HDI Regression Analysis

Data collection and goal of the analysis

The goal of my analysis is to examine how the Human Development Index (HDI) is influenced by various factors across the majority of countries in 2020 and how it varies in relation to some indicators. For this analysis, I will use two datasets: the first provides data on the HDI calculated for all countries, while the second contains a comprehensive list of global indicators that may be valuable for my analysis. Various variables can impact the HDI, particularly those related to education, health, social, economic and environmental aspects. The HDI is the geometric mean of normalized indices for three dimensions: the health dimension, assessed by life expectancy at birth; the education dimension, measured by the mean years of schooling for adults aged 25 years and older and expected years of schooling for children of school-entering age; and the standard of living dimension, measured by the gross national income per capita. My objective is to understand if other variables that do not directly enter into the HDI calculation can influence it. The HDI is published annually by the United Nations Development Programme (UNDP), while I used the World Bank website for the indicators. My dataset includes 177 countries after removing some countries due to excessive missing values for many variables. For other countries, I used "MissForest", which is a random forest-based imputation algorithm for missing data. Afterward, I examined the density functions before and after the imputation of the missing data and did not observe significant differences. After selecting variables based on their potential relevance for my analysis, I have chosen 10 predictors to start the analysis. Moreover, before starting the analysis, I centered all variables to their mean, except for CO_2 *emissions*. Since I will apply a logarithmic transformation to CO_2 *emissions*, centering would result in negative values, which are not suitable for logarithmic scaling. In conclusion, to better understand the distribution of the HDI across different regions, I have created a qualitative variable called "Geographical Region". The boxplot in Figure 1 summarizes the distribution of HDI scores across five geographical areas.

Exploratory analysis

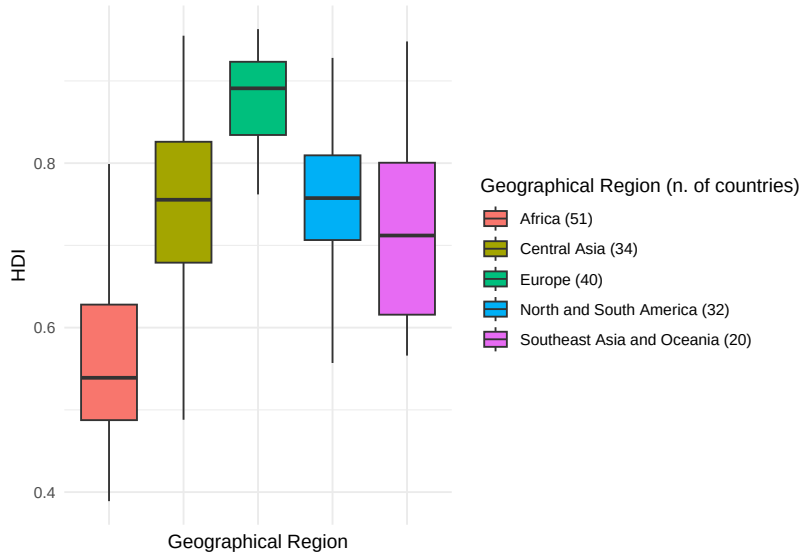

Figure 1: Boxplot for HDI by Geographical Region

From the boxplot, it is evident that Europe exhibits the highest HDI, indicating a better overall human development outcomes compared to other regions. In contrast, Africa displays the lowest HDI, reflecting more significant developmental issues. Central Asia and the Americas show similar distributions with nearly identical median values, while Southeast Asia and Oceania have a slightly lower median HDI than Central Asia and the Americas.

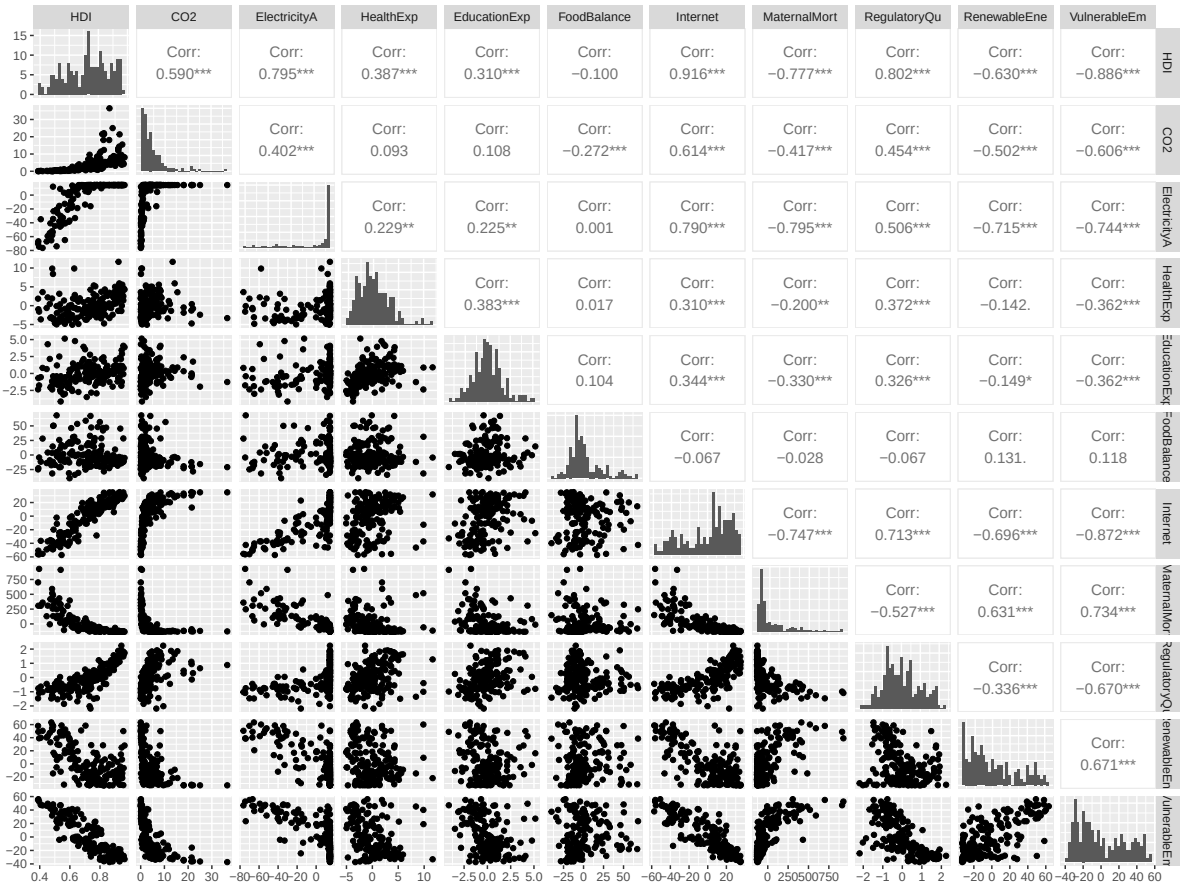

Figure 2: Scatter plot and correlation matrix for all variables

The variable *Electricity Access* (percentage of population with electricity access) is highly concentrated around 100, indicating that in many countries, electricity is universally available. It also shows a strong positive relationship with HDI. *Health Expenditure* and *Education Expenditure* (both as percentages of GDP) exhibit no clear linear relationship with HDI, with *Health Expenditure* also featuring some extreme values. Similarly, *Food Trade Balance* (difference between total food exports and imports) follows a comparable pattern. Conversely, *Individuals Using the Internet* (percentage of the population) displays a strong positive linear relationship with HDI. A similar trend is observed for *Regulatory Quality* (an estimate of quality of laws and regulations), though with greater dispersion between -2 and 0, beyond which the relationship becomes more linear. *Maternal Mortality Ratio* (per 100,000 live births) has a non-linear negative relationship with HDI. Moreover, most countries exhibit very low maternal mortality, as shown in the histogram. *Renewable Energy consumption* (percentage of total final energy use) is negatively correlated with HDI, maybe due to long times and significant investments that renewables need to impact a country's economy. *Vulnerable Employment* (percentage of workers in precarious jobs) shows a negative linear relationship with HDI, reinforcing that a higher share of vulnerable workers is associated with lower HDI. A strong positive correlation (about 0.8) exists between *Electricity Access* and *Individuals Using the Internet*, as stable electricity supply is essential for internet access. *Individuals Using the Internet* is also positively correlated with *Regulatory Quality* (0.713). *Renewable Energy Consumption* and CO_2 *Emissions* show a negative correlation, meaning countries with higher renewable energy use tend to have lower emissions. However, some industrialized nations with high HDI continue to exhibit high emissions despite investing in renewables.

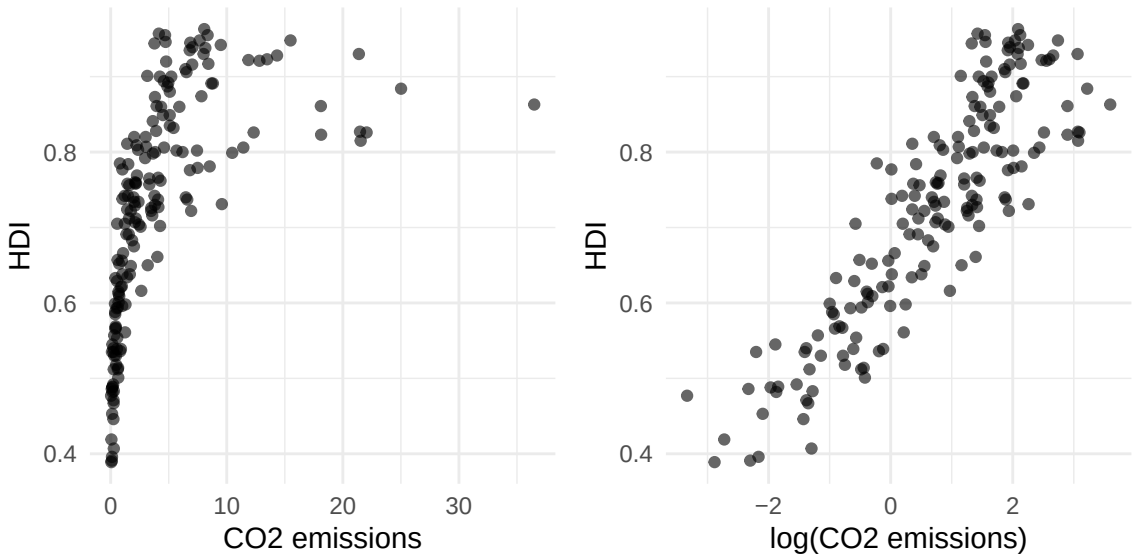

Figure 3: CO_2 emissions before and after the logarithmic transformation

Regarding CO_2 *emissions* (total tons per capita), I applied a logarithmic transformation since the data were highly dispersed. Some countries (such as Qatar or UAE) exhibit extreme values, as evident from the scatter plot on the right. The logarithmic transformation helped to make linear the relationship between HDI and this variable. Furthermore, I attempted to perform variable selection without applying the logarithmic transformation, but the variable is rarely included in the selected models. Since I think that this variable can be useful in explain variation in HDI I decided to apply the transformation before to start the analysis. To conclude, the relationship is positive, indicating that countries with a higher HDI tend to have higher CO_2 emissions (an expected outcome, given that more developed economies typically have greater industrial activity). Before starting the variable selection process, I also transformed my response variable using an inverse CDF transformation. This transformation was applied because the HDI is an index that ranges between 0 and 1, and using a standard linear regression model could lead to predictions outside this interval. The transformation makes the response variable take values on the entire real line (-∞, +∞) instead of being constrained to [0,1], making it more suitable for regression analysis. Moreover, the probit transformation helps to stabilize variance, but also it improves the normality assumption of the error terms (fundamental requirement for linear regression models) as showed in Figure 4.

```
Final_dataset$HDI_probit <- qnorm(Final_dataset$HDI)
```

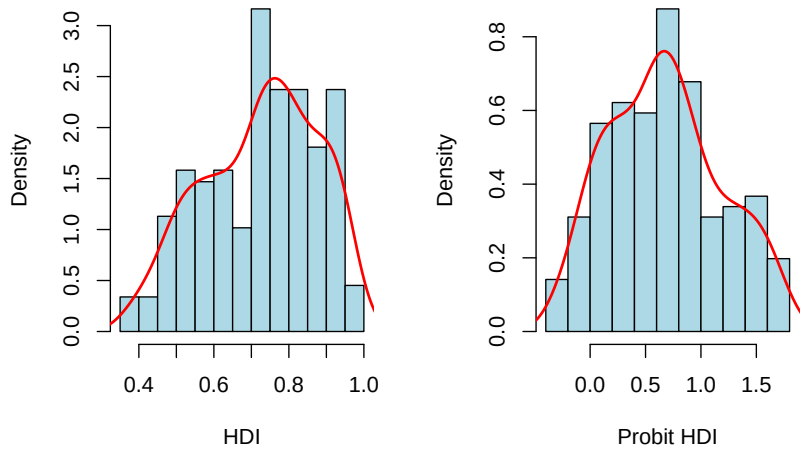

Figure 4: Histogram of HDI before and after probit transformation

Variable selection and model improvement

```
p <- 14
n <- 177
aic_values <- numeric(p)
for (k in 1:p) {
    model <- regsubsets(HDI_probit ~ . -Country -HDI +log(CO2) -CO2,
                        data = Final_dataset, nvmax = p)

    model_summary <- summary(model)
    aic_values[k] <- n * log(model_summary$rss[k]/n) + 2 * (k+2)
}
```

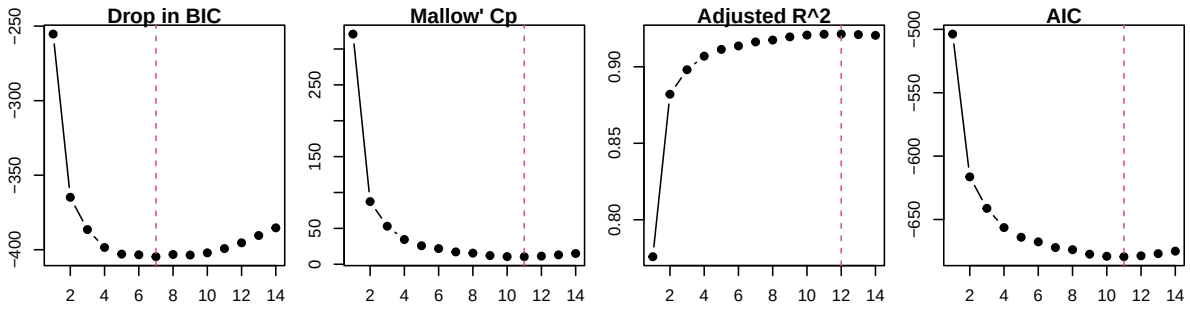

Figure 5: Criteria used for best subset selection

After the best subset selection, as indicated by the results of AIC and C_p statistic the optimal model includes 11 predictors. The BIC suggests a similar model but includes only one level of *Geographical region*, and also incorporates *Education expenditure*. The results of the forward selection are identical to those obtained from best subset selection, while the backward selection yields a very similar model. However, I prefer the model selected through best subset selection, suggested by AIC and C_p, because I want to retain the geographical variable as it account for variability across macro areas, which is valuable to check if the uncorrelated errors assumption is satisfied. Next, I will check for collinearity issues and perform model diagnostics to verify whether all regression assumptions are met. In conclusion, the model I selected after the best subset selection is the following:

```
model_reduced1 <- lm(HDI_probit ~ HealthExp + EducationExp + Internet + RegulatoryQuality
    + RenewableEnergy + VulnerableEmployment + GeoRegion + log(CO2), data = Final_dataset)
```

Table 1: GVIF values

Variable	GVIF	Df	GVIF_adj	Variable	GVIF	Df	GVIF_adj
HealthExp	1.447391	1	1.203076	RenewableEnergy	2.861608	1	1.691629
EducationExp	1.338621	1	1.156988	VulnerableEmployment	5.591491	1	2.364633
Internet	8.502558	1	2.915914	GeoRegion	3.357549	4	1.163464
RegulatoryQuality	2.699715	1	1.643081	log(CO2)	7.106710	1	2.665841

Looking at the GVIFs values there are not collinearity issues in the model since all values are under the threshold of 10.

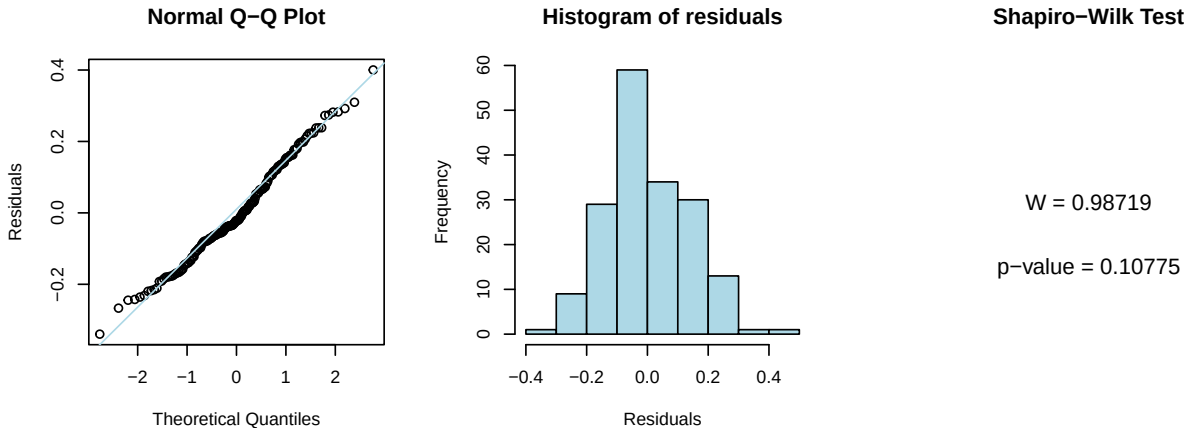

Figure 6: Q-Q plot, Histogram of the residuals and Shapiro-Wilk test

Regarding the normality assumption, the QQ-plot shows that most points align well with the theoretical quantiles, though some slight deviations occur at the tails, while the histogram displays a bell-shaped distribution. Overall, the residuals appear approximately normal, but the p-value of the Shapiro-Wilk test is low; however, since it is greater than 0.05, I can accept the null hypothesis of normality.

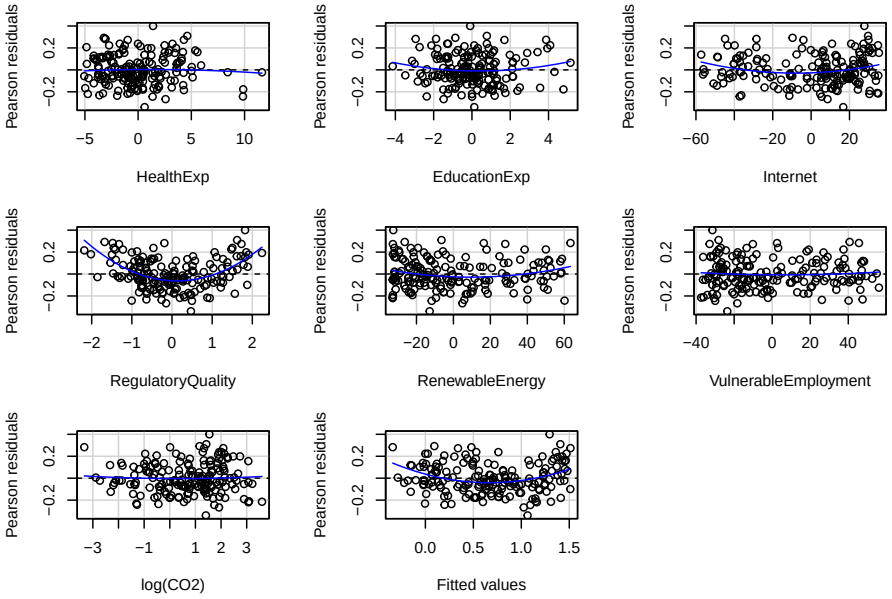

Figure 7: Residual plots

Looking at the residual plots, there is a strong non-linearity issue related to *Regulatory quality*. Other variables, such as *Education expenditure* and *Renewable energy*, also exhibit some degree of non-linearity, but the issue is not as severe. Moreover, the plot of the fitted values highlights a significant linearity problem, likely driven by the presence of *Regulatory quality* in the model. Indeed, with the addition of the quadratic term for this variable, keeping all other variables in the model, the results change significantly, as illustrated in Figure 8.

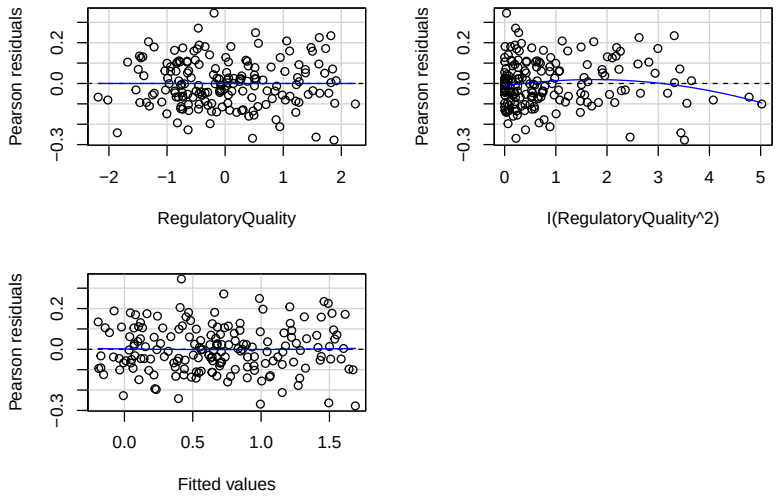

Figure 8: Effect of adding $(RegulatoryQuality)^2$

The fitted values now exhibit linearity, suggesting that the inclusion of this new predictor has improved the model. While the quadratic term shows some linearity issue, it is not severe. Moreover, the model summary confirms that the quadratic term is statistically significant. However, since an additional predictor has been included, it is necessary to repeat the variable selection procedure to determine which variables should be retained or excluded. Additionally, a new diagnostic analysis must be conducted.

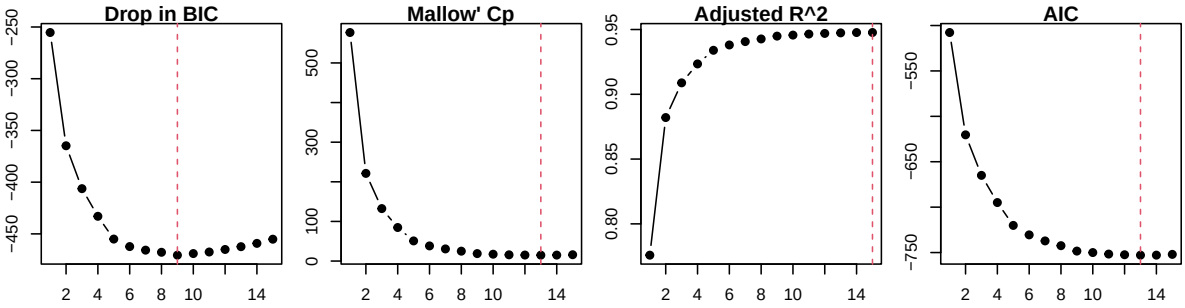

Figure 9: Criteria used for best subset selection of the model after the introduction of $RegulatoryQuality^2$

```r
p <- 15
set.seed (1)
folds <- sample (1:n, nrow(Final_dataset4), replace = FALSE)
cv.errors <- matrix (NA, n, p, dimnames = list(NULL, paste (1:p)))

for (j in 1:n){
  best.fit <- regsubsets(HDI_probit ~ .,  data = Final_dataset4[folds !=j,], nvmax = p)
  for (i in 1:p){
    mat <- model.matrix(as.formula(best.fit$call[[2]]), Final_dataset4[folds==j,])
    coefi <- coef(best.fit, id = i)
    xvars <- names(coefi)
    pred  <- mat[,xvars]%*%coefi
    cv.errors[j,i] <- mean((Final_dataset4$HDI_probit[folds==j]-pred)^2)
  }
}
```

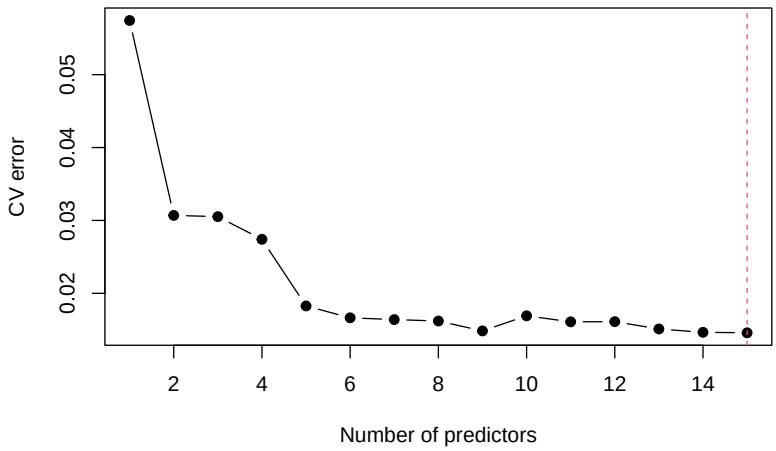

Figure 10: LOOCV plot

Table 2: Selected Variables by AIC, BIC, and Cp

Criterion	Selected_Variables
AIC	Internet / RegulatoryQuality / log(CO2) / EducationExp / ElectricityAccess / HealthExp / MaternalMortalityRatio / RenewableEnergy / VulnerableEmployment / GeoRegionAfrica / I(RegulatoryQuality^2) / GeoRegionCentral Asia / GeoRegionNorth and South America / GeoRegionSoutheast Asia and Oceania
BIC	Internet / RegulatoryQuality / log(CO2) / VulnerableEmployment / GeoRegionAfrica / I(RegulatoryQuality^2) / GeoRegionCentral Asia / GeoRegionNorth and South America / GeoRegionSoutheast Asia and Oceania
Cp	Internet / RegulatoryQuality / log(CO2) / EducationExp / ElectricityAccess / HealthExp / MaternalMortalityRatio / RenewableEnergy / VulnerableEmployment / GeoRegionAfrica / I(RegulatoryQuality^2) / GeoRegionCentral Asia / GeoRegionNorth and South America / GeoRegionSoutheast Asia and Oceania

Looking at the variables selected with all criteria, the quadratic term is included in all the models, confirming its relevance. The results remain largely similar to the previous analysis; however, C_p statistic and AIC now include also *Electricity access* and *Maternal mortality ratio*, even though these variables are not significant. The LOOCV and the adjusted R^2 instead suggest to take the full model. The results remain the same regardless of whether best subset selection, backward selection, or forward selection is used. However, I prefer the smaller model, as suggested by BIC, for two main reasons: a smaller model reduces overall complexity, improving interpretability and avoiding unnecessary predictors; and although AIC and C_p suggest a larger model, the difference between models with 9 or more predictors is minimal according to these two criteria, as observed in Figure 9. Therefore, I have chosen the smaller model (with 9 predictors), which includes the following variables:

```
Final_model <- lm(HDI_probit ~ Internet + RegulatoryQuality + VulnerableEmployment +
 GeoRegion + log(CO2) + I(RegulatoryQuality^2), data = Final_dataset)
summary(Final_model)
```

Checking for collinearity issues

Table 3: GVIF values of the final model

Variable	GVIF	Df	GVIF_adj	Variable	GVIF	Df	GVIF_adj
Internet	8.162598	1	2.857026	GeoRegion	2.811630	4	1.137941
RegulatoryQuality	2.504899	1	1.582687	log(CO2)	6.444320	1	2.538567

VulnerableEmployment	5.193321	1	2.278886	I(RegulatoryQuality^2)	1.129552	1	1.062804

Looking at the GVIFs values there are not collinearity issues in the final model since all values are under the threshold of 10.

Diagnostics

Homoskedasticity assumption

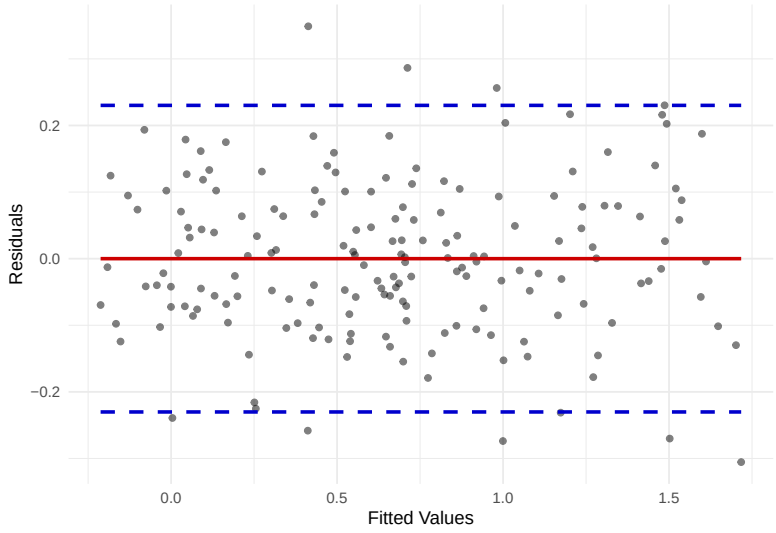

Figure 11: Fitted values vs Residuals plot

Examining the plot of fitted values against residuals, it appears that the assumption of constant variance (homoskedasticity) holds. The residuals are randomly distributed around zero, with no precise pattern.

Linearity assumption

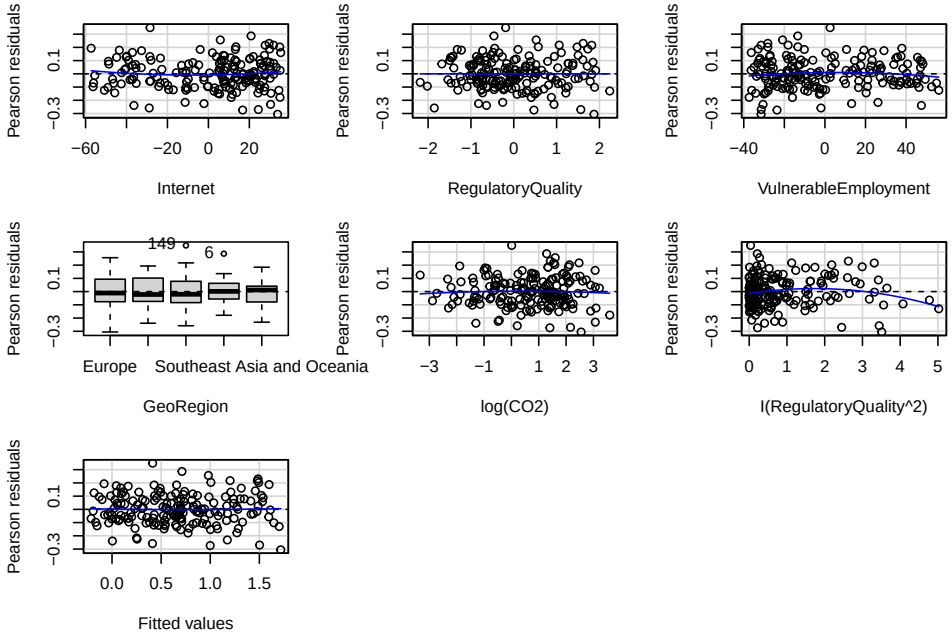

Figure 12: Residual plots of the final model

The plots display the Pearson residuals against all predictors and fitted values to assess the model's linearity assumption. Most predictors exhibit a random scatter of residuals around zero, suggesting that the linear assumption holds well, even if *Regulatoryquality*^{2} shows a slight curvature. *Vulnerable employment* and *Individuals using the internet* display mild patterns, indicating potential non-linearity. The boxplot of residuals across geographical regions suggests that the medians are close to zero and there is similar variance across groups; however, a few extreme residuals (e.g., 149, 6) indicate some observations deviate more than expected.

Normality assumption

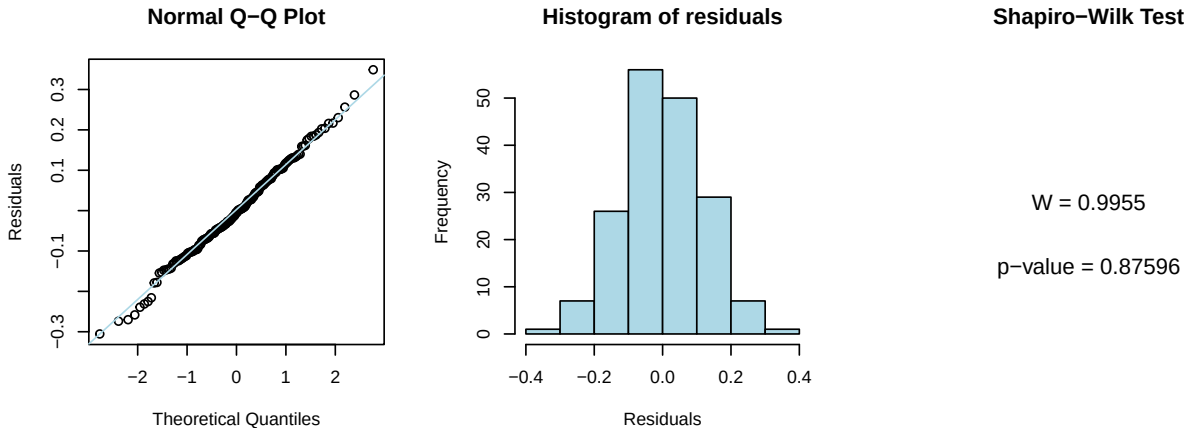

Figure 13: Q-Q plot, Histogrsm of the residuals and Shapiro-Wilk test

In the Q-Q plot most points closely follow the reference line, suggesting that the residuals are approximately normal. However, there are slight deviations in the tails. In the histogram the residuals appear approximately symmetric, suggesting that the assumption of normality is met. The p-value of the Shapiro-Wilk test is very high, indicating that the null hypothesis of normality can be accepted.

Uncorrelation of the errors assumption

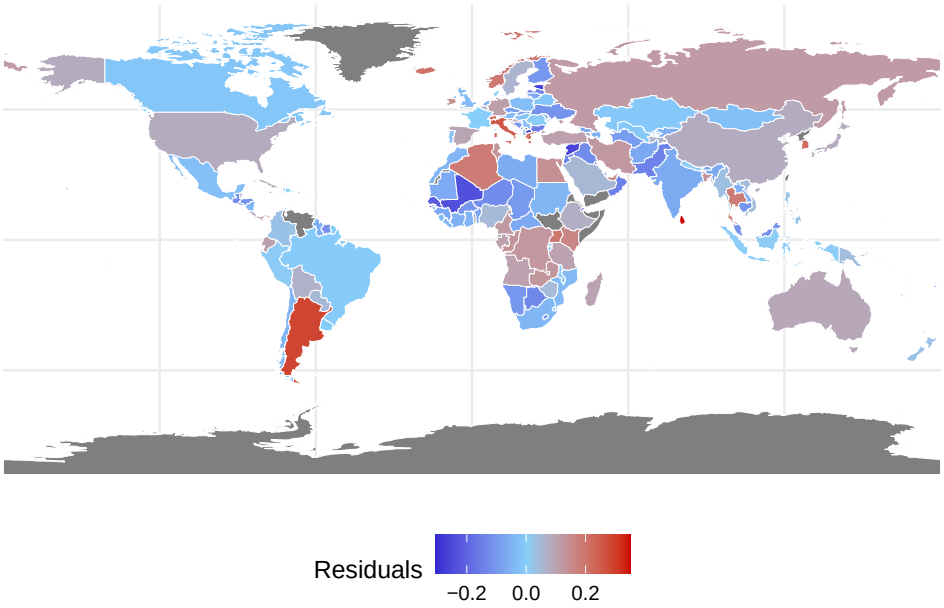

Figure 14: Spatial distribution of the residuals

Variations in colors suggest potential differences in residuals across geographical regions. Eastern Europe exhibits a consistent blue pattern, suggesting correlation of the residuals, while Western Europe shows more variations in residuals. South America presents spatial dependency, with clusters of blue and red areas indicating regional trends. North and Central America display mixed residuals, suggesting a less pronounced spatial correlation. Africa also shows clear patterns, with a red cluster in Central Africa and a large blue area in the North, suggesting that certain regional effects might not be fully captured. Also Central Asia similarly exhibits signs of spatial correlation. These patterns show that spatial dependencies could be influencing the residuals. The presence of geographical clusters of positive or negative residuals suggests that some important regional factors might not be fully captured in the model. For instance, cultural factors, political stability or conflicts can affect development beyond economic indicators; indeed, some countries might not fit well with the general model trends due to unique economic, social or political conditions.

Checking for unusual observations

```
lev <- hatvalues(Final_model)  #Leverage
stud_resid <- rstudent(Final_model)  #Standardized residuals
threshold_leverage <- 2 * (length(coef(Final_model)) / n)

outliers <- abs(stud_resid) > 3
high_leverage <- lev > threshold_leverage
```

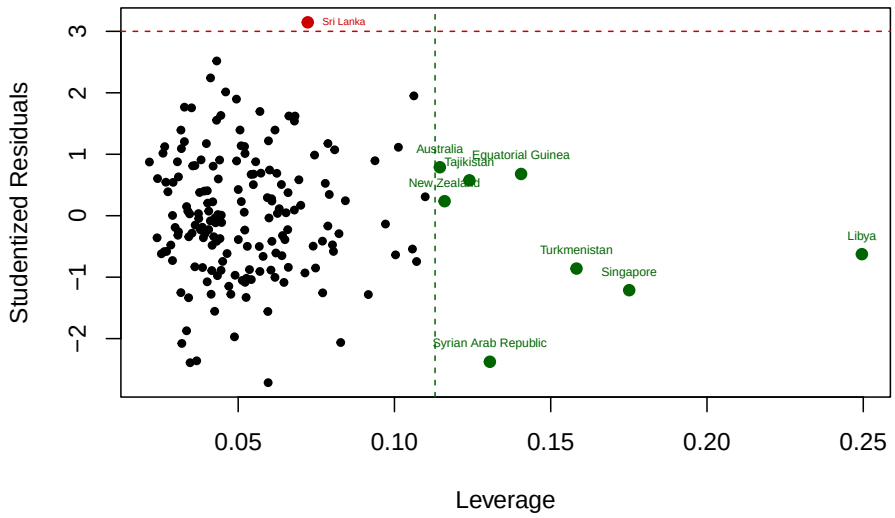

Figure 15: Residuals vs Leverage plot

Sri Lanka stands out as an outlier since its standardized residual falls outside the typical ± 3 range, which is commonly used to detect extreme observations. The green points represent high leverage points, indicating that these observations have extreme values in one or more covariates. Indeed, leverage measures how far an observation's predictor values deviate from the rest of the dataset. After reviewing summary statistics, I observed that some of these countries exhibit extreme values: Libya has the lowest value for *Regulatory Quality*, whereas Singapore has the highest. However, it is essential to verify whether these data points are also influential by examining Cook's distance.

```
alpha <-  0.05
bonferroni_quantile <- qt(1 - alpha/(2*n), df = Final_model$df.residual)
```

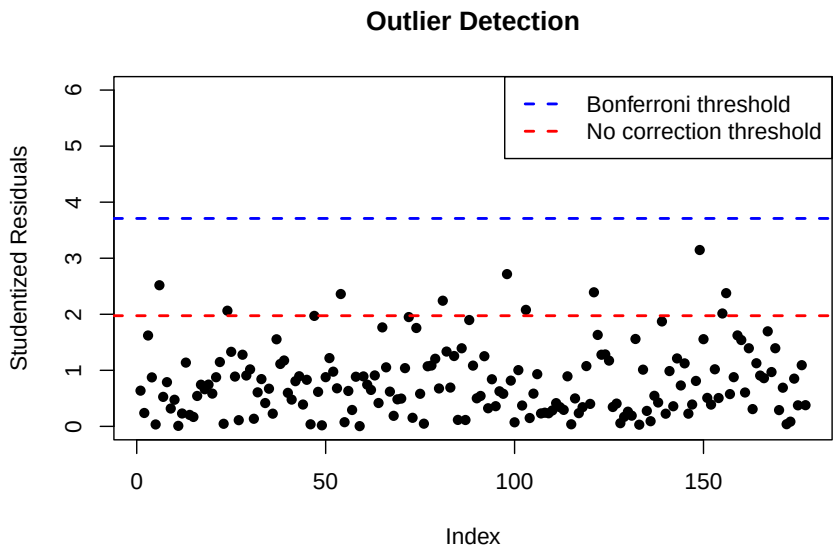

Figure 16: Bonferroni correction

Moreover, analyzing the Bonferroni correction plot, no data point falls outside the critical range. So, there are not extreme values of studentized residuals, so no extreme outliers.

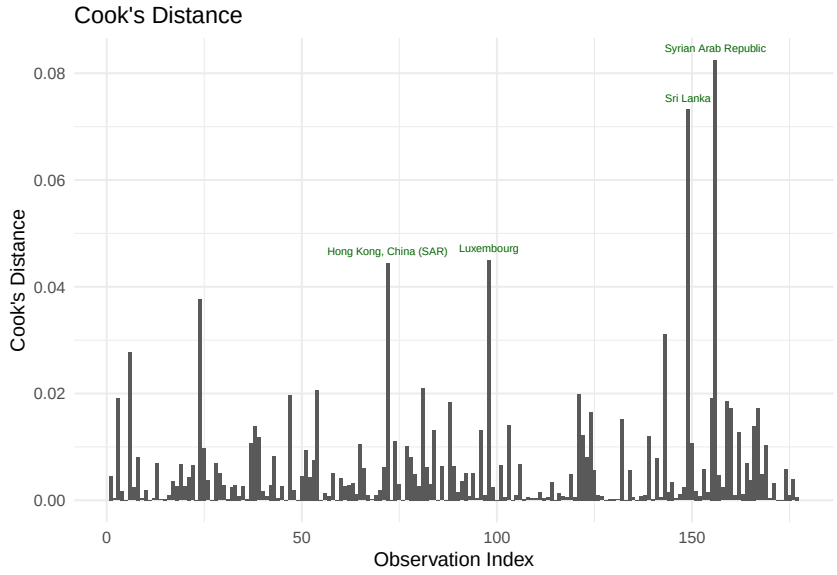

Figure 17: Cook's distance

There are no influential points as all the Cook's distance values are under 0.5. However, Sri Lanka, the only identified outlier, exhibits the second-highest Cook's distance among all countries. This suggests that, while not highly influential, it still has a relatively greater impact on the model compared to other observations.

Interpretation of the results

Now, I examine the results of the final model.

$$HDI_{probit} = \beta_0 + \beta_1 \cdot \text{Internet} + \beta_2 \cdot \text{RegulatoryQuality} + \beta_3 \cdot \text{VulnerableEmployment} + \beta_4 \cdot \text{GeoRegion} + \beta_5 \cdot \log(\text{CO}_2) + \beta_6 \cdot \text{RegulatoryQuality}^2 + \epsilon$$

Table 4: Regression Results

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	0.6724197	0.0265090	25.365744	0.0000000	0.6200838	0.7247556
Internet	0.0038102	0.0009891	3.852290	0.0001666	0.0018575	0.0057630
RegulatoryQuality	0.1910935	0.0149156	12.811697	0.0000000	0.1616462	0.2205409
VulnerableEmployment	-0.0021170	0.0007637	-2.771929	0.0062053	-0.0036248	-0.0006092
GeoRegionAfrica	-0.2466173	0.0346691	-7.113451	0.0000000	-0.3150636	-0.1781710
GeoRegionCentral Asia	-0.1129341	0.0311848	-3.621448	0.0003883	-0.1745014	-0.0513669
GeoRegionNorth and South America	-0.0934330	0.0313804	-2.977432	0.0033394	-0.1553864	-0.0314796
GeoRegionSoutheast Asia and Oceania	-0.1619347	0.0353430	-4.581807	0.0000090	-0.2317113	-0.0921581
log(CO2)	0.0739145	0.0160438	4.607053	0.0000081	0.0422397	0.1055892
I(RegulatoryQuality^2)	0.0862403	0.0089569	9.628370	0.0000000	0.0685569	0.1039236

The regression coefficients represent the effect of the independent variables on the probit-transformed Y. So, for example the average probit-HDI is equal to 0.6724197, for a country in Europe (baseline category), when all other variable are equal to their mean and CO_2 *emissions* is equal to 1 (since log(1) = 0). While for example the increase of one unit (one percentage point) in the percentage of people using the internet leads to a 0.0038 unit change in the probit-HDI. A one-unit increase in *Regulatory Quality* from its mean leads to a (0.191 + 2(0.0862) x *Regulatory Quality*) increase in probit-HDI. The positive quadratic term suggests a nonlinear relationship: higher values of *Regulatory Quality* lead to an accelerating effect on probit-HDI. So, if *Regulatory Quality* is at its mean (centered at 0), its effect on probit-HDI is just 0.191, while if it increases, its effect on HDI grows. Conversely, if *Regulatory Quality* decreases, the effect weakens and could even become negative for very low values. Regarding *Vulnerable Employment* a 1 percentage point increase in vulnerable employment relative to the mean leads to a 0.0021 decrease in probit-HDI. While for *log(CO_2)* a 1% increase in CO_2 emissions leads to a 0.0739 increase in probit-HDI. For the geographical variable the interpretation is the following:

- *Africa*: being in Africa decreases probit-HDI by 0.2466 relative to Europe;
- *Central Asia*: being in Central Asia decreases probit-HDI by 0.1129 relative to Europe;
- *North and South America*: probit-HDI is 0.0934 lower than Europe;
- *Southeast Asia and Oceania*: probit-HDI is 0.1619 lower than Europe.

However, using a probit transformation makes the interpretation of the coefficients more difficult, and the results can only be interpreted on the probit scale.

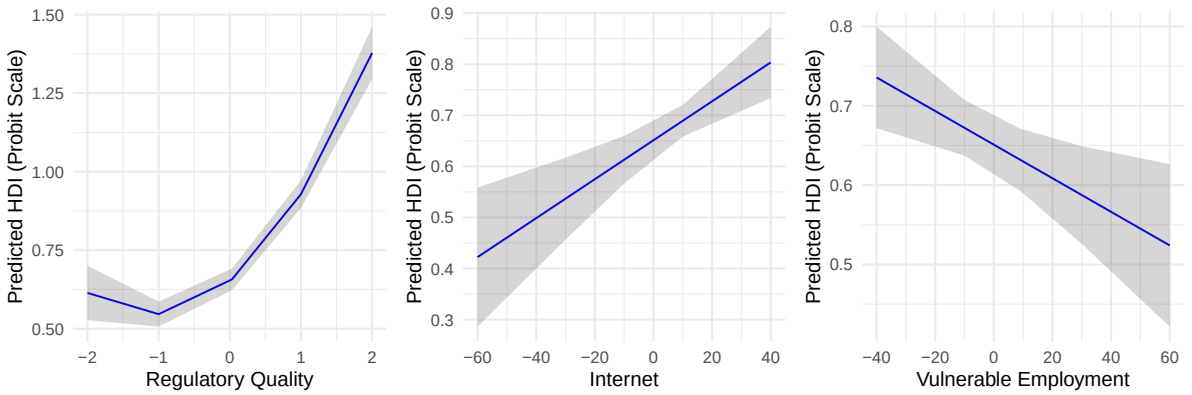

Figure 18: Effect plots of some variables

An effects plot displays the estimated impact of each predictor on the response variable while averaging out the effects of the other predictors. It helps interpret how changes in a specific predictor influence

the predicted response, holding all other variables constant. As expected, *Regulatory Quality* exhibits a quadratic effect, meaning that while low levels may not strongly impact HDI, higher values significantly contribute to human development. The curve bend upward: indeed the quadratic term is positive. *Internet* access positively affects HDI, confirming its crucial role in development, while *Vulnerable Employment* has a negative impact, emphasizing the importance of stable job opportunities in enhancing human development. Confidence intervals are wider at extreme values, indicating greater uncertainty in those regions.

t-tests

The t-statistic perform the following hypotheses:

$$\begin{cases} H_0 : \beta_i = 0 & \text{(The coefficient is not significant)} \\ H_1 : \beta_i \neq 0 & \text{(The coefficient is significant)} \end{cases}$$

Looking at the p-values in the summary of the regression in Table 4, all coefficients are statistically significant, meaning they are different from zero at any conventional significance level. This indicates that each predictor has a significant impact on the probit-HDI and contributes to explaining its variability within the model. This result aligns with my expectations since none of the confidence intervals in Table 4 contain 0.

Testing a group of regressors

Then, I decided to test a smaller model without $RegulatoryQuality^2$ and *Geographical Region* to determine whether these two variables contribute significantly to the model. This was done by performing an ANOVA test with the following hypotheses:

$$\begin{cases} H_0 : \beta_4 = 0, \quad \beta_6 = 0 & (\textit{Regulatory Quality}^2 \text{ and Geographical Region are not significant}) \\ H_1 : \beta_4 \neq 0 \text{ or } \beta_6 \neq 0 & \text{(At least one of the variables contributes significantly to the model)} \end{cases}$$

```
anova(restricted_model, Final_model)
```

```
## Analysis of Variance Table
##
## Model 1: HDI_probit ~ Internet + log(CO2) + VulnerableEmployment + RegulatoryQuality
## Model 2: HDI_probit ~ Internet + RegulatoryQuality + VulnerableEmployment +
##     GeoRegion + log(CO2) + I(RegulatoryQuality^2)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    172 4.3170
## 2    167 2.3311  5    1.9859 28.454 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is very low and so I reject the null hypothesis and I conclude that the removed variables ($Regulatoryquality^2$ and *Geographical region*) contribute significantly to the model. This further contributes to demonstrating that the quadratic term is significant.

Goodness of fit

```
## R-squared: 0.9477203
```

```
## Adjusted R-squared: 0.9449028
```

The values of R^2 and Adjusted R^2 are very similar and both quite high. Considering the Adjusted R^2, which accounts for the number of predictors in the model, the model explains 94.49% of the variability in probit-HDI. This indicates that the predictors capture almost all the variance in the response variable.

Prediction

Now, suppose there is a new observation in the dataset with the following values:

```
new_data <- data.frame(
  GeoRegion = "Africa",
  RegulatoryQuality = 0.5 - mean(Final_dataset$RegulatoryQuality),
  `I(RegulatoryQuality^2)` = (0.5 - mean(Final_dataset$RegulatoryQuality))^2,
  CO2 = log(7,65),
  Internet = 83 - mean(Final_dataset$Internet),
  VulnerableEmployment = 20 - mean(Final_dataset$VulnerableEmployment))

predict(Final_model, newdata = new_data, interval = "prediction", level = 0.95)
```

```
##         fit       lwr      upr
## 1 0.7604051 0.4508263 1.069984
```

The value 0.7604 represents the predicted HDI (transformed using the probit transformation) when the predictors take the specified values in the code above. Additionally, the displayed interval (0.4508, 1.0699) represents the confidence interval for this prediction, indicating the range within which the true predicted value is likely to fall.

Data simulation

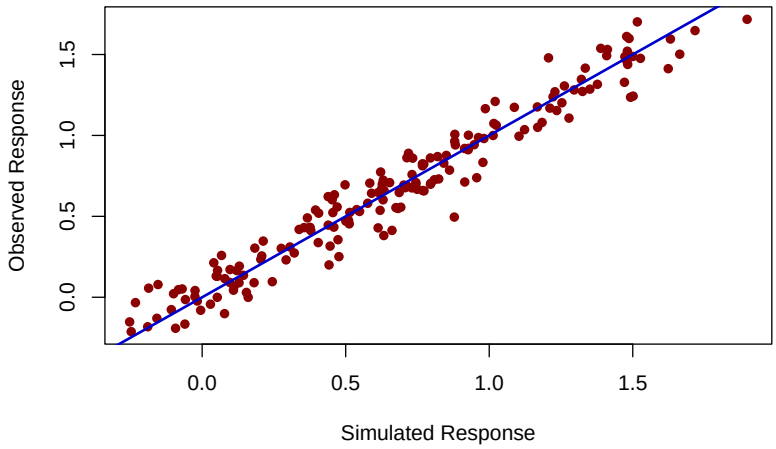

Figure 19: Simulated vs Observed response plot

The plot above illustrates that most points align closely with the blue line, indicating that the model effectively captures the observed data. While minor deviations suggest some prediction errors, they do not appear severe.

Conclusion

Overall, the model performs well in explaining HDI variability, demonstrating that factors beyond its standard computation, such as CO_2 emissions or internet access, can influence development. However, interpretability may be somewhat complex due to the probit transformation. The assumptions of linear regression seem to hold, except for spatial correlation of residuals, likely driven by similar economic and social conditions in certain regions. There is only one outlier (Sri Lanka), but it is not an influential point. Additionally, the high R^2 and the strong significance of all coefficients reinforce the model's explanatory power. Nevertheless, it is important to acknowledge that HDI is a simplified measure of human development, omitting aspects such as inequality, poverty, and human security. While the included variables do not directly address these dimensions (except *Regulatory Quality*), they remain highly relevant in capturing socio-economic and environmental factors influencing HDI.

References

- United Nation Development programs, 2020, Human Development Index Dataset, https://www.undp.org/
- World Bank, last update: 01/28/2025, World Development Indicators Dataset, https://databank.worldbank.org/source/world-development-indicators
- Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). *An introduction to statistical learning with applications in R*. New York, Springer.