.expectancy           3.903948e-01           1.249634e-01
## Generosity                        2.697418e-01           2.713824e-01
## Perceptions.of.corruption         4.388433e-01           3.464215e-01
## Social.support.sq                 4.510500e-01           4.027675e-02
## Freedom.life.choices.1            1.000000e+00          -3.073374e-17
## Freedom.life.choices.2           -3.073374e-17           1.000000e+00
## Freedom.life.choices.3            9.557920e-18          -1.788256e-17
##                          Freedom.life.choices.3
## Score                             1.473224e-01
## GDP.per.capita                    5.992387e-02
## Social.support                    1.655432e-01
## Healthy.life.expectancy           8.962565e-02
## Generosity                       -1.038109e-02
## Perceptions.of.corruption         1.579564e-01
## Social.support.sq                 1.749233e-01
## Freedom.life.choices.1            9.557920e-18
## Freedom.life.choices.2           -1.788256e-17
## Freedom.life.choices.3            1.000000e+00
```

```
# 02 Model Selection ####

# Best subset selection

happy192_sq_cu_exFLC_subset <- regsubsets(Score ~.,
                           data = happy_192_sq_cu_exFLC)


summary(happy192_sq_cu_exFLC_subset)
```

```
## Subset selection object
## Call: regsubsets.formula(Score ~ ., data = happy_192_sq_cu_exFLC)
## 9 Variables  (and intercept)
##                            Forced in Forced out
## GDP.per.capita                 FALSE      FALSE
## Social.support                 FALSE      FALSE
## Healthy.life.expectancy        FALSE      FALSE
## Generosity                     FALSE      FALSE
## Perceptions.of.corruption      FALSE      FALSE
## Social.support.sq              FALSE      FALSE
## Freedom.life.choices.1         FALSE      FALSE
## Freedom.life.choices.2         FALSE      FALSE
## Freedom.life.choices.3         FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          GDP.per.capita Social.support Healthy.life.expectancy Generosity
## 1  ( 1 ) " "            " "            " "                     " "
## 2  ( 1 ) " "            " "            "*"                     " "
## 3  ( 1 ) "*"            " "            " "                     " "
## 4  ( 1 ) "*"            " "            "*"                     " "
## 5  ( 1 ) "*"            "*"            "*"                     " "
## 6  ( 1 ) "*"            "*"            "*"                     " "
## 7  ( 1 ) "*"            "*"            "*"                     " "
## 8  ( 1 ) "*"            "*"            "*"                     "*"
##          Perceptions.of.corruption Social.support.sq Freedom.life.choices.1
## 1  ( 1 ) " "                       "*"               " "
## 2  ( 1 ) " "                       "*"               " "
## 3  ( 1 ) " "                       "*"               "*"
## 4  ( 1 ) " "                       "*"               "*"
## 5  ( 1 ) " "                       "*"               "*"
## 6  ( 1 ) "*"                       "*"               "*"
## 7  ( 1 ) "*"                       "*"               "*"
## 8  ( 1 ) "*"                       "*"               "*"
##          Freedom.life.choices.2 Freedom.life.choices.3
## 1  ( 1 ) " "                    " "
## 2  ( 1 ) " "                    " "
## 3  ( 1 ) " "                    " "
## 4  ( 1 ) " "                    " "
## 5  ( 1 ) " "                    " "
## 6  ( 1 ) " "                    " "
## 7  ( 1 ) " "                    "*"
## 8  ( 1 ) " "                    "*"
```

```
# Adding some models using nvmax function

happy192_sq_cu_exFLC_full <- regsubsets(Score ~ ., data = happy_192_sq_cu_exFLC ,
                           nvmax = 11)

happy192_sq_cu_exFLC_reg_summary <- summary(happy192_sq_cu_exFLC_full)

happy192_sq_cu_exFLC_reg_summary
```

```
## Subset selection object
## Call: regsubsets.formula(Score ~ ., data = happy_192_sq_cu_exFLC, nvmax = 11)
## 9 Variables  (and intercept)
##                            Forced in Forced out
## GDP.per.capita                 FALSE      FALSE
## Social.support                 FALSE      FALSE
## Healthy.life.expectancy        FALSE      FALSE
## Generosity                     FALSE      FALSE
## Perceptions.of.corruption      FALSE      FALSE
## Social.support.sq              FALSE      FALSE
## Freedom.life.choices.1         FALSE      FALSE
## Freedom.life.choices.2         FALSE      FALSE
## Freedom.life.choices.3         FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          GDP.per.capita Social.support Healthy.life.expectancy Generosity
## 1  ( 1 ) " "            " "            " "                     " "
## 2  ( 1 ) " "            " "            "*"                     " "
## 3  ( 1 ) "*"            " "            " "                     " "
## 4  ( 1 ) "*"            " "            "*"                     " "
## 5  ( 1 ) "*"            "*"            "*"                     " "
## 6  ( 1 ) "*"            "*"            "*"                     " "
## 7  ( 1 ) "*"            "*"            "*"                     " "
## 8  ( 1 ) "*"            "*"            "*"                     "*"
## 9  ( 1 ) "*"            "*"            "*"                     "*"
##          Perceptions.of.corruption Social.support.sq Freedom.life.choices.1
## 1  ( 1 ) " "                       "*"               " "
## 2  ( 1 ) " "                       "*"               " "
## 3  ( 1 ) " "                       "*"               "*"
## 4  ( 1 ) " "                       "*"               "*"
## 5  ( 1 ) " "                       "*"               "*"
## 6  ( 1 ) "*"                       "*"               "*"
## 7  ( 1 ) "*"                       "*"               "*"
## 8  ( 1 ) "*"                       "*"               "*"
## 9  ( 1 ) "*"                       "*"               "*"
##          Freedom.life.choices.2 Freedom.life.choices.3
## 1  ( 1 ) " "                    " "
## 2  ( 1 ) " "                    " "
## 3  ( 1 ) " "                    " "
## 4  ( 1 ) " "                    " "
## 5  ( 1 ) " "                    " "
## 6  ( 1 ) " "                    " "
## 7  ( 1 ) " "                    "*"
## 8  ( 1 ) " "                    "*"
## 9  ( 1 ) "*"                    "*"
```

```
# Taking a look at what we can see

names(happy192_sq_cu_exFLC_reg_summary)
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

```
# Having a look at R squared statistic

happy192_sq_cu_exFLC_reg_summary$rsq
```

```
## [1] 0.6505580 0.7282170 0.7686928 0.7839514 0.7902280 0.7935024 0.7951206
## [8] 0.7958454 0.7961165
```

```
# Looking at the coefficients

# Model 4

coef(happy192_sq_cu_exFLC_full, 4)
```

```
##             (Intercept)          GDP.per.capita Healthy.life.expectancy
##               3.0718027               0.6998333               1.0704885
##       Social.support.sq  Freedom.life.choices.1
##               0.5970147               3.1495323
```

```
# Model 5

coef(happy192_sq_cu_exFLC_full, 5)
```

```
##             (Intercept)          GDP.per.capita          Social.support
##               3.8234732               0.6866967              -1.5275908
## Healthy.life.expectancy       Social.support.sq  Freedom.life.choices.1
##               1.0879370               1.3028319               3.1821818
```

```
# Model 6

coef(happy192_sq_cu_exFLC_full, 6)
```

```
##               (Intercept)           GDP.per.capita           Social.support
##                 3.6248762                0.6505008               -1.2542244
##   Healthy.life.expectancy Perceptions.of.corruption        Social.support.sq
##                 1.0515827                0.7899417                1.1995574
##    Freedom.life.choices.1
##                 2.7977719
```

```
# Forwards stepwise selection

happy_192_sq_cu_exFLC_regfit_fwd <- regsubsets (Score ~ ., data = happy_192_sq_cu_exFLC ,
                          nvmax = 11, method = "forward")

summary (happy_192_sq_cu_exFLC_regfit_fwd)
```

```
## Subset selection object
## Call: regsubsets.formula(Score ~ ., data = happy_192_sq_cu_exFLC, nvmax = 11,
##     method = "forward")
## 9 Variables  (and intercept)
##                            Forced in Forced out
## GDP.per.capita                 FALSE      FALSE
## Social.support                 FALSE      FALSE
## Healthy.life.expectancy        FALSE      FALSE
## Generosity                     FALSE      FALSE
## Perceptions.of.corruption      FALSE      FALSE
## Social.support.sq              FALSE      FALSE
## Freedom.life.choices.1         FALSE      FALSE
## Freedom.life.choices.2         FALSE      FALSE
## Freedom.life.choices.3         FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: forward
##          GDP.per.capita Social.support Healthy.life.expectancy Generosity
## 1  ( 1 ) " "            " "            " "                     " "
## 2  ( 1 ) " "            " "            "*"                     " "
## 3  ( 1 ) " "            " "            "*"                     " "
## 4  ( 1 ) "*"            " "            "*"                     " "
## 5  ( 1 ) "*"            "*"            "*"                     " "
## 6  ( 1 ) "*"            "*"            "*"                     " "
## 7  ( 1 ) "*"            "*"            "*"                     " "
## 8  ( 1 ) "*"            "*"            "*"                     "*"
## 9  ( 1 ) "*"            "*"            "*"                     "*"
##          Perceptions.of.corruption Social.support.sq Freedom.life.choices.1
## 1  ( 1 ) " "                       "*"               " "
## 2  ( 1 ) " "                       "*"               " "
## 3  ( 1 ) " "                       "*"               "*"
## 4  ( 1 ) " "                       "*"               "*"
## 5  ( 1 ) " "                       "*"               "*"
## 6  ( 1 ) "*"                       "*"               "*"
## 7  ( 1 ) "*"                       "*"               "*"
## 8  ( 1 ) "*"                       "*"               "*"
## 9  ( 1 ) "*"                       "*"               "*"
##          Freedom.life.choices.2 Freedom.life.choices.3
## 1  ( 1 ) " "                    " "
## 2  ( 1 ) " "                    " "
## 3  ( 1 ) " "                    " "
## 4  ( 1 ) " "                    " "
## 5  ( 1 ) " "                    " "
## 6  ( 1 ) " "                    " "
## 7  ( 1 ) " "                    "*"
## 8  ( 1 ) " "                    "*"
## 9  ( 1 ) "*"                    "*"
```

```
# Looking at the coefficients of model 4

coef(happy_192_sq_cu_exFLC_regfit_fwd, 4)
```

```
##             (Intercept)          GDP.per.capita Healthy.life.expectancy
##               3.0718027               0.6998333               1.0704885
##       Social.support.sq  Freedom.life.choices.1
##               0.5970147               3.1495323
```

```
# Looking at the coefficients of model 5

coef(happy_192_sq_cu_exFLC_regfit_fwd, 5)
```

```
##             (Intercept)          GDP.per.capita          Social.support
##               3.8234732               0.6866967              -1.5275908
## Healthy.life.expectancy       Social.support.sq  Freedom.life.choices.1
##               1.0879370               1.3028319               3.1821818
```

```
# Backwards Selection

happy_192_sq_cu_exFLC_regfit_bwd <- regsubsets (Score ~ ., data = happy_192_sq_cu_exFLC ,
                            nvmax = 11, method = "backward")

summary (happy_192_sq_cu_exFLC_regfit_bwd)
```

```
## Subset selection object
## Call: regsubsets.formula(Score ~ ., data = happy_192_sq_cu_exFLC, nvmax = 11,
##     method = "backward")
## 9 Variables  (and intercept)
##                           Forced in Forced out
## GDP.per.capita                FALSE      FALSE
## Social.support                FALSE      FALSE
## Healthy.life.expectancy       FALSE      FALSE
## Generosity                    FALSE      FALSE
## Perceptions.of.corruption     FALSE      FALSE
## Social.support.sq             FALSE      FALSE
## Freedom.life.choices.1        FALSE      FALSE
## Freedom.life.choices.2        FALSE      FALSE
## Freedom.life.choices.3        FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: backward
##          GDP.per.capita Social.support Healthy.life.expectancy Generosity
## 1  ( 1 ) " "            " "            " "                     " "
## 2  ( 1 ) "*"            " "            " "                     " "
## 3  ( 1 ) "*"            " "            " "                     " "
## 4  ( 1 ) "*"            " "            "*"                     " "
## 5  ( 1 ) "*"            "*"            "*"                     " "
## 6  ( 1 ) "*"            "*"            "*"                     " "
## 7  ( 1 ) "*"            "*"            "*"                     " "
## 8  ( 1 ) "*"            "*"            "*"                     "*"
## 9  ( 1 ) "*"            "*"            "*"                     "*"
##          Perceptions.of.corruption Social.support.sq Freedom.life.choices.1
## 1  ( 1 ) " "                       "*"               " "
## 2  ( 1 ) " "                       "*"               " "
## 3  ( 1 ) " "                       "*"               "*"
## 4  ( 1 ) " "                       "*"               "*"
## 5  ( 1 ) " "                       "*"               "*"
## 6  ( 1 ) "*"                       "*"               "*"
## 7  ( 1 ) "*"                       "*"               "*"
## 8  ( 1 ) "*"                       "*"               "*"
## 9  ( 1 ) "*"                       "*"               "*"
##          Freedom.life.choices.2 Freedom.life.choices.3
## 1  ( 1 ) " "                    " "
## 2  ( 1 ) " "                    " "
## 3  ( 1 ) " "                    " "
## 4  ( 1 ) " "                    " "
## 5  ( 1 ) " "                    " "
## 6  ( 1 ) " "                    " "
## 7  ( 1 ) " "                    "*"
## 8  ( 1 ) " "                    "*"
## 9  ( 1 ) "*"                    "*"
```

```
# Model 4

coef(happy_192_sq_cu_exFLC_regfit_bwd, 4)
```

```
##             (Intercept)          GDP.per.capita Healthy.life.expectancy
##               3.0718027               0.6998333               1.0704885
##       Social.support.sq  Freedom.life.choices.1
##               0.5970147               3.1495323
```

```
# Model 5

coef(happy_192_sq_cu_exFLC_regfit_bwd, 5)
```

```
##             (Intercept)          GDP.per.capita          Social.support
##               3.8234732               0.6866967              -1.5275908
## Healthy.life.expectancy       Social.support.sq  Freedom.life.choices.1
##               1.0879370               1.3028319               3.1821818
```

```
# All

coef(happy192_sq_cu_exFLC_full, 5)
```

```
##             (Intercept)          GDP.per.capita          Social.support
##               3.8234732               0.6866967              -1.5275908
## Healthy.life.expectancy       Social.support.sq  Freedom.life.choices.1
##               1.0879370               1.3028319               3.1821818
```

```
# A look at the coefficents of model 9 (previous models were looked at, but this has been rem
oved for clarity and readability)

coef(happy_192_sq_cu_exFLC_regfit_bwd, 9)
```

```
##               (Intercept)            GDP.per.capita            Social.support
##                 3.5658205                 0.7015232                -1.2636506
##   Healthy.life.expectancy                Generosity Perceptions.of.corruption
##                 1.0587913                 0.3957734                 0.6514321
##         Social.support.sq    Freedom.life.choices.1    Freedom.life.choices.2
##                 1.1745291                 2.7432510                -0.2596570
##    Freedom.life.choices.3
##                 0.5857637
```

```
coef(happy_192_sq_cu_exFLC_regfit_bwd, 9)
```

```
##               (Intercept)            GDP.per.capita            Social.support
##                 3.5658205                 0.7015232                -1.2636506
##   Healthy.life.expectancy                Generosity Perceptions.of.corruption
##                 1.0587913                 0.3957734                 0.6514321
##         Social.support.sq    Freedom.life.choices.1    Freedom.life.choices.2
##                 1.1745291                 2.7432510                -0.2596570
##    Freedom.life.choices.3
##                 0.5857637
```

```
coef(happy192_sq_cu_exFLC_full, 9)
```

```
##              (Intercept)            GDP.per.capita            Social.support
##                3.5658205                 0.7015232                -1.2636506
##   Healthy.life.expectancy               Generosity Perceptions.of.corruption
##                1.0587913                 0.3957734                 0.6514321
##         Social.support.sq    Freedom.life.choices.1    Freedom.life.choices.2
##                1.1745291                 2.7432510                -0.2596570
##    Freedom.life.choices.3
##                0.5857637
```

```
# Validation set approach

# making a training set

set.seed (1)

train <- sample (c(TRUE , FALSE), nrow (happy_192_sq_cu_exFLC),
                 replace = TRUE)
test <- (!train)

# Subset selection

happy_subset <- regsubsets (Score ~ .,
                            data = happy_192_sq_cu_exFLC[train , ], nvmax = 9)

summary(happy_subset)
```

```
## Subset selection object
## Call: regsubsets.formula(Score ~ ., data = happy_192_sq_cu_exFLC[train,
##     ], nvmax = 9)
## 9 Variables  (and intercept)
##                           Forced in Forced out
## GDP.per.capita                FALSE      FALSE
## Social.support                FALSE      FALSE
## Healthy.life.expectancy       FALSE      FALSE
## Generosity                    FALSE      FALSE
## Perceptions.of.corruption     FALSE      FALSE
## Social.support.sq             FALSE      FALSE
## Freedom.life.choices.1        FALSE      FALSE
## Freedom.life.choices.2        FALSE      FALSE
## Freedom.life.choices.3        FALSE      FALSE
## 1 subsets of each size up to 9
## Selection Algorithm: exhaustive
##          GDP.per.capita Social.support Healthy.life.expectancy Generosity
## 1  ( 1 ) " "            " "            " "                     " "
## 2  ( 1 ) " "            " "            "*"                     " "
## 3  ( 1 ) " "            " "            "*"                     " "
## 4  ( 1 ) "*"            " "            "*"                     " "
## 5  ( 1 ) "*"            " "            "*"                     " "
## 6  ( 1 ) "*"            " "            "*"                     "*"
## 7  ( 1 ) "*"            "*"            "*"                     "*"
## 8  ( 1 ) "*"            "*"            "*"                     "*"
## 9  ( 1 ) "*"            "*"            "*"                     "*"
##          Perceptions.of.corruption Social.support.sq Freedom.life.choices.1
## 1  ( 1 ) " "                       "*"               " "
## 2  ( 1 ) " "                       "*"               " "
## 3  ( 1 ) " "                       "*"               "*"
## 4  ( 1 ) " "                       "*"               "*"
## 5  ( 1 ) "*"                       "*"               "*"
## 6  ( 1 ) "*"                       "*"               "*"
## 7  ( 1 ) "*"                       "*"               "*"
## 8  ( 1 ) "*"                       "*"               "*"
## 9  ( 1 ) "*"                       "*"               "*"
##          Freedom.life.choices.2 Freedom.life.choices.3
## 1  ( 1 ) " "                    " "
## 2  ( 1 ) " "                    " "
## 3  ( 1 ) " "                    " "
## 4  ( 1 ) " "                    " "
## 5  ( 1 ) " "                    " "
## 6  ( 1 ) " "                    " "
## 7  ( 1 ) " "                    " "
## 8  ( 1 ) " "                    "*"
## 9  ( 1 ) "*"                    "*"
```

```
# model matrix from test data

happy_test_mat <- model.matrix (Score ~ ., data = happy_192_sq_cu_exFLC[test , ])

# Validation

happy_val_errors = rep(NA,9)
for (i in 1:9) {
  coefi = coef(happy_subset, id = i)
  pred = happy_test_mat[,names(coefi)] %*% coefi
  happy_val_errors[i] = mean(((happy_192$Score[test]) - pred)^2)
}

# Taking a look

happy_val_errors
```

```
## [1] 0.4618958 0.3742432 0.3177228 0.2937440 0.2938859 0.3087733 0.3036925
## [8] 0.3013946 0.3108608
```

```
# How many variables are in the best performing model?

which.min(happy_val_errors)
```

```
## [1] 4
```

```
# There are four variables in the best performing model

coef(happy_subset, 4)
```

```
##             (Intercept)          GDP.per.capita Healthy.life.expectancy
##               2.8565983               0.6561909               1.3606188
##       Social.support.sq  Freedom.life.choices.1
##               0.6415986               2.9168856
```

```
# Writing a prediction method for regsubsets() (As it doesn't have one)

predict.regsubsets <- function (object, newdata, id, ...) {
  form <- as.formula (object$call[[2]])
  mat <- model.matrix (form, newdata)
  coefi <- coef (object, id = id)
  xvars <- names (coefi)
  mat[, xvars] %*% coefi
}

# Best subset selection on the full dataset, and four variable model

happy_regfit_best <- regsubsets(Score ~ ., data = happy_192_sq_cu_exFLC,
                           nvmax = 11)

coef(happy_regfit_best, 4)
```

```
##            (Intercept)          GDP.per.capita Healthy.life.expectancy
##              3.0718027               0.6998333               1.0704885
##      Social.support.sq  Freedom.life.choices.1
##              0.5970147               3.1495323
```

```
# Cross-validation to choose a model in the group of different sized models

# making folds and allocating observations

k <- 10
n <- nrow (happy_192_sq_cu_exFLC)
set.seed(1)
folds <- sample(rep (1:k, length = n))

cv.errors <- matrix (NA, k, 11,
                     dimnames = list (NULL , paste (1:11)))


# Writing a loop to perform the cross validation


for (j in 1:k) {
  happy_best_fit <- regsubsets(Score ~ .,
                          data = happy_192_sq_cu_exFLC[folds != j, ],
                          nvmax = 9)
  for (i in 1:9) {
    pred <- predict(happy_best_fit, happy_192_sq_cu_exFLC[folds == j, ], id = i)
    cv.errors[j, i] <-
      mean((happy_192_sq_cu_exFLC$Score[folds == j] - pred)^2)
  }
}


# Using apply to average over the columns to see a vector
# where ith element is the cross validation error for the i-variable model

mean.cv.errors <- apply (cv.errors , 2, mean)

# taking a look

mean.cv.errors
```

```
##         1         2         3         4         5         6         7         8
## 0.4362209 0.4140006 0.3335294 0.3122921 0.3118353 0.3112585 0.3095902 0.3089214
##         9        10        11
## 0.3042268        NA        NA
```

Graphs and Figures

Fig.1. Correlation Matrix of the dependant variable Score and independant variables

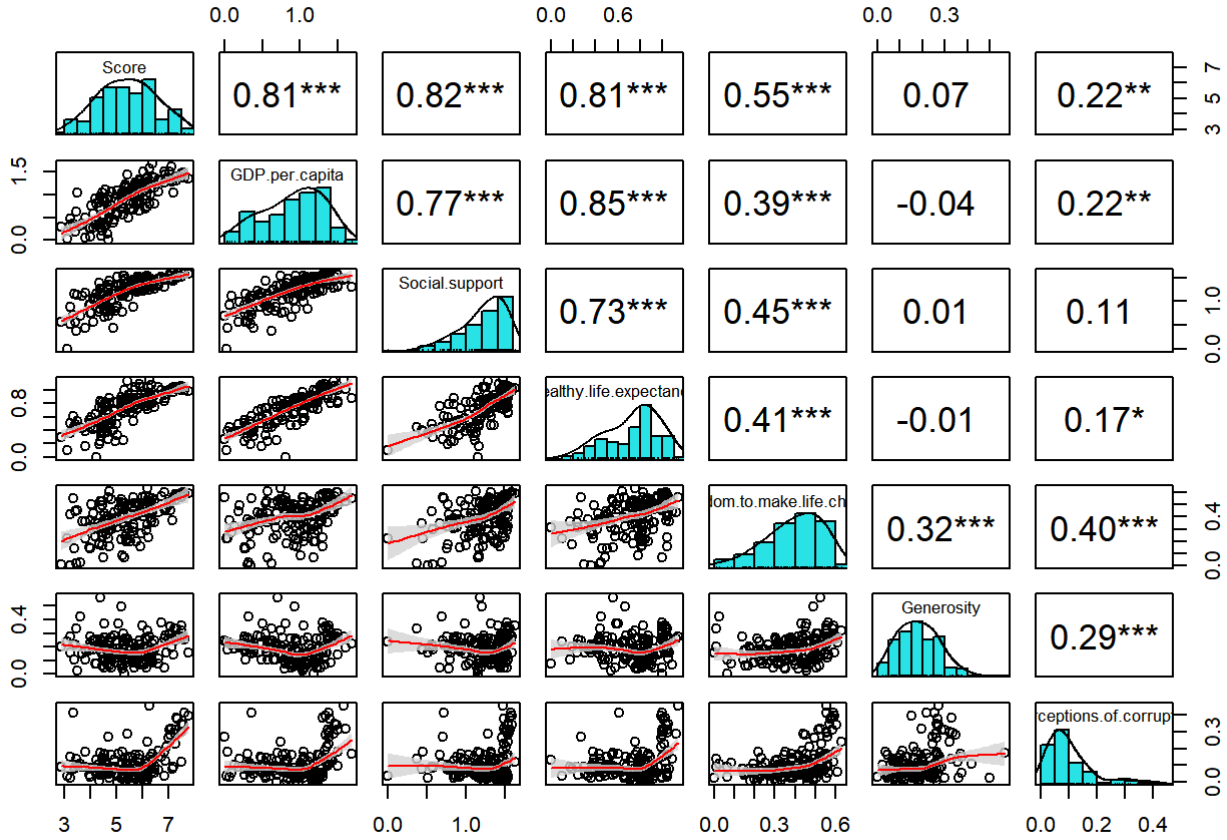


Fig. 2. Four plots for RSS, Adjusted Rsq, Cp, and BIC. For Adjusted Rsq the red dot identifies the model with the highest value. For Cp and BIC, the red dot identifies the models with the lowest value

```
## [1] 7
```

```
## [1] 6
```

```
## [1] 4
```

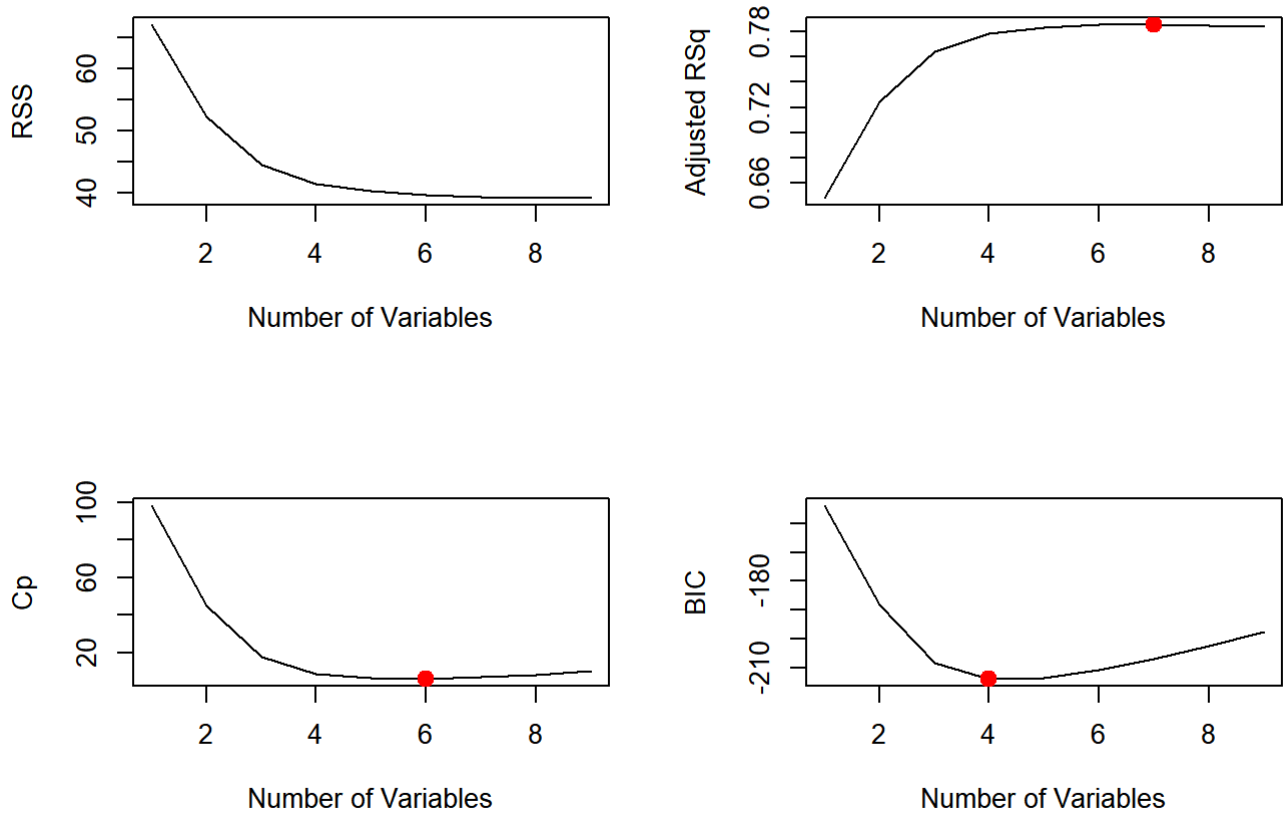


Fig. 4. Four plots for RSS, Adjusted Rsq, Cp, and BIC, The models starting with the first being from the top of the y axis, with the values pertaining to the particular statistics on the y axis. The black and grey squares identifying the independant variables present in each model, with the white squares being an absence of the independant variable in the model from the perspective of that statistic represented in the plot.

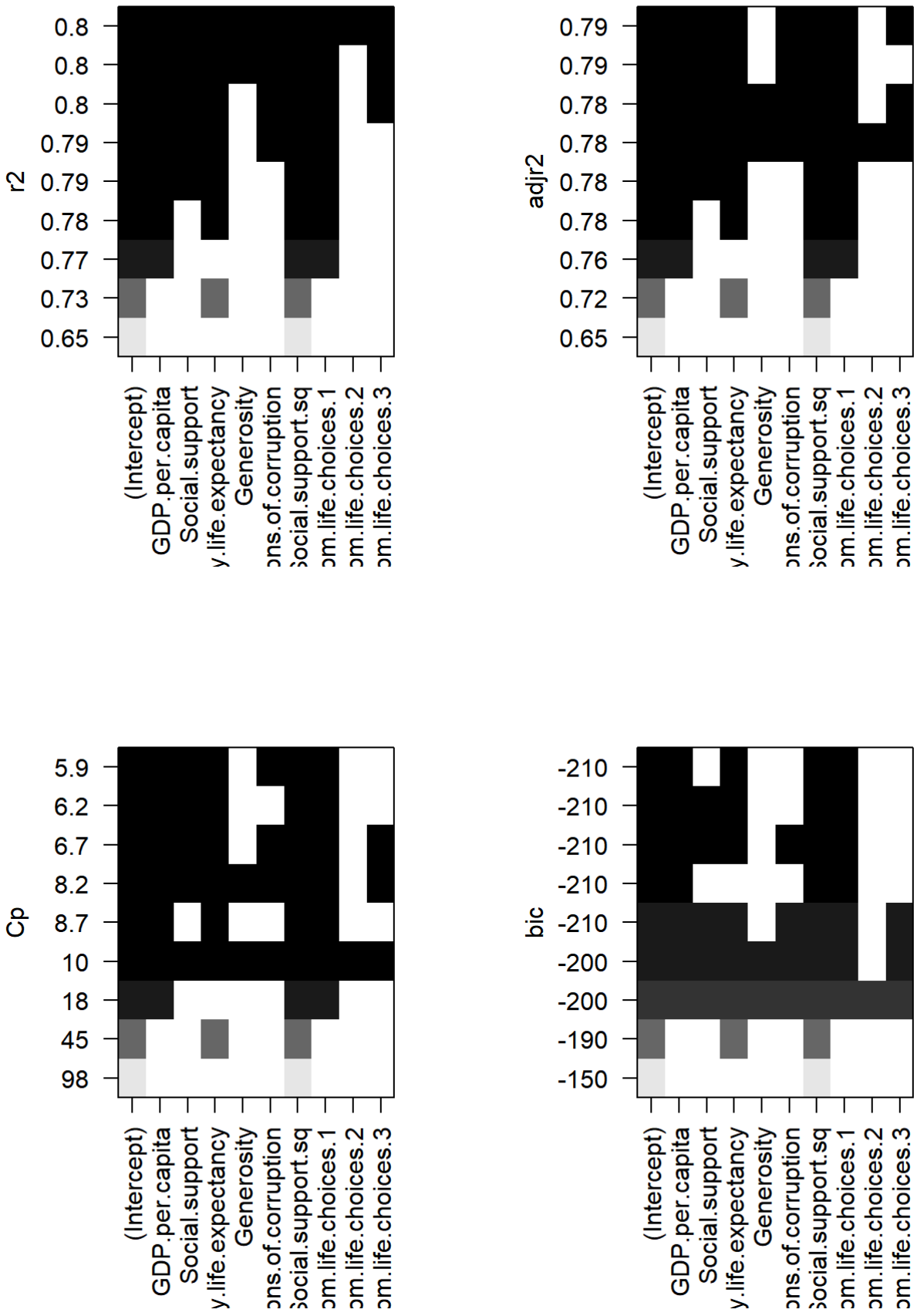
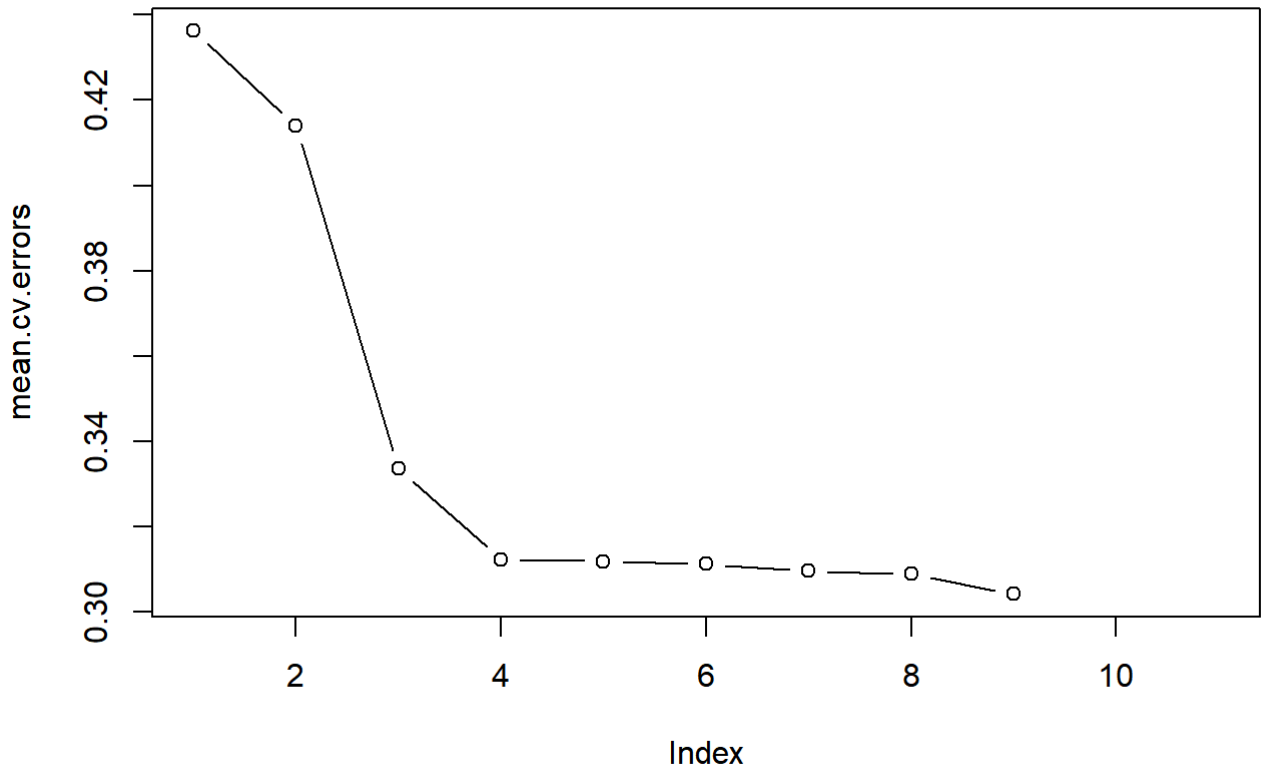


Fig. 5. A plot of the four variable model following the validation set approach and cross validation



References

Ma, Y., Liu, A., Hu, X. and Shao, Y. (2020) Happiness Score Identification: a Regression Approach. EDP Sciences, . doi 10.1051/e3sconf/202021801051

Nakamura, J.S., Delaney, S.W., Diener, E., VanderWeele, T.J. and Kim, E.S. (2021a) 'Are all domains of life satisfaction equal? Differential associations with health and well-being in older adults', Quality of life research : an international journal of quality of life aspects of treatment, care and rehabilitation, . doi: 10.1007/s11136-021-02977-0 [doi].

Network, U.S. (2019) 'World Happiness Report 2019', Retreived from: https://worldhappiness.report/ed (https://worldhappiness.report/ed), 2019, pp. 76.

Gareth, J., Daniela, W., Trevor, H. and Robert, T., 2013. '3 Linear Regression' An introduction to statistical learning: with applications in R. Spinger. pp 59-128

Gareth, J., Daniela, W., Trevor, H. and Robert, T., 2013. '6 Linear Model Selection and Regularization' An introduction to statistical learning: with applications in R. Spinger. pp 225-228

Acevedo, M.F., 2012. 'Chapter 6 Regression Data' analysis and statistics for geography, environmental science, and engineering. Crc Press. Pp177-223