Measuring Joy Before the Storm: A Regression Analysis of Pre-COVID Happiness

Pablo Hernández Fernández
Luis Albertos Serrano
Alberto Marín Redondo
Miguel Rodríguez Losada

May 28, 2025

Contents

1 Introduction

In recent years, subjective well-being and happiness have become central topics in social and economic policy discussions. Quantifying happiness allows researchers and policymakers to evaluate the quality of life across countries beyond traditional economic indicators such as GDP per capita.

This analysis aims to identify and quantify the key drivers of national happiness using the 2019 data from the *World Happiness Report*. We apply multiple linear regression techniques to model the happiness score of each country as a function of social, economic, and health-related variables.

One of the aspects that attracted us to this dataset is the opportunity to study what influenced national happiness immediately before a historical global event: the COVID-19 pandemic. As 2019 represents the last "normal" year before the crisis, this analysis offers a valuable baseline for understanding how countries differed in perceived well-being and which factors contributed most to it.

Our main research question is:

> *Can we predict the happiness score of a country in 2019 using measurable national indicators, and which of these indicators are the most influential?*

2 Dataset

The dataset used in this report is obtained from Kaggle, under the World Happiness dataset published by the United Nations Sustainable Development Solutions Network (UN SDSN) [1]. It contains data for 156 countries and includes several quantitative indicators believed to be associated with national happiness.

Below we load the dataset and inspect the variable structure:

```
data <- read.csv("2019.csv", header=TRUE, sep=",", stringsAsFactors=TRUE)

for (i in seq_along(data)) {
  var_name <- names(data)[i]
  var_type <- class(data[[i]])[1]
  cat(paste0("- ", var_name, ": ", var_type, "\n"))
}
```

```
## - Overall.rank: integer
## - Country.or.region: factor
## - Score: numeric
## - GDP.per.capita: numeric
## - Social.support: numeric
## - Healthy.life.expectancy: numeric
## - Freedom.to.make.life.choices: numeric
## - Generosity: numeric
## - Perceptions.of.corruption: numeric
```

[1] https://www.kaggle.com/datasets/unsdsn/world-happiness?select=2019.csv

Note: Instead of using the `str()` *function, we manually display the names and types of the variables below. This decision was made to avoid compilation errors associated with long or complex R output in the* `Sweave` + `pdflatex` *pipeline.*

We now provide a detailed explanation of the variables considered in this analysis.

- **Score**: The average life evaluation score based on the Cantril ladder question, used as the response variable. According to the report's methodology, this score is approximated by the sum of six major factors, each derived from a regression model, plus a residual component:

$$\begin{aligned}\texttt{Score} \approx\ & \texttt{GDP.per.capita} + \texttt{Social.support} + \texttt{Healthy.life.expectancy} \\ & + \texttt{Freedom.to.make.life.choices} + \texttt{Generosity} + \texttt{Perceptions.of.corruption} + \varepsilon\end{aligned}$$

- **GDP per capita**: Logarithmic estimate of income per person.
- **Social support**: Perceived availability of someone to rely on in times of trouble.
- **Healthy life expectancy**: Life expectancy at birth, adjusted for health conditions.
- **Freedom to make life choices**: Measure of perceived freedom in life decisions.
- **Generosity**: Average donations relative to GDP, based on survey responses.
- **Perceptions of corruption**: Trust in public institutions, based on perceived corruption levels in government and business.

No missing values were found in the dataset. All predictor variables are numeric and well-suited for regression analysis.

3 The Model

To understand the factors influencing happiness across countries, we aim to:

- Construct a statistical model to explain the variation in national happiness based on socioeconomic and perception-based indicators.
- Estimate this model using the World Happiness Report 2019 data.
- Identify which predictors have the most substantial impact on the happiness score.

To achieve these objectives, we propose a multiple linear regression model of the form:

$$Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ik}\beta_k + \varepsilon_i, \quad i = 1, \ldots, n$$

where $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are independent random errors, and each x_{ij} denotes the value of the j-th predictor for the i-th observation.

In our case, the response variable Y is the happiness score (`Score`), and the explanatory variables are the following $k = 6$ features from the dataset:

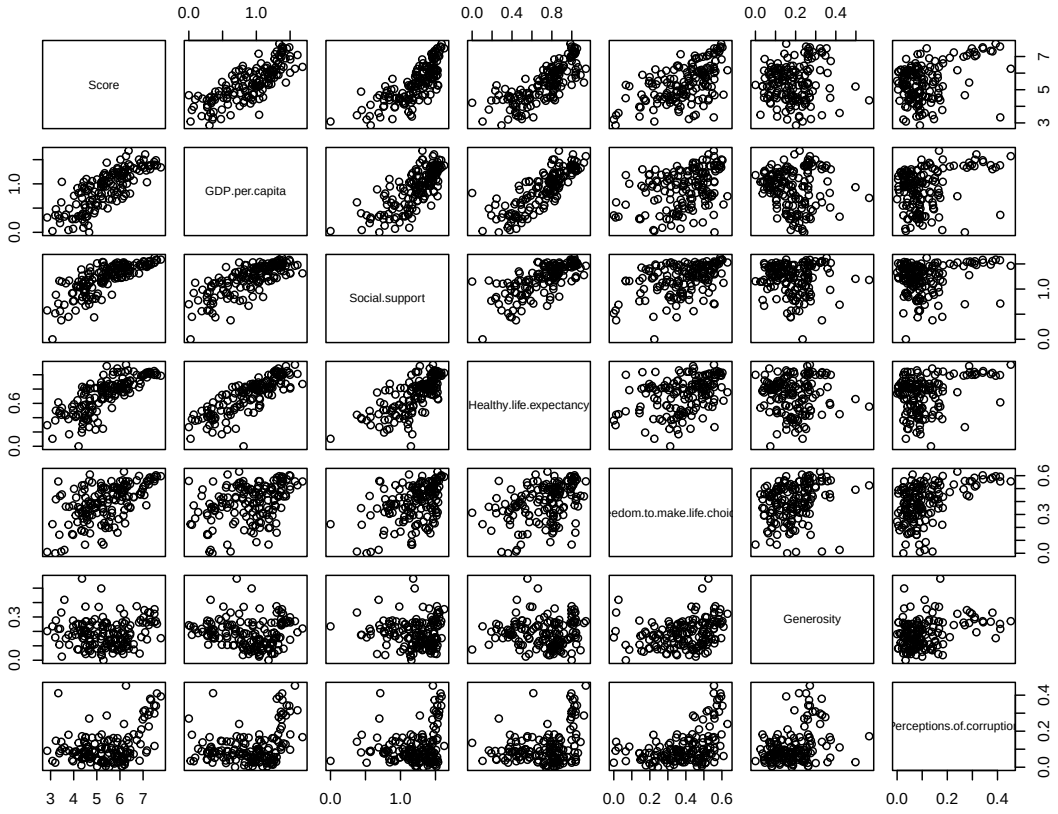
```
GDP.per.capita
Social.support
Healthy.life.expectancy
Freedom.to.make.life.choices
Generosity
Perceptions.of.corruption
```

This model allows us to estimate the relative contribution of each of these variables to the reported happiness levels and to identify those with the greatest explanatory power.

3.1 Exploratory Visualization of the Data

Before fitting the model, it is essential to visually examine the pairwise relationships between the response and explanatory variables. This helps to validate assumptions of linearity and identify potential multicollinearity.

We use a scatterplot matrix to display the relationships among the response variable (`Score`) and the six predictors under consideration:



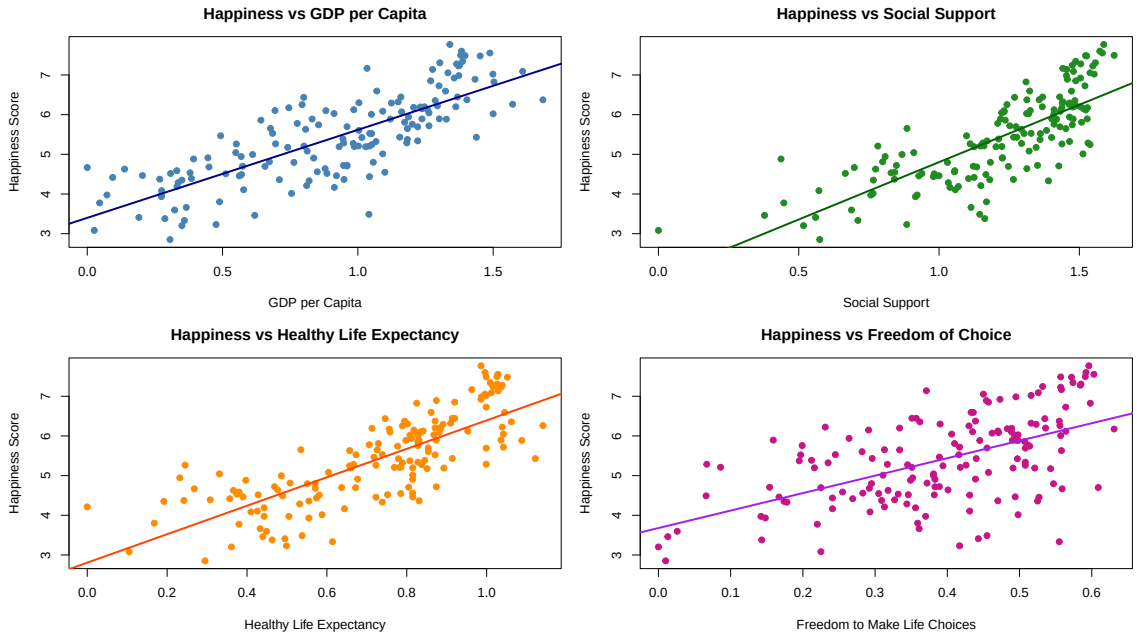
From the scatterplot matrix, we can draw the following initial insights:

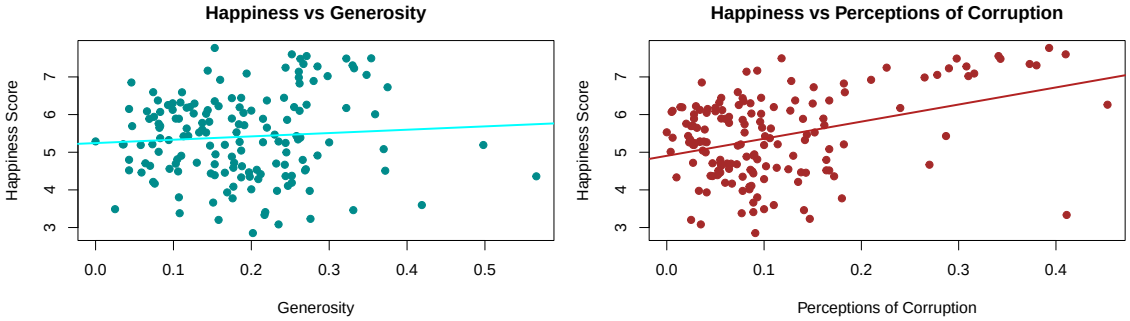
- There is a strong positive linear association between the response variable `Score` and several predictors, especially `GDP.per.capita`, `Social.support`, and `Healthy.life.expectancy`. This visual pattern supports the use of a linear regression framework.
- Clear collinearity is observed between certain predictors. Notably, `GDP.per.capita` and `Healthy.life.expectancy` display a near-linear relationship, as do `Social.support` and `Freedom.to.make.life.choices`. This redundancy may compromise the stability of coefficient estimates and will be addressed through multicollinearity diagnostics.
- Some predictors, such as `Generosity` and `Perceptions.of.corruption`, exhibit weaker or less structured relationships with `Score`. Their explanatory power may therefore be limited, though this will be formally assessed during model fitting.
- No major outliers or non-linear trends are evident at first glance, suggesting the data meet initial expectations for linear regression. However, residual analysis will be conducted to validate these assumptions.

4 Statistical Analysis

We now proceed to estimate simple and multiple linear regression models in order to statistically explain the variability in the national happiness score (`Score`) using six explanatory variables derived from the dataset.

As a preliminary step, we explore the empirical associations between the response and each covariate through pairwise scatterplots, which serve to assess potential linear relationships and justify the use of a linear modeling framework.





We begin our statistical analysis by fitting a series of simple linear regression models, each involving the response variable `Score` and a single explanatory variable.

This approach allows us to individually assess the marginal effect of each predictor on national happiness without accounting for potential multicollinearity or interactions with other covariates. These models serve as a baseline for understanding the strength and direction of each variable's relationship with the response before proceeding to a full multivariate analysis.

```
lm1 <- lm(Score ~ GDP.per.capita, data = data)
lm2 <- lm(Score ~ Social.support, data = data)
lm3 <- lm(Score ~ Healthy.life.expectancy, data = data)
lm4 <- lm(Score ~ Freedom.to.make.life.choices, data = data)
lm5 <- lm(Score ~ Generosity, data = data)
lm6 <- lm(Score ~ Perceptions.of.corruption, data = data)
```

For each fitted model, we examine the estimated regression coefficients, their associated standard errors, and the corresponding p-values to assess statistical significance. This is carried out using the `summary()` function in `R`, which also reports the R^2 and adjusted R^2 values. These metrics allow us to evaluate the proportion of variance in the happiness score explained by each predictor and to identify which covariates exhibit strong linear associations with the response variable when considered in isolation.

- **GDP per capita**: Shows the highest explanatory power among all variables. The model yields an R^2 of 0.6303, and we have no evidence to reject a linear relation ($p < 2 \times 10^{-16}$). On average, an increase of one unit in GDP per capita is associated with a 2.2181-point increase in happiness score.
- **Social support**: Also exhibits a strong positive association with happiness, with an R^2 of 0.6038 we have no evidence to reject a linear relation ($p < 2 \times 10^{-16}$). The estimated effect is approximately 2.8910 units.
- **Healthy life expectancy**: Performs similarly well, with an R^2 of 0.6082. Also there is no evidence to reject a linear relation ($p < 2.2 \times 10^{-16}$). The happiness score increases by 3.5854 units per unit increase in life expectancy.
- **Freedom to make life choices**: Although we have no evidence to reject ($p < 2 \times 10^{-14}$), this variable shows a noticeably lower explanatory capacity ($R^2 = 0.3212$), indicating a weaker individual influence.

- **Generosity**: Displays negligible explanatory power ($R^2 = 0.0057$) and is not statistically significant ($p = 0.347$), so we can reject a linear relation between perceived generosity of a country and its hapiness score.
- **Perceptions of corruption**: While can't reject a linear relation between the variables ($p < 10^{-6}$), its explanatory power is limited ($R^2 = 0.1487$), indicating that trust in public institutions has a moderate role in explaining national happiness.

Overall, economic and health-related indicators are the most informative individual predictors of national happiness, while subjective measures such as generosity appear to contribute less in isolation.

Having examined the individual effect of each predictor through simple linear regression, we now fit a multiple linear regression model that includes all six explanatory variables simultaneously.

This comprehensive model allows us to evaluate the joint effect of the covariates on the happiness score (`Score`), while accounting for potential correlations among them. The estimated coefficients in this setting represent the marginal effect of each variable on the response, adjusted for the presence of the others. This model serves as the foundation for subsequent diagnostic analysis and potential variable selection procedures.

Specifically, the corresponding code chunk is presented below, detailing the procedure used to carry out this step of the analysis.

```r
# Multiple regression with all covariates
lm_full <- lm(Score ~ GDP.per.capita + Social.support +
                Healthy.life.expectancy +
              Freedom.to.make.life.choices + Generosity +
                Perceptions.of.corruption,
              data = data)
```

```r
summary(lm_full)
```

```
##
## Call:
## lm(formula = Score ~ GDP.per.capita + Social.support + Healthy.life.expectancy +
##     Freedom.to.make.life.choices + Generosity + Perceptions.of.corruption,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75304 -0.35306  0.05703  0.36695  1.19059
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.7952     0.2111   8.505 1.77e-14 ***
## GDP.per.capita                  0.7754     0.2182   3.553 0.000510 ***
```

```
## Social.support                  1.1242     0.2369   4.745 4.83e-06 ***
## Healthy.life.expectancy         1.0781     0.3345   3.223 0.001560 **
## Freedom.to.make.life.choices    1.4548     0.3753   3.876 0.000159 ***
## Generosity                      0.4898     0.4977   0.984 0.326709
## Perceptions.of.corruption       0.9723     0.5424   1.793 0.075053 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5335 on 149 degrees of freedom
## Multiple R-squared:  0.7792, Adjusted R-squared:  0.7703
## F-statistic: 87.62 on 6 and 149 DF,  p-value: < 2.2e-16
```

We are able to confirm that some variables, such as `Generosity` and `Perceptions.of.corruption`, may not be significant in the full model. To refine our model and potentially improve its interpretability and predictive performance, we apply a stepwise variable selection procedure. This technique iteratively adds or removes covariates based on the Akaike Information Criterion (AIC), aiming to retain only those predictors that provide substantial explanatory power.

```
lm_step <- step(lm_full, direction = "both")

## Start:  AIC=-189.18
## Score ~ GDP.per.capita + Social.support + Healthy.life.expectancy +
##     Freedom.to.make.life.choices + Generosity + Perceptions.of.corruption
##
##                                Df Sum of Sq    RSS     AIC
## - Generosity                    1    0.2756 42.687 -190.17
## <none>                                      42.412 -189.18
## - Perceptions.of.corruption     1    0.9148 43.326 -187.85
## - Healthy.life.expectancy       1    2.9564 45.368 -180.67
## - GDP.per.capita                1    3.5934 46.005 -178.49
## - Freedom.to.make.life.choices  1    4.2764 46.688 -176.19
## - Social.support                1    6.4099 48.822 -169.22
##
## Step:  AIC=-190.17
## Score ~ GDP.per.capita + Social.support + Healthy.life.expectancy +
##     Freedom.to.make.life.choices + Perceptions.of.corruption
##
##                                Df Sum of Sq    RSS     AIC
## <none>                                      42.687 -190.17
## + Generosity                    1    0.2756 42.412 -189.18
## - Perceptions.of.corruption     1    1.3053 43.993 -187.47
## - Healthy.life.expectancy       1    2.9896 45.677 -181.61
## - GDP.per.capita                1    3.3872 46.075 -180.26
## - Freedom.to.make.life.choices  1    4.9836 47.671 -174.94
## - Social.support                1    6.3443 49.032 -170.55
```

The algorithm starts by removing `Generosity` (the least significant for the model). Once done, it tries again, but all the remaining covariates are significant enough, so the only one removed is `Generosity`.

Below we present the summary of the final model selected through this process:

```
summary(lm_step)

##
## Call:
## lm(formula = Score ~ GDP.per.capita + Social.support + Healthy.life.expectancy +
##     Freedom.to.make.life.choices + Perceptions.of.corruption,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.82997 -0.35344  0.05803  0.35977  1.17522
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     1.8689     0.1973   9.471  < 2e-16 ***
## GDP.per.capita                  0.7455     0.2161   3.450 0.000728 ***
## Social.support                  1.1180     0.2368   4.722 5.33e-06 ***
## Healthy.life.expectancy         1.0840     0.3344   3.241 0.001467 **
## Freedom.to.make.life.choices    1.5340     0.3666   4.185 4.84e-05 ***
## Perceptions.of.corruption       1.1176     0.5218   2.142 0.033839 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5335 on 150 degrees of freedom
## Multiple R-squared:  0.7777, Adjusted R-squared:  0.7703
## F-statistic:   105 on 5 and 150 DF,  p-value: < 2.2e-16
```

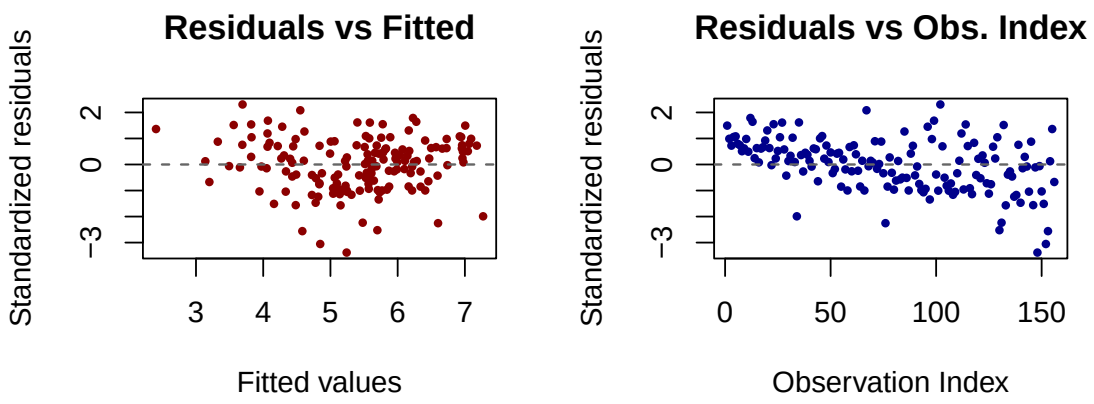
4.1 Regression Diagnostics

Regression diagnostics allow us to check whether the assumptions listed in Section 3 hold in our fitted model. These include linearity, independence, homoscedasticity (constant variance), and normality of residuals, as well as detecting potential outliers or influential points.

4.1.1 Constant Variance and Independence of Residuals

We first examine the assumption of homoscedasticity (constant variance of the residuals) by plotting the standardized residuals against the fitted values. A random spread of points around zero suggests that the variance is constant across levels of the fitted values.

To check for independence of residuals, we plot them against the observation index. We also apply the Durbin-Watson test to statistically assess autocorrelation, with the null hypothesis being that residuals are uncorrelated.



We apply the Durbin-Watson test for autocorrelation:

```
library(lmtest)
dwtest(lm_full)

##
##  Durbin-Watson test
##
## data:  lm_full
## DW = 1.6484, p-value = 0.01097
## alternative hypothesis: true autocorrelation is greater than 0
```

We can draw the following conclusions:

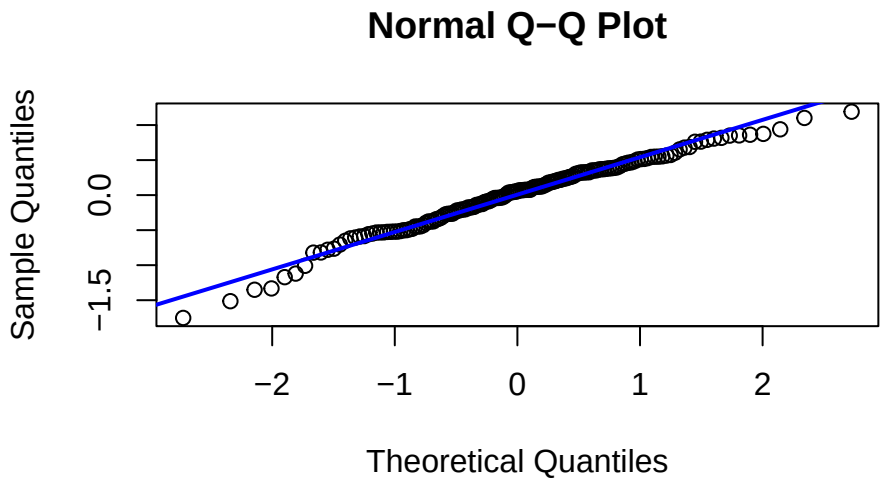
- The **Residuals vs Fitted Values** plot shows no clear funnel shape or systematic pattern, suggesting that the assumption of *homoscedasticity* (constant variance) is reasonably satisfied. The residuals appear to be randomly scattered around zero across all fitted values.
- The **Residuals vs Observation Index** plot reveals a slight trend and possible clustering, which may indicate minor dependence between residuals. This visual impression motivates formal testing.
- The **Durbin-Watson test** returns a test statistic of approximately 1.6484 with a p-value of 0.011.

Note: Although the Durbin-Watson test indicates some positive autocorrelation in residuals (DW = 1.6484, p = 0.011), we note that the dataset is cross-sectional in nature, and countries are likely ordered by overall rank rather than time or space. As such, residual dependence may result from unmeasured regional similarities. Given that other assumptions (homoscedasticity, linearity, normality) are met and the residuals are relatively stable, we proceed with the model, while acknowledging this limitation and suggesting that future studies incorporate spatial or multilevel modeling techniques.

4.1.2 Normality of Residuals

To assess whether residuals follow a normal distribution, we use a Q-Q plot and apply the Kolmogorov-Smirnov test. A linear pattern in the Q-Q plot and a large p-value from the test support the assumption of normality.

```
qqnorm(resid(lm_full))
qqline(resid(lm_full), col = "blue", lwd = 2)
```



We test normality using the Kolmogorov-Smirnov test:

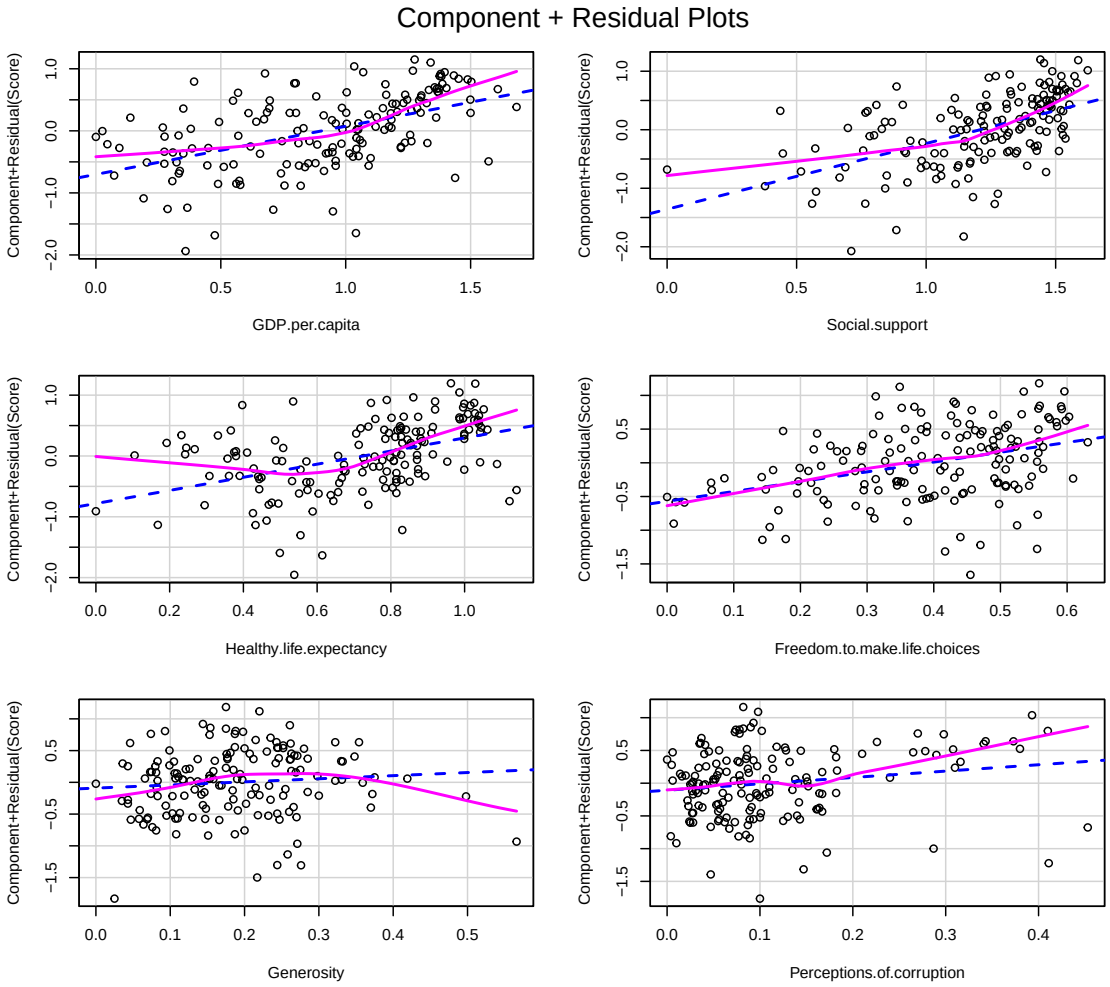
```
ks.test(resid(lm_full), "pnorm", mean = mean(resid(lm_full)),
        sd = sd(resid(lm_full)))

##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  resid(lm_full)
## D = 0.058212, p-value = 0.6658
## alternative hypothesis: two-sided
```

The Q-Q plot displays a largely linear pattern, with only minor deviations in the tails, indicating that the residuals follow an approximately normal distribution. This impression is confirmed by the Kolmogorov-Smirnov test, which returns a test statistic $D = 0.0582$ and a p-value of 0.6658. Since this p-value far exceeds the conventional 5% significance level, we find no statistical evidence to reject the null hypothesis of normality. Both the visual and formal diagnostics thus suggest that the normality assumption is + adequately met for the residuals of the fitted model.

4.1.3 Linearity Check with Component+Residual Plots

To assess the linearity assumption, we use component-plus-residual (partial residual) plots. These plots help visualize whether the relationship between the response and each covariate is approximately linear, controlling for the effect of other variables.

```
library(car)
crPlots(lm_full)
```



Based on the component-plus-residual plots, the linearity assumption appears reasonably satisfied for all predictors. The relationships are approximately linear, especially for `GDP.per.capita`, `Social.support`, and `Freedom.to.make.life.choices`. Although some slight nonlinearity is visible in `Generosity` and `Healthy.life.expectancy`, the deviations are not severe enough to undermine the validity of a linear regression approach.

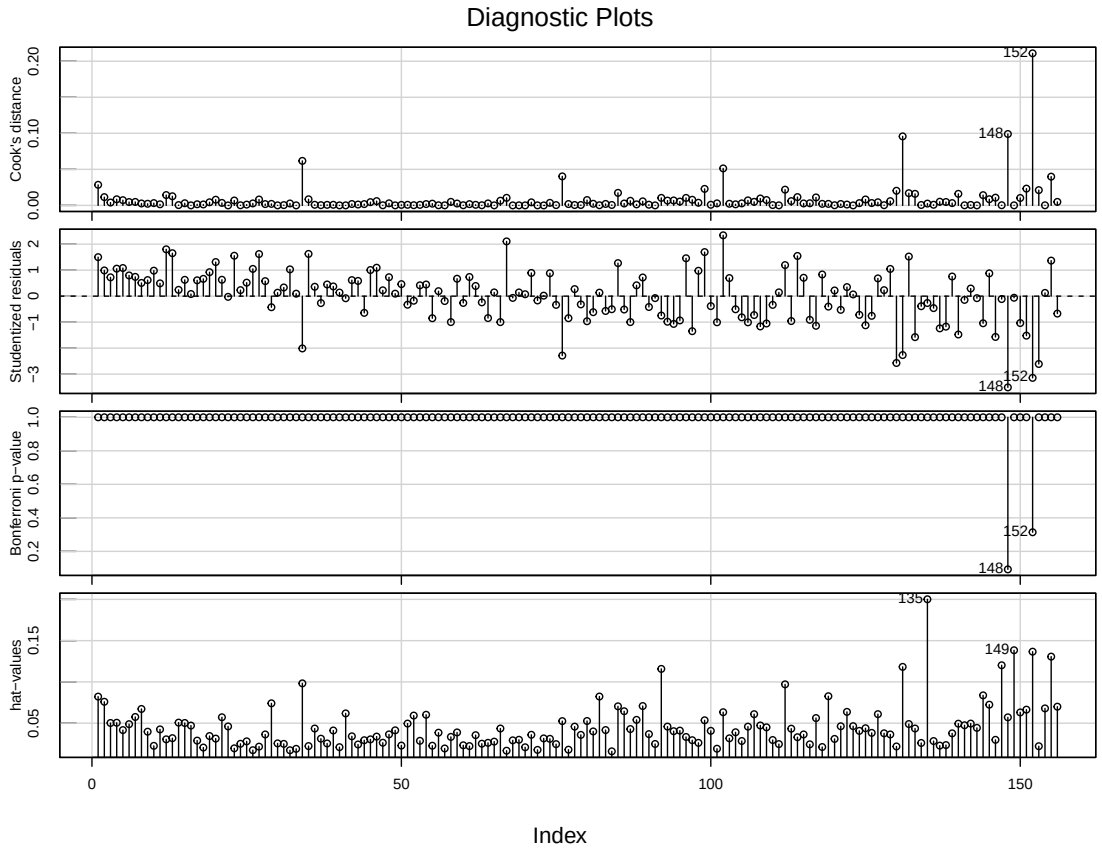
4.1.4 Outliers and Influential Observations

We identify potential outliers using standardized residuals greater than ± 2.5. In addition, we use Cook's distance and leverage measures (via the influence index plot) to detect influential or high-leverage observations.

```
outliers <- which(abs(resid_std) > 2.5)
outliers

## 130 148 152 153
## 130 148 152 153
```

We now visualize the influence measures for all observations:

```
influenceIndexPlot(lm_full)
```



4.1.5 Refining the Model: Removing Unusual Observations

Outliers and influential observations can distort the accuracy and stability of regression estimates. These atypical data points might stem from measurement inconsistencies, data entry errors, or genuine but exceptional cases that deviate from general trends. Proper identification and handling of these points is essential for obtaining robust and interpretable models.

In this section, we applied an iterative outlier removal strategy using standardized residuals obtained from the `rstandard()` function in textttR. The model used at each iteration was a multiple linear regression fitted via `lm()`, with the following six predictors:

```
GDP.per.capita   Social.support   Healthy.life.expectancy
Freedom.to.make.life.choices   Generosity   Perceptions.of.corruption
```

For each fitted model, we computed standardized residuals with `rstandard()` and identified outliers as those with absolute values greater than 2.5. We removed the most extreme observation and refitted the model using `lm()` on the reduced dataset. We repeated this procedure until all standardized residuals were within the acceptable range.

Each iteration was evaluated using `summary()`, tracking the adjusted R^2, residual standard error, and statistical significance of covariates. A particular focus was placed on the variable `Generosity`, whose coefficient repeatedly failed to reach significance.

Removed Observation	**Adjusted R^2**	**`Generosity` Significance**
152	0.7796	Not Significant
148	0.7938	Not Significant
130	0.8023	Not Significant
76	0.8113	Not Significant
34	0.8199	Not Significant
128	0.8268	Not Significant
147	0.8352	Model Stabilized

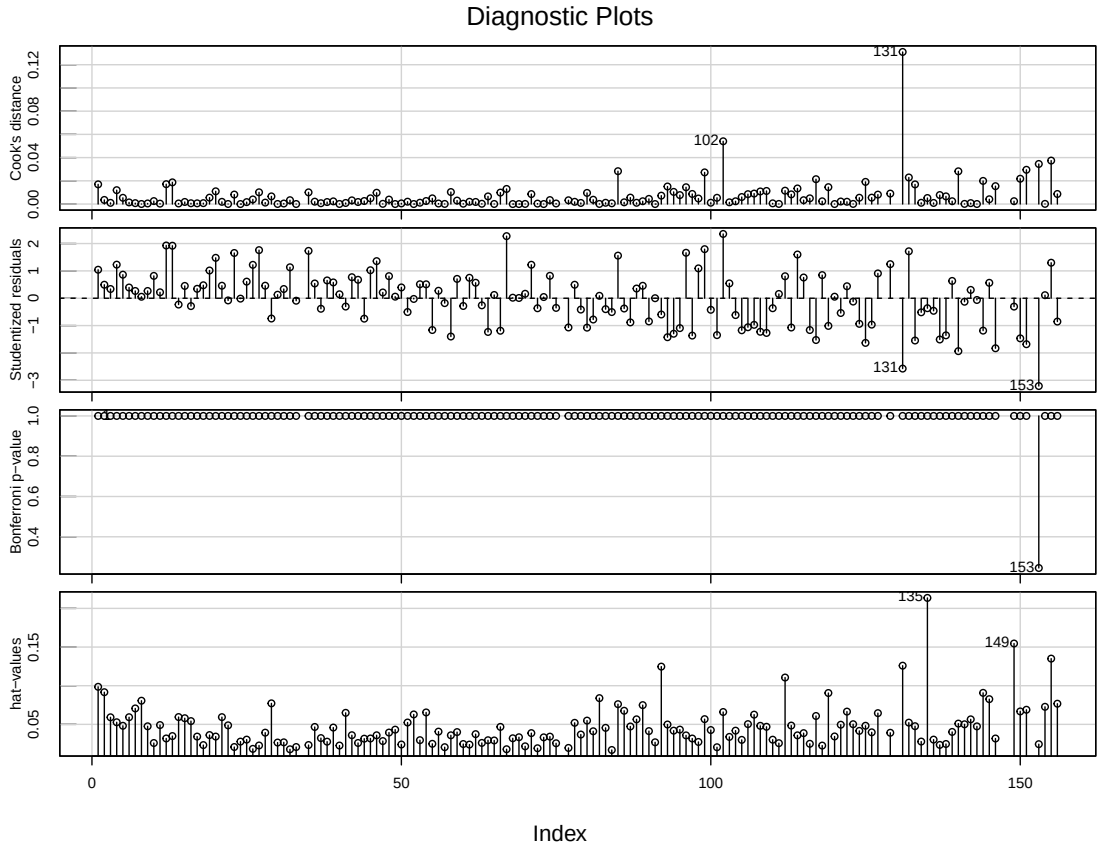
Table 1: Summary of the Iterative Model Refinement Process

The final model, fitted after excluding observations 152, 148, 130, 76, 34, 128, and 147, achieved the following:

- $R^2_{adj} = 0.8352$: about 83.5% of the variability in `Score` is explained.
- Residual standard error = 0.4427.
- F-statistic = 126 (df = 6, 142), $p < 2.2 \times 10^{-16}$.

Notably, `Generosity` was not statistically significant in any iteration, suggesting limited explanatory value for this response.

The influence diagnostics of the final model were visually inspected using the `influenceIndexPlot()` function from the `car` package:



This plot confirms that no new high-leverage or extreme residual values remain. The model can therefore be considered a stable and trustworthy basis for inference in the study of national happiness determinants.

5 Conclusions

This report has explored the key national factors associated with happiness levels in the 2019 *World Happiness Report*, applying multiple linear regression to model the variable `Score` across 156 countries.

The primary objective was to uncover which socio-economic and perception-based indicators best explain national well-being. This was achieved through a sequence of exploratory analysis, regression modeling, assumption verification, and outlier refinement.

Main empirical findings:

- The strongest positive associations in the initial regressions were found for `GDP.per.capita`, `Social.support`, and `Healthy.life.expectancy`.
- In the full model built using the `lm()` function in `R`, all variables remained significant except `Generosity`, which did not provide sufficient explanatory power.
- Regression diagnostics confirmed that linearity, constant variance, and normality of residuals held reasonably well. Slight autocorrelation was + detected through the Durbin-Watson test, though considered manageable given the cross-sectional nature of the data.

Model refinement and robustness:

- Outliers were identified and removed based on standardized residuals and influence metrics. This step, implemented via `rstandard()` and `influenceIndexPlot()`, enhanced the model's stability.
- The final refined model achieved improved metrics:
  - Adjusted $R^2 = 0.8352$
  - Residual standard error = 0.4427
- Notably, `Generosity` remained statistically non-significant throughout all iterations.

Key insight: Among all variables considered, national income, health outcomes, and social support stood out as the most reliable predictors of happiness in 2019, demonstrating consistent and strong effects across all phases of analysis.

Considerations for future work:

- Linear models are easy to interpret but may miss nonlinearities or complex interactions in social data.
- Regularized methods (e.g., Ridge, LASSO), as well as nonlinear or longitudinal models, could reveal deeper patterns.
- To address the mild autocorrelation detected in Section 4.1.1, future work could consider **cluster-based models** to account for potential regional or cultural groupings among countries.

The results emphasize that fostering economic prosperity, public health, and social connectedness may offer the most effective pathways for improving happiness at the national level.

6 References

- Crawley, M. J. (2007). *The R Book*. Wiley. Available at: The R Book PDF
- Wickham, H. (2023). *R Packages*. Available at: r-pkgs.org