Final assignment Applied

Sara Dellacasa

The aim of this project is to analyze, through multiple regression, the relationship between perceived happiness ('Happiness Score') and a set of psychological and socioeconomic variables, building upon the work of the World Happiness Report.The analysis could be useful to identify possible effective political interventions aimed at improving the quality of life in different national contexts.

The dataset refers to the year 2019 and was downloaded and adapted from Kaggle, the link to the original file is the following: https://www.kaggle.com/datasets/mirkoferretti/data-visualization-project-trendspotters

The data in this CSV come from two main sources:

1. World Happiness Report: Provides variables related to happiness and perceived quality of life (e.g., "Happiness Score" and "Social support").
2. World Bank: Integrates additional socioeconomic variables, expanding the analysis compared to the original study of the World Happiness Report with variables of a different nature.

The dataset includes 139 observations referring to as many countries of the world and contains the following variables:

Happiness Score the response variable measures overall well-being and life satisfaction of individuals in each country based on responses to the Cantril ladder question, where individuals rate their current life on a scale from 0 (worst possible life) to 10 (best possible life).

Other variables of my dataset are the followings: Country name, Region (geographical area), Log GDP per capita, Health Expenditure (Percentage of GDP spent on health by the country), Education Expenditure(Percentage of GDP), Unemployment (proportion of the working age population),Log CO2 Emissions (tons per capita), Annual population growth rate(%), Old population(equal to 1 when the percentage of people over 65 exceeds the mean, 0 otherwise).The other variables are national averages of binary survey responses (0 or 1) from the World Happiness Report 2019, measuring aspects such as social support, freedom of choice, generosity (adjusted for GDP), perceived corruption, and the frequency of positive or negative emotions experienced by respondents.

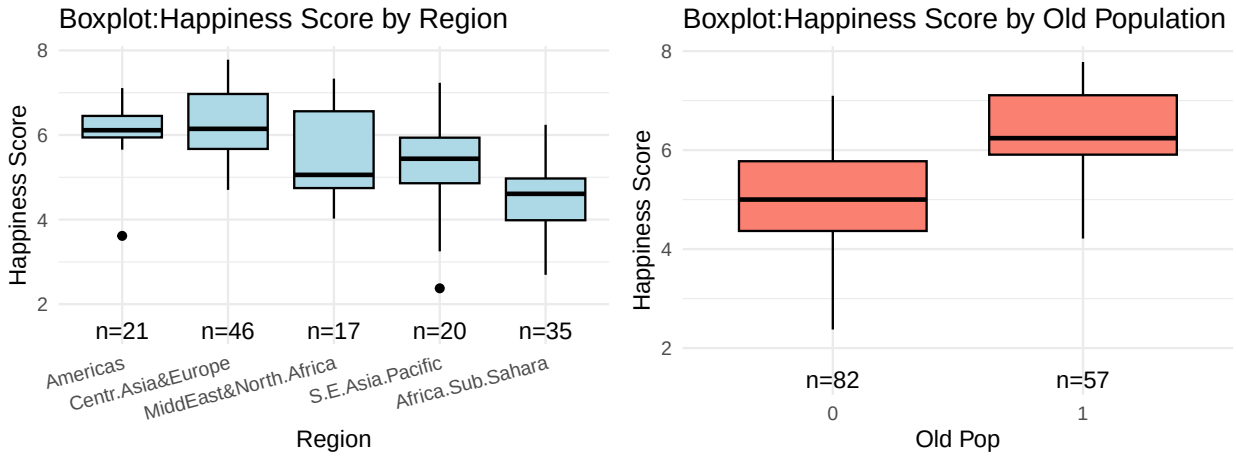
After addressing missing data (four NA values replaced with the respective variable means) and recoding the categorical variable "Region" by merging certain heterogeneous categories (North, South and Central America=Americas), the distribution of both continuous and discrete variables is as follows:

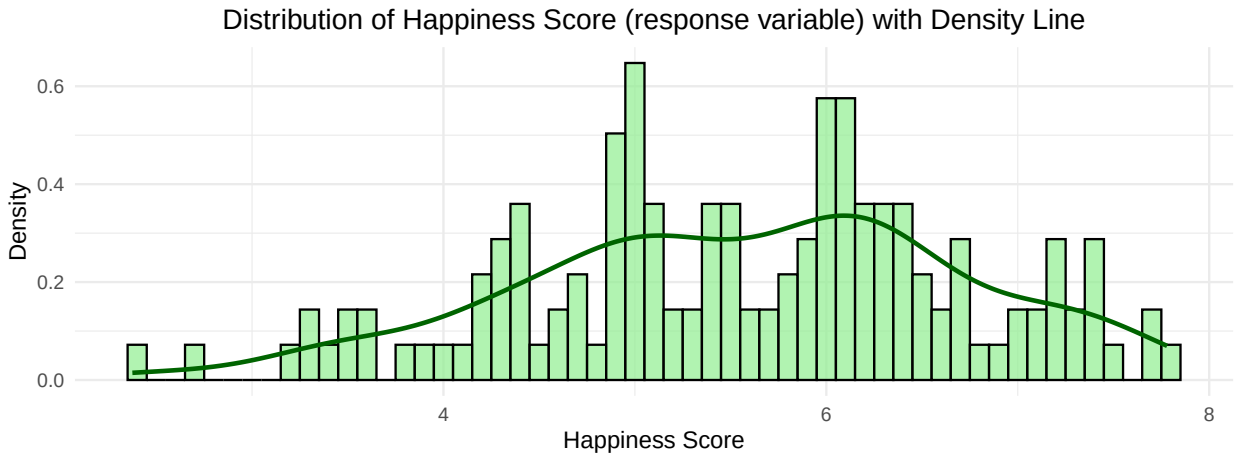
Exploratory analysis

```
not_numcols <- c("Old_pop", "Region_factor", "Country.Code", "Year", "Country")
summary(happscoredb[ , !names(happscoredb) %in% not_numcols])
```

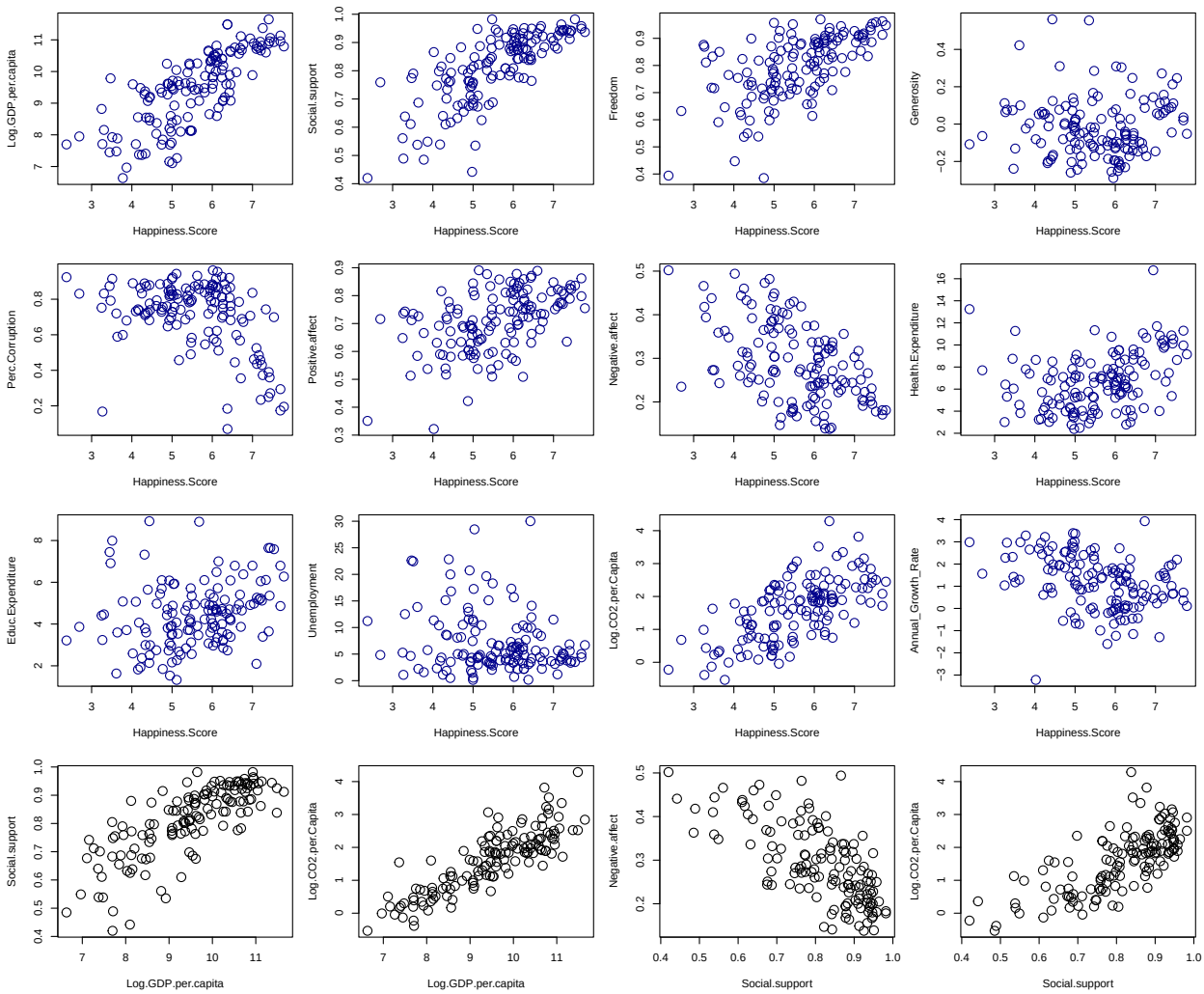
```
##  Happiness.Score Log.GDP.per.capita Social.support      Freedom
##  Min.   :2.375   Min.   : 6.635     Min.   :0.4200   Min.   :0.3850
##  1st Qu.:4.889   1st Qu.: 8.560     1st Qu.:0.7570   1st Qu.:0.7170
##  Median :5.626   Median : 9.592     Median :0.8420   Median :0.8220
##  Mean   :5.564   Mean   : 9.467     Mean   :0.8115   Mean   :0.7946
##  3rd Qu.:6.330   3rd Qu.:10.432     3rd Qu.:0.9120   3rd Qu.:0.8890
##  Max.   :7.780   Max.   :11.648     Max.   :0.9820   Max.   :0.9700
##    Generosity       Perc.Corruption  Positive.affect  Negative.affect
##  Min.   :-0.28900   Min.   :0.0700   Min.   :0.3220   Min.   :0.1380
##  1st Qu.:-0.12950   1st Qu.:0.6515   1st Qu.:0.6345   1st Qu.:0.2225
##  Median :-0.04500   Median :0.7570   Median :0.7250   Median :0.2750
##  Mean   :-0.01829   Mean   :0.7108   Mean   :0.7097   Mean   :0.2902
##  3rd Qu.: 0.07350   3rd Qu.:0.8480   3rd Qu.:0.7955   3rd Qu.:0.3575
##  Max.   : 0.56100   Max.   :0.9630   Max.   :0.8910   Max.   :0.5020
##  Health.Expenditure Educ.Expenditure  Unemployment    Annual_Growth_Rate
##  Min.   : 2.388     Min.   :1.326    Min.   : 0.147   Min.   :-3.2180
##  1st Qu.: 4.510     1st Qu.:3.389    1st Qu.: 3.581   1st Qu.: 0.3405
##  Median : 6.578     Median :4.256    Median : 5.020   Median : 1.2160
##  Mean   : 6.681     Mean   :4.428    Mean   : 7.061   Mean   : 1.1651
##  3rd Qu.: 8.483     3rd Qu.:5.244    3rd Qu.: 9.080   3rd Qu.: 1.9825
##  Max.   :16.767     Max.   :8.927    Max.   :30.010   Max.   : 3.9310
##  Log.CO2.per.Capita
##  Min.   :-0.535
##  1st Qu.: 0.827
##  Median : 1.805
##  Mean   : 1.601
##  3rd Qu.: 2.241
##  Max.   : 4.289
```

For categorical and binary variables, the distribution can be represented using boxplots of the Happiness Score across the categories defined by each dummy variable:





Here in the previous plot we have the distribution of our response variable, we can study also the relationship between the explanatory variables and the response variable using scatterplots (in blue). In the black plots, we display the relationship between pairs of explanatory variables that show a high correlation (greater than 0.65) based on the correlation matrix.



As highlighted by the correlation matrix, a strong relationship between logGDP and logCO2 is clearly visible.The other correlations identified in the matrix also appear to be confirmed by the scatterplots. After performing variable selection, it will be necessary to compute the Variance Inflation Factors (VIF) to assess potential multicollinearity issues in the regression model.

Variable selection:

First we compute a linear regression with all the predictors:

```
ols = lm( Happiness.Score ~ Log.GDP.per.capita + Social.support  + Freedom + Generosity + Perc.Corruptio
summary(ols)
```

The summary output shows that several variables are not statistically significant, likely due to the inclusion of too many predictors in the model, some of which are highly correlated. This multicollinearity suggests that the model, in its current form, may not be appropriate to reliably explain the response variable.

Considering the assumed linear regression model we can perform a best subset selection in order to explore all the possible best combinations for the specific numer of covariates we want to keep.

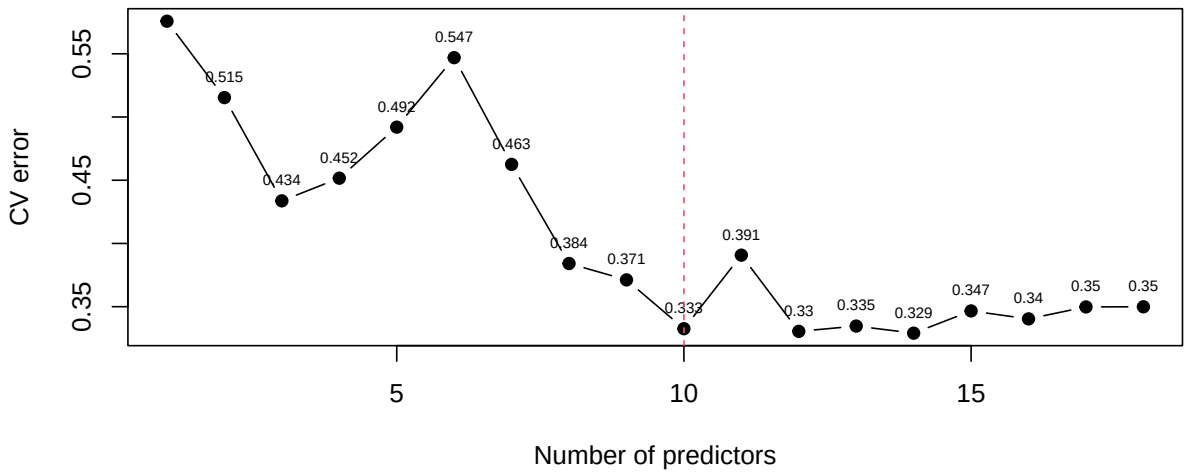
```
bestsubsett=regsubsets(Happiness.Score ~ ., data = happscore, nvmax = ncol(happscore)-1+3)
summary1=summary(bestsubsett)
```

The results of the best subset selection are saved in summary1, and we use plots of the selection criteria to assess which model is recommended by the method.

The following methods were analyzed:
- Best subset selection, which suggests retaining between 8 and 10 predictors
- Backward selection suggested retaining between 8 and 10 predictors, largely overlapping with those identified by the best subset selection method.
- Forward selection led to similar conclusions, with BIC, AIC, and Mallow's Cp plots all recommending models with approximately 9 to 10 variables. However, in this case, the variable "Old population" was preferred over "Perception of corruption."
- For shrinkage methods, only Lasso Regression was considered, as it is the only approach that performs variable selection. The Lasso model included 13 predictors, all of which were already identified by the other methods.

Considering all selection techniques, the best subset selection was ultimately chosen, as it compares all 2^p possible combinations of variables, unlike forward and backward methods which follow a sequential process and do not evaluate all possible subsets for each model size.

The best subset method suggested retaining between 9 and 11 predictors. Although a model with fewer parameters would have been preferred, the cross-validation error plot showed that a model with 9 variables had a noticeably higher CV error (around 0.40) compared to a model with 10 variables (minimum at approximately 0.33).



The ten best predictors selected through the best subset method are as follows:: - LogGDP per Capita - Social Support - Freedom - Perception of Corruption - Positive Affect - Negative Affect - Unemplyment - Region factor: South&East Asia & Pacific - Region Factor: Africa Sub-Sahara - Health Expenditure * OldPop. To include only the significant levels of the categorical variable Region_factor, two binary variables were created: "Sub-Saharan Africa" and "South & East Asia & Pacific." With these adjustments, we can now proceed to estimate the regression model using the selected variables.

After performing variable selection, we assess the presence of potential multicollinearity issues among the selected variables by analyzing the Variance Inflation Factors (VIF).

```r
vif(bestregres)
```

```
## Log.GDP.per.capita     Social.support            Freedom    Perc.Corruption
##           4.595802           3.875324           2.371474           1.496482
##    Positive.affect    Negative.affect       Unemployment  Africa.Sub.Sahara
##           1.844340           2.191130           1.293876           2.358796
##    HealthExpOldpop   S.E.Asia.Pacific
##           1.886575           1.258569
```

From this, we can observe that LogGDP remains slightly more problematic than the other variables but it does not represent a problem. So all the vifs are acceptable.

Regression diagnostics and unusual observations

The following diagnostic analysis reports the results of the assumption checks carried out on the first regression model. 1.Homoschedasticity: refers to the assumption that the variance of the errors (residuals) is constant across all levels of the independent variables. This means that the spread of residuals does not change as the fitted values increase or decrease.The homoscedasticity assumption was tested using a residual plot, which revealed mild heteroscedasticity issues.
To address this, a log transformation of the response variable (Happiness Score) was applied, but the heteroscedasticity did not significantly improve, and the model's overall diagnostic performance worsened.
A second attempt using Weighted Least Squares (WLS) improved the residual plot but resulted in a distorted Q-Q plot, displaying an S-shaped pattern inconsistent with the linear model assumptions.
Since neither transformation fully resolved the issues and both introduced additional complications, we opted to retain the simpler multivariate linear regression model, which also offers clearer interpretability.

2.Linear structure of the relationship between the predictors and the response: this assumption was previously assessed using scatterplots in the initial phase of the project and further confirmed through residual plots for each explanatory variable, so the assumption is respected.

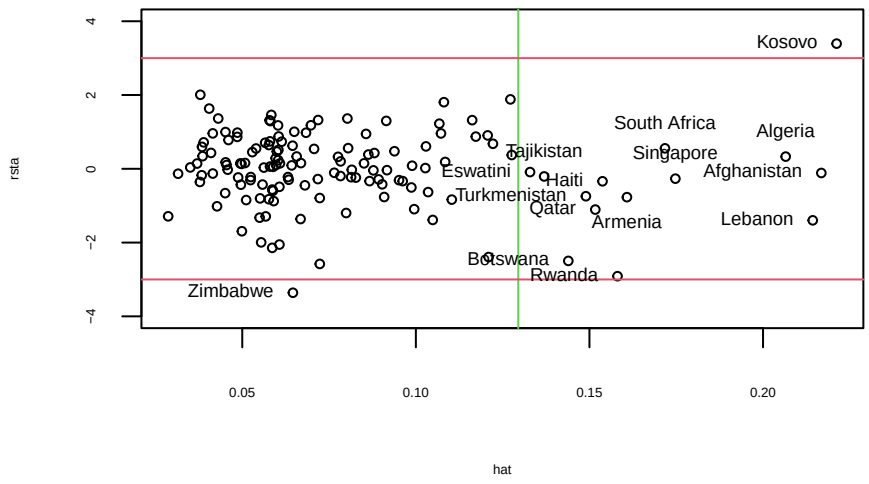
3.Normality assumptions of errors: We checked this assumption using a QQplot of residuals and the Shapiro-test.Although the Q-Q plot suggested a distribution roughly consistent with normality, the histogram appears slightly skewed to the right. Additionally, the Shapiro-Wilk test returned a p-value of 0.007817 that does not allow us to accept the null hypothesis of normality.

4.Uncorrelation of errors: To assess the assumption, residuals were plotted on a world map to detect potential spatial clustering. Some associations between residuals were observed, suggesting the presence of clusters. This may lead to an underestimation of the standard errors of the coefficients.

5.Unusual Observation: We proceed by analyzing high leverage points, outliers, and influential observations

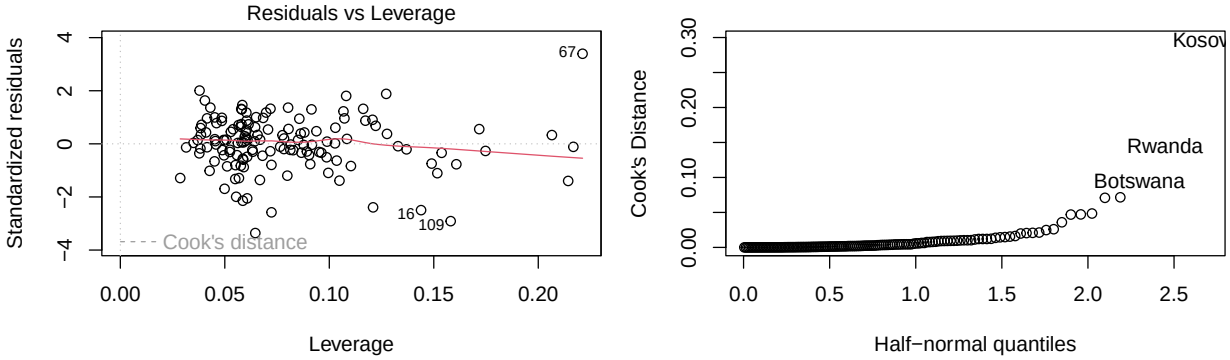
High leverage points and Outliers:

To visualize all leverage points and Outliers, it is useful to use the plot.



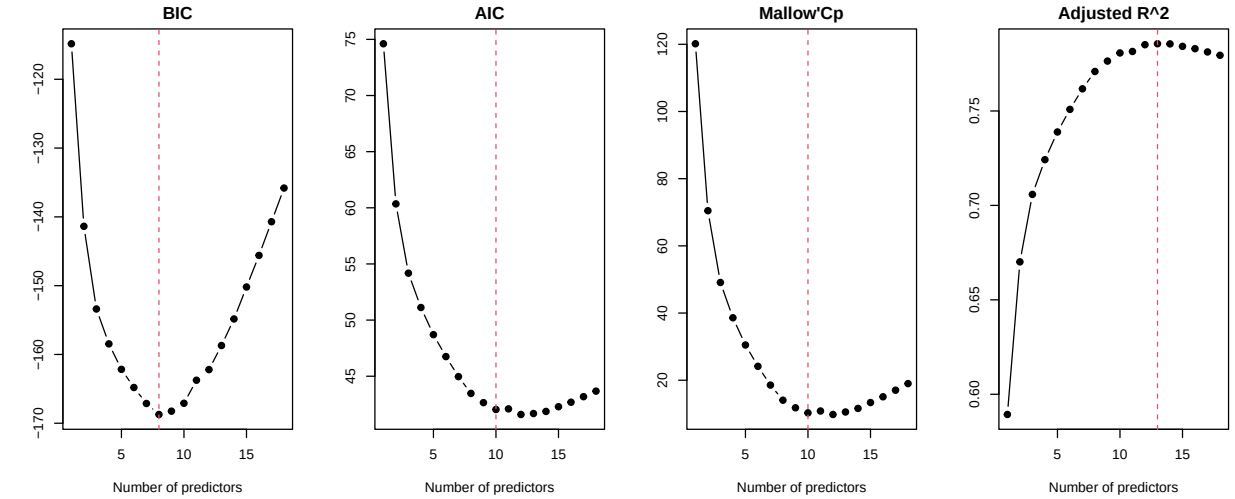
Here we can see all the high leverage points (except for Zimbabwe that is not) and the outliers: Zimbabwe and Kosovo(that is both outlier and high leverage)

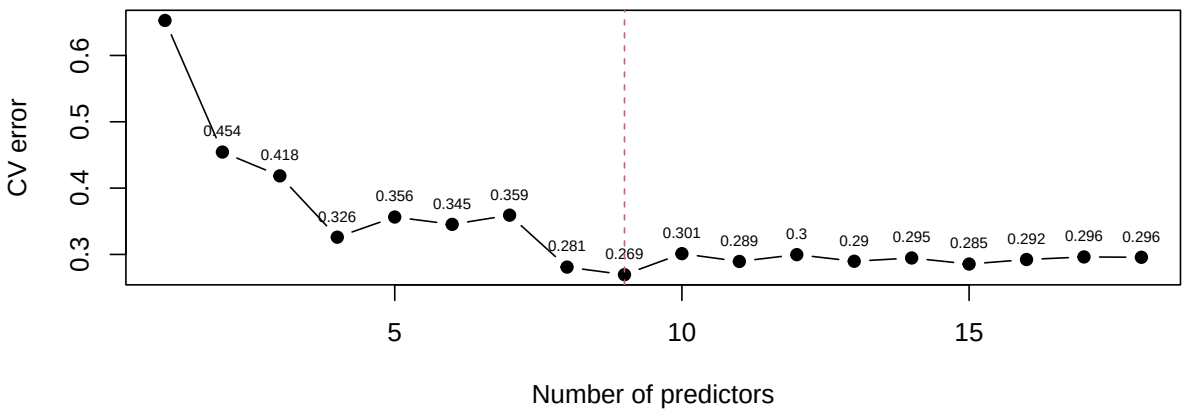
Influential point: Using Cook's distance, we will identify which points have the greatest influence on the regression.



Given that Kosovo was identified as both an outlier and an influential observation, and considering that previous model adjustments did not yield satisfactory improvements, Kosovo and Zimbabwe were excluded from the analysis to mitigate their undue influence on the model estimates.

We then repeated the variable selection and diagnostic process. For the same reason exposed for the previous model we will use the best method.
This time, AIC, BIC, and Mallow's Cp consistently suggested a model with 10 predictors, while the cross-validation results indicated a preference for a model with 9 predictors. Since the tenth predictor, "Freedom," was previously found to be only marginally significant, we proceed by estimating the model excluding this variable.





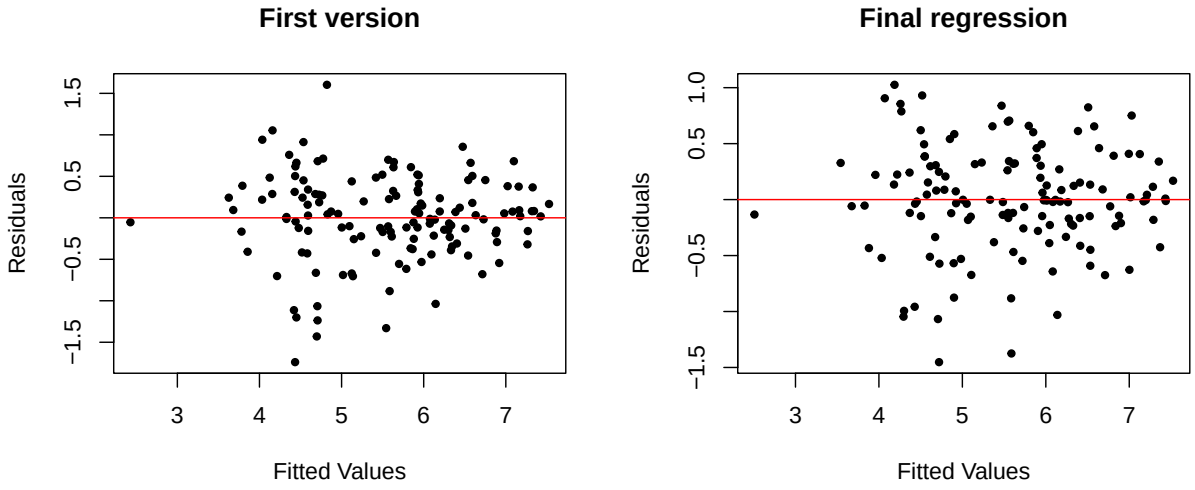
This time variables are the same as before (only twist in the Freedom one that we decided to exclude), these are: Log.GDP.per.capita - Social.support - Perc.Corruption - Positive.affect - Negative.affect - Unemployment - Region_factor: S.E.Asia.Pacific - Region_factorAfrica: Sub.Sahara - HealthExp * Oldpop.

We perform the regression using the selected variables and we check again the VIF's.

```
## Log.GDP.per.capita     Social.support    Perc.Corruption    Positive.affect
##           4.576307           3.855466           1.359747           1.375525
##    Negative.affect       Unemployment    HealthExpOldpop  S.E.Asia.Pacific1
##           2.292146           1.273634           1.874042           1.248593
## Africa.Sub.Sahara1
##           2.325916
```

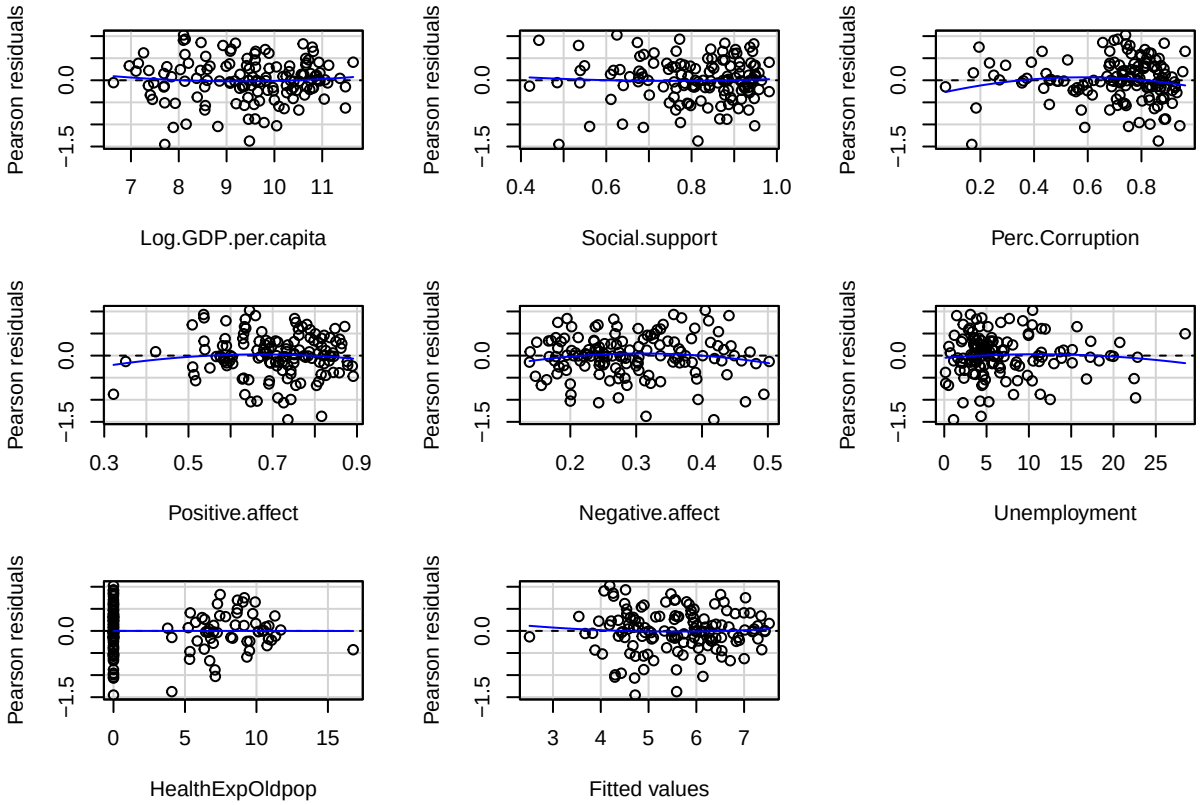
We therefore re-ran the diagnostic analysis, presenting the updated plots.

1.Homoscedasticity:The final plot still displays a slight megaphone-shaped pattern. However, the reduced influence of unusual observations suggests that the model specification has improved, and a more stable variance structure compared to the initial model.

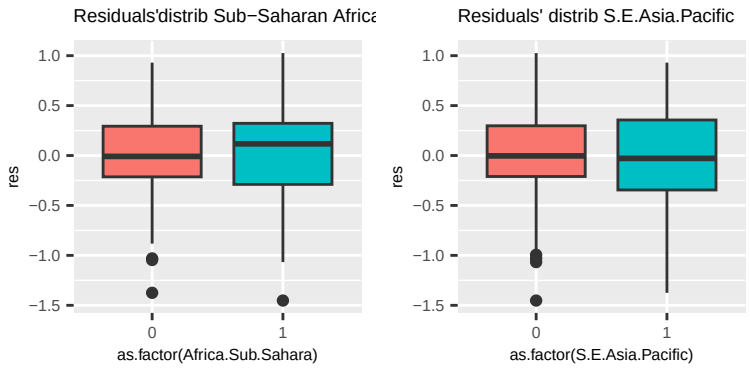


2. The linear structure between the predictors and the response is assessed using the following plots. To formally test linearity, we also employed an R function that performs a hypothesis test where H0 assumes a linear relationship, and H1 indicates non-linearity. The p-values for all predictors were greater than 0.05; therefore, we fail to reject the null hypothesis of linearity.

```
library(car)
residualPlots(finalreg,terms = ~ . - Africa.Sub.Sahara - S.E.Asia.Pacific)
```

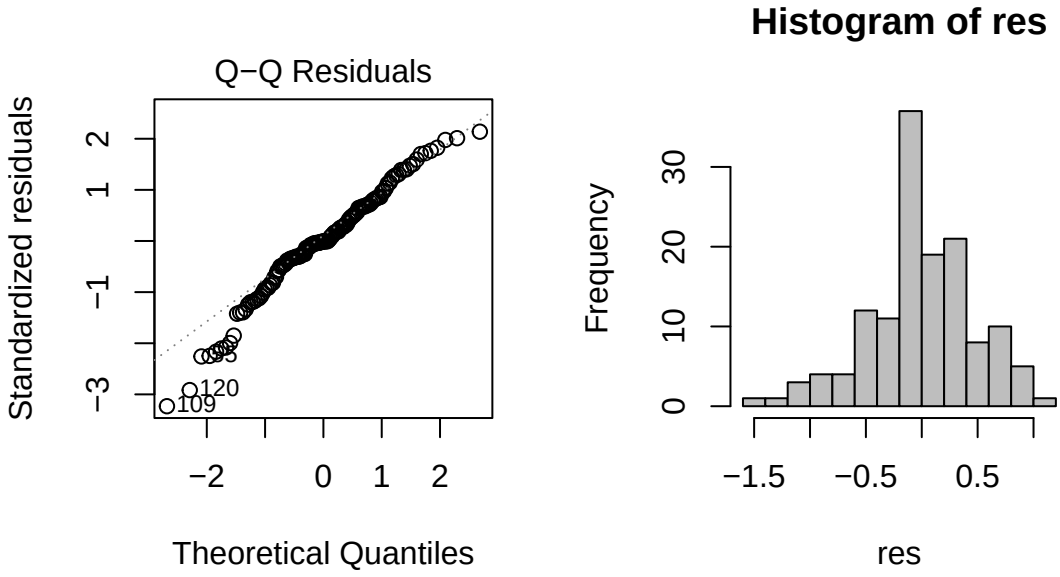


For the categorical variables, we can use boxplots.



3.Normality assumption of errors: The normality assumption is assessed using Q-Q plots, the Shapiro-Wilk test, and additional plots. The residuals in the Q-Q plot and histogram appear approximately normal; however, the Shapiro-Wilk test still does not allow us to accept the null hypothesis of normality, despite a p-value that is very close to the conventional threshold.

```
par(mfrow=c(1, 2))
plot(finalreg, which = 2)
hist(res,col = "grey")
```

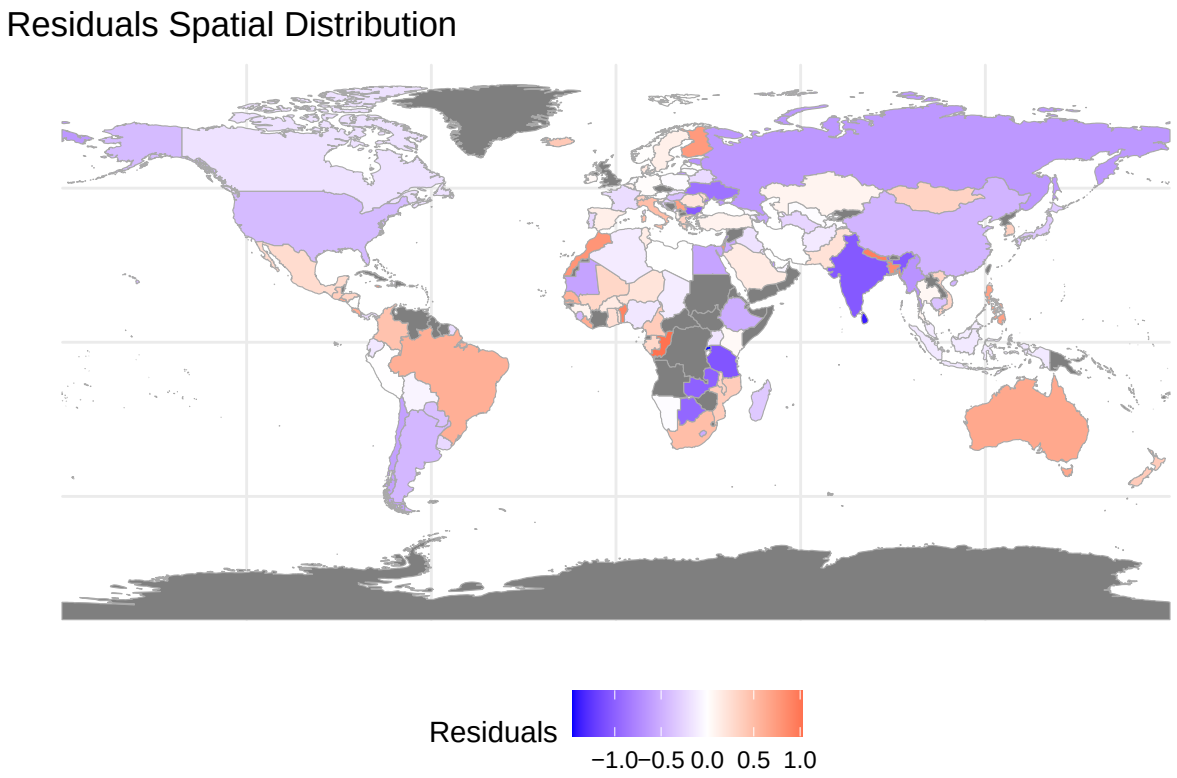



```
shapiro.test(residuals(finalreg))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(finalreg)
## W = 0.97947, p-value = 0.03673
```

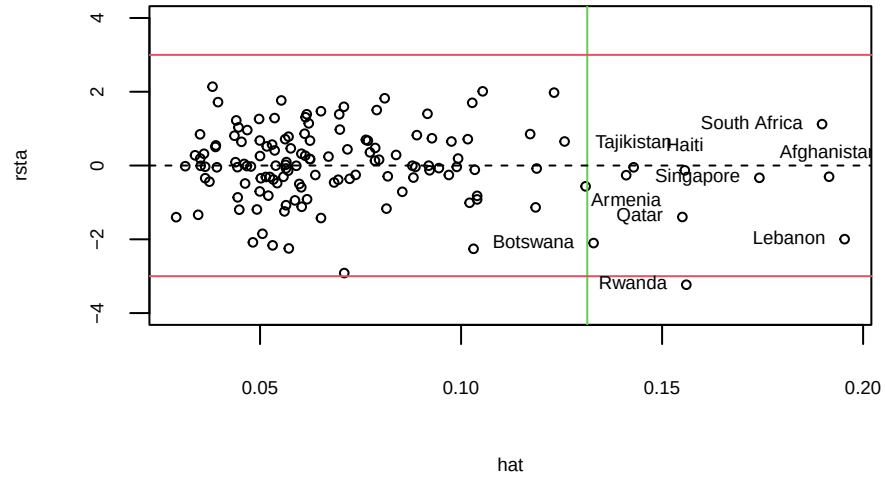
4.Uncorrelation of errors: To assess this assumption, residuals were plotted on a world map to identify potential spatial clusters. Due to the lack of temporal data, no time-related analysis was performed. Compared to

the initial model, residual values have generally decreased, although some spatial clustering is still present.



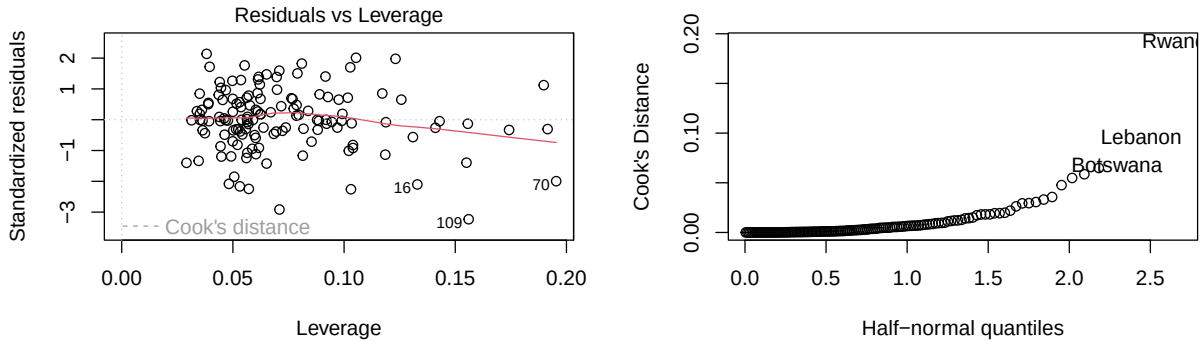
5. Unusual Observations: We now proceed with the analysis of high leverage points, outliers, and influential observations.

To visualize all leverage points, it is useful to use the plot.



In order to check the influential points we analyze Cook's distance:

```
par(mfrow=c(1, 2))
plot(finalreg, which=5)
cook= cooks.distance(finalreg)
halfnorm(cook, 3, labs = paste(happscoredb_filt$Country), ylab = "Cook's Distance")
```

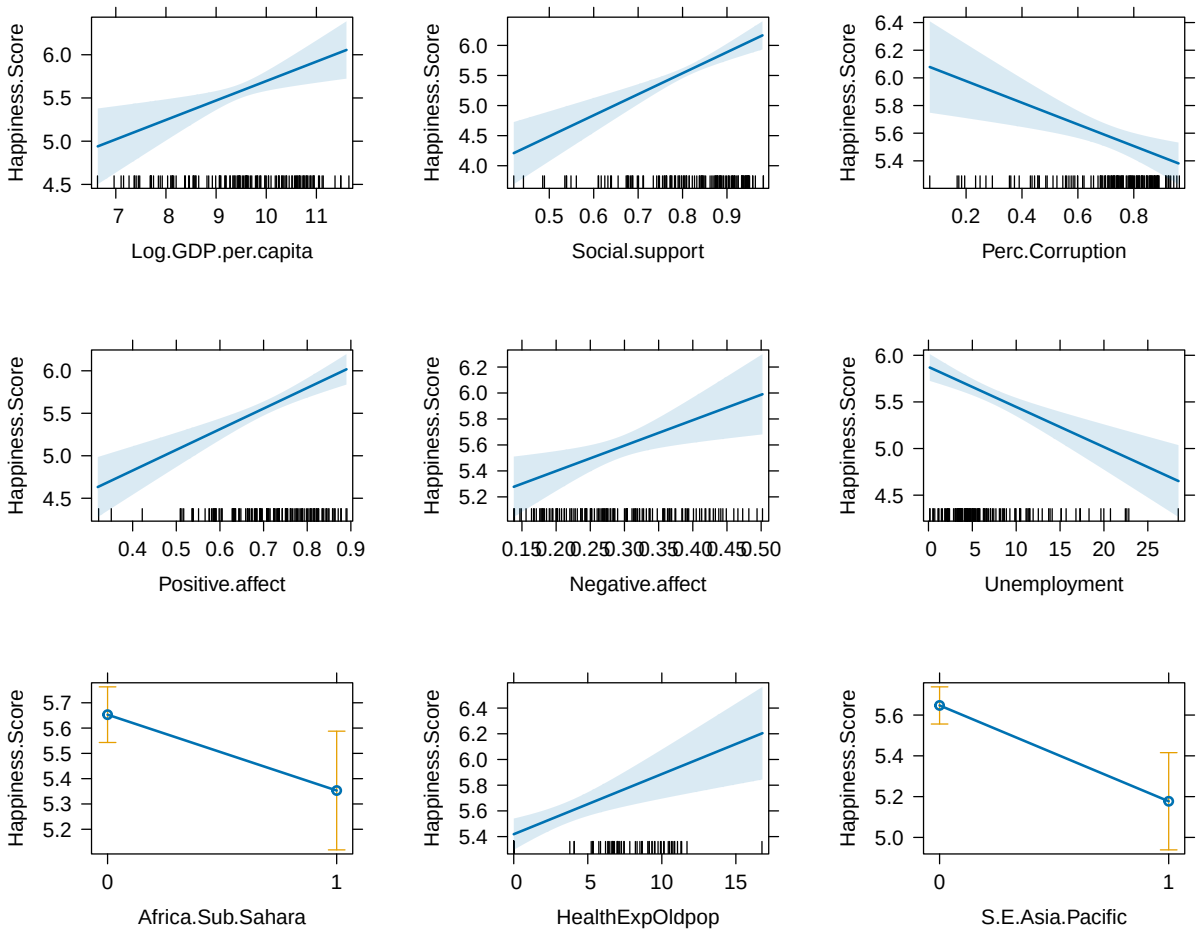


Final results

```
summary(finalreg)
```

```
##
## Call:
## lm(formula = Happiness.Score ~ ., data = happinessdb[best_variables])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.45243 -0.21719 -0.00387  0.31677  1.02567
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -0.851519   0.870148  -0.979 0.329643
## Log.GDP.per.capita   0.224873   0.076659   2.933 0.003979 **
## Social.support       3.496327   0.659076   5.305 4.85e-07 ***
## Perc.Corruption     -0.783320   0.253616  -3.089 0.002471 **
## Positive.affect      2.430519   0.446065   5.449 2.53e-07 ***
## Negative.affect      1.958639   0.716478   2.734 0.007157 **
## Unemployment        -0.042959   0.008824  -4.868 3.28e-06 ***
## HealthExpOldpop      0.046786   0.013248   3.531 0.000577 ***
## S.E.Asia.Pacific1   -0.469933   0.132275  -3.553 0.000536 ***
## Africa.Sub.Sahara1 -0.299884   0.147575  -2.032 0.044232 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4892 on 127 degrees of freedom
## Multiple R-squared:  0.8226, Adjusted R-squared:   0.81
## F-statistic: 65.43 on 9 and 127 DF,  p-value: < 2.2e-16
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```



The interpretation of Beta coefficients varies depending on the units of measurement of the covariates and should be interpreted according to these for the regressors: Social support, perceptions of corruption, positive affect, and negative affect. An increase of 1 point corresponds to an increase/decrease in the happiness score equal to Beta. However, an increase of 1 point is unlikely for the average values of dummy variables. To check, for example, the increase in happiness score when one of these indices increases by 0.01, it is sufficient to multiply Beta by 0.01.

After this premise, we can analyze each of the variables in the regression:

Intercept (-0-851519): This is the predicted happiness score when all independent variables are equal to zero. However, this value may not be meaningful in isolation because not all predictors can realistically take zero values (e.g., GDP per capita, Unemplyment or Health Expenditure). Given its p-value it is not significantly different from zero, this suggests that the intercept itself might not be meaningful.

Log.GDP.per.capita (0.224873, p-value =0.003979 **): The log of GDP per capita, as expected has a clear positive effect on happiness. For every unit increase in logGDP per capita, the happiness score is predicted to increase by about 0.224873 units.

Social.support (0.046786 , p-value = 0.000577 ***): Social support has a strong positive effect on happiness.

For each unit increase in social support, happiness is predicted to increase by 2.6558 units, with a very low p-value indicating statistical significance.

Perceptions.of.corruption (-0.783320, p-value 0.002471 **): Perceptions of corruption has a negative impact on happiness. For a unit augment of Perceptions of corruption the happiness score will decrease by 0.7833.

Positive.affect (2.430519, p-value 2.53e-07 ***): Positive affect has a positive and statistically significant relationship with happiness. An increase of 1 point in positive emotions is associated with an increase in happiness by 1.4461 units.

Negative.affect (1.958639, 0.007157 **): Negative affect is lesser significant, with an increase of 1 in negative emotions being associated with an increase in happiness by 1.958 units. An increase of 0.01 in negative affect will increase the happiness score by 0.01958 units. This result is unexpected and it can be due from correlation with other explanatory variables, despite the VIF analysis not indicating problematic multicollinearity. The only notable correlation, as observed in the scatterplots, appears to be with social support.

Unemployment (-0.0429, 3.28e-06 ***): Unemployment has a negative effect of 0.0429 meaning that a 1 percentage point increase in the unemployment rate will decrease the happiness score by 0.0429 units. The p-value (0.0380) indicates that this effect is statistically significant.

Africa sub- Sahara (-0.299884, p-value 0.044232 *) Even if it has a small p-value this variable is still slightly significant, it has a negative effect on the happiness score with coef -0.299 this means that if the Country is part of Sub-Saharan Africa we expect an happiness score decreased by 0.299884 unit with other variables fixed.

S.E.Asia.Pacific (-0.46993, 0.000536 ***) This variable has a negative effect on the happiness score with coef -0.4699, this means that if the Country is part of South and East Asia and Pacific area we expect an happiness score decreased by 0.469933 unit with respect to countries in the rest of the world.

Health.ExpenditureOld_Pop (0.046786, 0.000577 ***): Health expenditure has a significant positive coefficient of 0.04678 in the world's nations that have an eldest population. The low p-value suggests that this effect is significant, indicating that a 1% percentage of health expenditure in GDP is associated with 0.046786 increase in happiness for countries with older population.

The confidence intervals (shaded areas) are relatively narrow for most predictors, reflecting stable estimates. However, wider intervals appear for some categorical variables (e.g., Africa Sub-Sahara, S.E. Asia Pacific) due to group variability

Testing a group of regressors

The reduced model includes Log GDP per capita, Social support, and Unemployment as these are well-established predictors of subjective well-being in the literature. They capture key economic, social, and labor market dimensions strongly associated with happiness. The selection balances explanatory power and model simplicity

The F-test ($F = 11.12$, $p < 0.001$) shows that the full model significantly outperforms the reduced model. The inclusion of additional covariates leads to a relevant improvement in model fit. Therefore, we reject the null hypothesis and confirm the importance of the additional variables

```
anova(finalreg, altreg)
```

```
## Analysis of Variance Table
##
## Model 1: Happiness.Score ~ Log.GDP.per.capita + Social.support + Perc.Corruption +
##     Positive.affect + Negative.affect + Unemployment + HealthExpOldpop +
##     S.E.Asia.Pacific + Africa.Sub.Sahara
## Model 2: Happiness.Score ~ Log.GDP.per.capita + Social.support + Unemployment
```

```
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    127 30.397
## 2    133 46.372 -6   -15.975 11.124 5.992e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Goodness of Fit Evaluation

To assess how well the model fits the data, we can rely on R-squared, Adjusted R-squared and the Residual Standard Error (RSE).

```
## [1] 0.8226009
```

$R^2 = 0.8226$ indicate that approximately 82% of the variance in the Happiness Score is explained by the model. So the model explain in a proper way the variance of the response variable.

Prediction

```
new_data = data.frame(Log.GDP.per.capita = 10.5, Social.support = 0.85,
                      Perc.Corruption = 0.75, Positive.affect = 0.7,
                      Negative.affect = 0.3, Unemployment = 6, Africa.Sub.Sahara = "0",
                      HealthExpOldpop = 4.9 ,S.E.Asia.Pacific = "1")
predict(finalreg, newdata = new_data, interval = "prediction")
```

```
##        fit      lwr     upr
## 1 5.684553 4.681616 6.68749
```

Fake data simulation

Considering the multiple linear regression model fitted at the previous point we can simulate n data points from the fitted model, assuming the estimated parameters as the true parameters.We provide a scatterplot of the simulated response vs the observed response and comment it.