CO2, Growth, and Life Expectancy

Hedayatullah Hakimi

10/05/2023

Contents

1 Abstract

This project analyzes the relationships between socioeconomic factors, CO2 emissions per capita, and life expectancy at birth using data from the World Bank's World Development Indicators. Linear regression models are employed to explore these relationships, considering factors such as GDP per capita, access to electricity, secondary school enrollment, agriculture's share of GDP, and income inequality. The analysis reveals complex interactions between these variables, highlighting the challenges in balancing economic development, environmental sustainability, and human wellbeing.

2 Introduction

The World Development Indicators database provides a comprehensive collection of data on global development. This study aims to understand the complex relationships between economic, social, and environmental factors that shape our world. By examining these interconnections, valuable insights can be gained to inform sustainable development strategies and policy decisions.

The research focuses on two key aspects of development: environmental impact (measured by CO2 emissions) and human wellbeing (measured by life expectancy). It investigates how these outcomes are influenced by various socioeconomic factors, including economic output, access to electricity, education, agricultural dependence, and income inequality.

Understanding these relationships is crucial for addressing global challenges such as climate change, poverty, and health disparities. The findings may help policymakers navigate the often conflicting goals of economic growth, environmental protection, and social welfare.

3 Data Description and Preparation

Data from the World Bank's World Development Indicators database is used, extracted using R's WDI package. The dataset covers all countries from 2010 to 2020 and includes the following variables:

- Gross Domestic Product per capita (GDP_pc)
- Life expectancy at birth (life_exp)
- Gross secondary school enrollment (sec_enroll)
- Access to electricity in urban areas (elec_access)
- Agriculture, forestry, and fishing, value added as % of GDP (agri_gdp)
- GINI index (gini)
- CO2 emissions per capita (CO2)

```
# Set the country codes and years of interest
countries <- "all"
years <- c(2010:2020)

# Extract the data for the variables of interest
indicators <- c("NY.GDP.PCAP.KD", "SP.DYN.LE00.IN", "SE.SEC.ENRR",
                "EG.ELC.ACCS.UR.ZS", "NV.AGR.TOTL.ZS", "SI.POV.GINI", "EN.ATM.CO2E.PC")
data <- WDI(country = countries, indicator = indicators, start = years[1], end = years[length(years)])

# Rename columns for clarity
names(data)[names(data) %in% indicators] <- c("GDP_pc", "life_exp", "sec_enroll",
                                              "elec_access", "agri_gdp", "gini", "CO2")

# Remove rows with missing values
data <- na.omit(data)

# Display summary statistics
summary(data[,c("GDP_pc", "life_exp", "sec_enroll", "elec_access", "agri_gdp", "gini", "CO2")])
```

```
##      GDP_pc            life_exp       sec_enroll      elec_access
##  Min.   :   263.4   Min.   :50.01   Min.   : 13.54   Min.   : 13.60
##  1st Qu.:  5053.8   1st Qu.:72.89   1st Qu.: 90.07   1st Qu.: 99.70
##  Median : 11728.7   Median :76.50   Median :100.33   Median :100.00
##  Mean   : 20117.9   Mean   :75.72   Mean   : 97.08   Mean   : 97.27
##  3rd Qu.: 30242.4   3rd Qu.:80.99   3rd Qu.:108.13   3rd Qu.:100.00
##  Max.   :108351.4   Max.   :83.90   Max.   :164.08   Max.   :100.00
##     agri_gdp            gini            CO2
##  Min.   : 0.1991   Min.   :23.20   Min.   : 0.0361
##  1st Qu.: 1.9019   1st Qu.:30.00   1st Qu.: 2.0802
##  Median : 4.0454   Median :34.40   Median : 4.4117
##  Mean   : 6.7163   Mean   :35.44   Mean   : 5.1674
##  3rd Qu.: 8.5421   3rd Qu.:40.00   3rd Qu.: 7.1169
##  Max.   :58.9344   Max.   :63.40   Max.   :32.2566
```

4 Exploratory Data Analysis

4.1 Correlation Analysis

The analysis begins by examining the correlations between the variables:

```
cor_matrix <- cor(data[,c("GDP_pc", "life_exp", "sec_enroll", "elec_access", "agri_gdp", "gini", "CO2")]
corrplot(cor_matrix, method = "color", type = "upper", order = "hclust",
         tl.col = "black", tl.srt = 45, addCoef.col = "black", number.cex = 0.7)
```

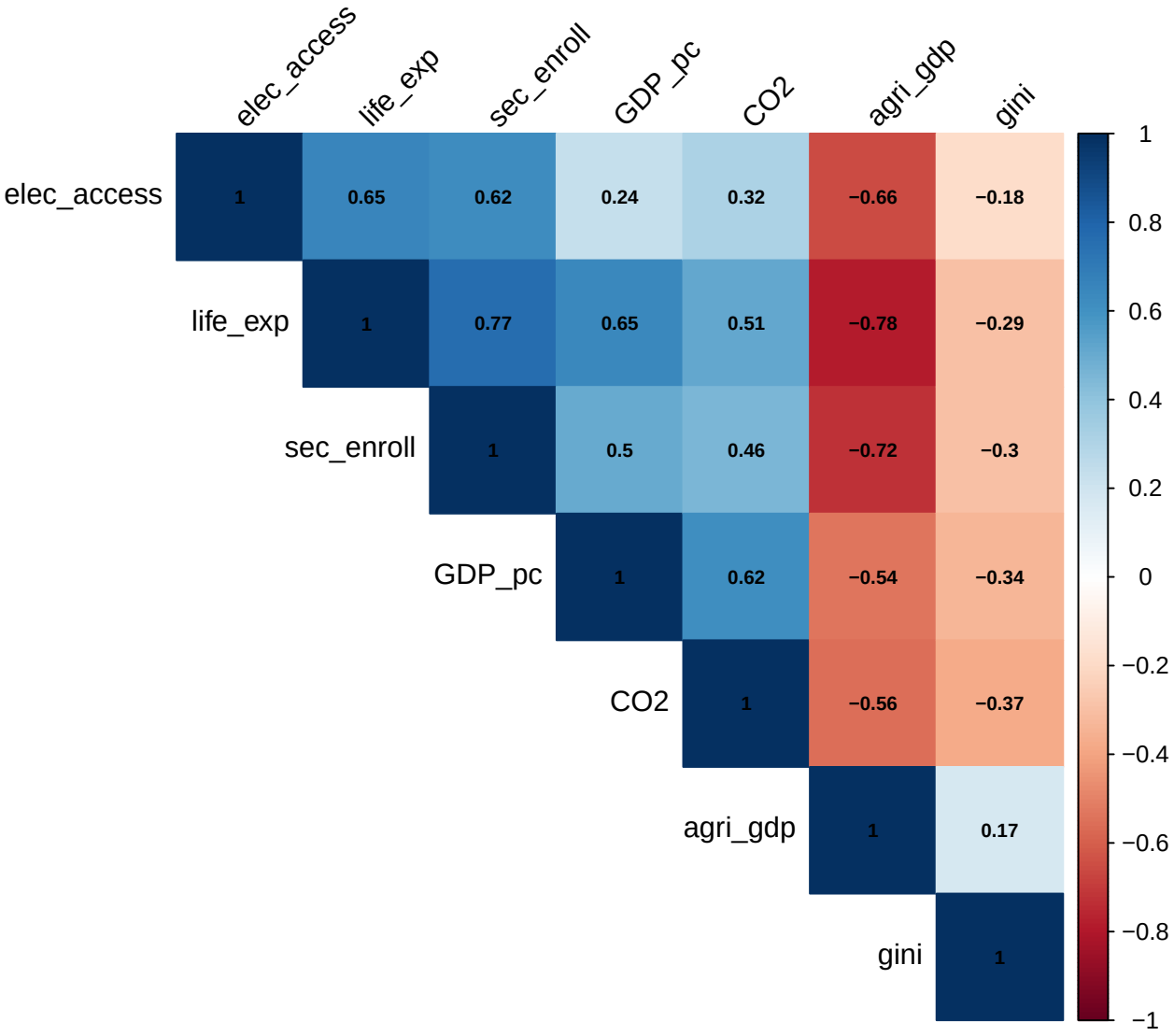

Figure 1: Figure 1: Correlation Matrix of Socioeconomic Indicators

Key observations from the correlation matrix:

1. GDP per capita shows strong positive correlations with life expectancy (0.65) and access to electricity (0.62), suggesting that economic development is associated with improved health outcomes and infrastructure.
2. CO2 emissions have a moderate positive correlation with GDP per capita (0.62) and life expectancy (0.51), which is somewhat counterintuitive and warrants further investigation.

3. The GINI index exhibits negative correlations with most variables, particularly strong with life expectancy (-0.78) and secondary school enrollment (-0.72), indicating that higher income inequality is associated with poorer outcomes in these areas.

It's important to note that correlation does not imply causation, and these relationships may be influenced by confounding factors not captured in this analysis.

4.2 Scatterplot Matrix

To visualize these relationships, a scatterplot matrix is created:

```
pairs(data[,c("GDP_pc", "life_exp", "sec_enroll", "elec_access", "agri_gdp", "gini", "CO2")])
```

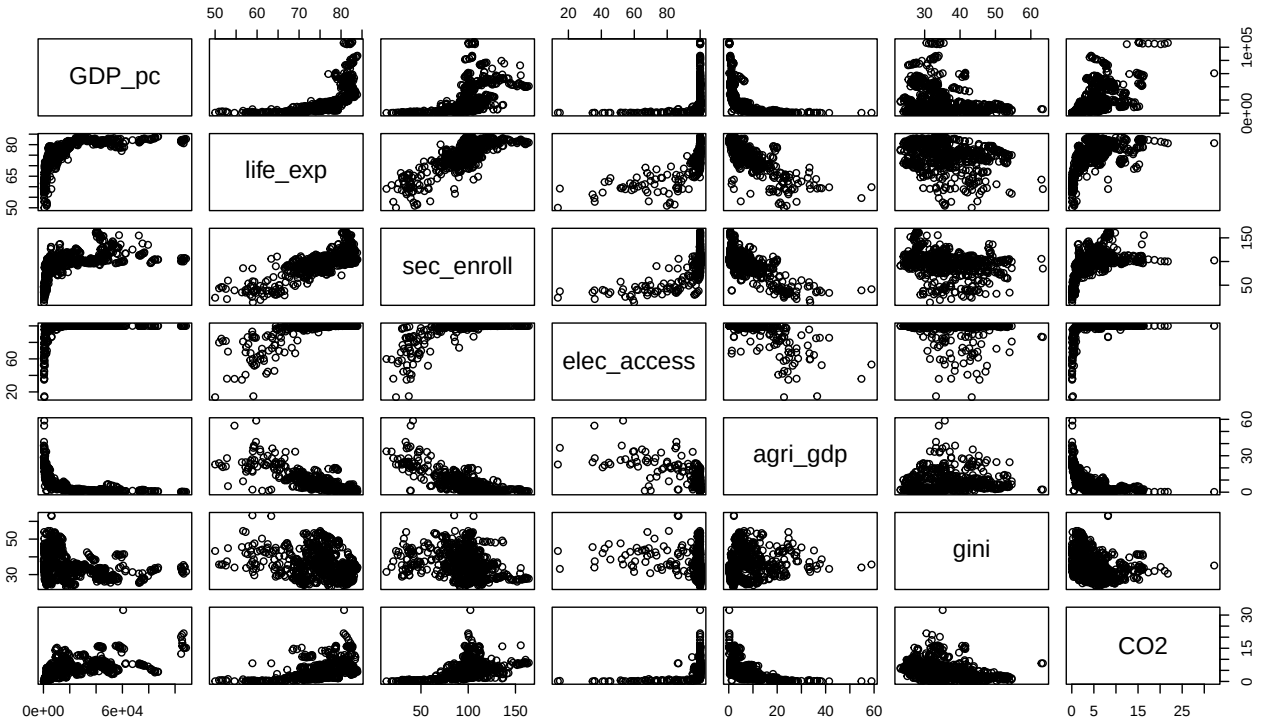

Figure 2: Figure 2: Scatterplot Matrix of Socioeconomic Indicators

The scatterplot matrix provides additional insights:

1. The relationship between GDP per capita and CO2 emissions appears non-linear, with emissions increasing more rapidly at higher GDP levels. This suggests that economic growth may lead to disproportionate increases in emissions as countries develop.
2. Life expectancy shows a positive trend with GDP per capita, but the relationship appears to plateau at higher GDP levels, indicating diminishing returns to life expectancy from economic growth beyond a certain point.
3. The relationships involving the GINI index show considerable scatter, suggesting complex interactions that may not be fully captured by linear models.

5 Regression Analysis

5.1 Model 1: CO2 Emissions

The relationship between CO2 emissions and GDP per capita, access to electricity, and secondary school enrollment is examined:

```r
model1 <- lm(CO2 ~ GDP_pc + elec_access + sec_enroll, data = data)
summary(model1)
```

```
##
## Call:
## lm(formula = CO2 ~ GDP_pc + elec_access + sec_enroll, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.5349 -1.7409 -0.6161  1.0716 23.0084
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.346e+00  1.113e+00  -3.005 0.002744 **
## GDP_pc       9.514e-05  5.817e-06  16.354  < 2e-16 ***
## elec_access  4.835e-02  1.421e-02   3.403 0.000702 ***
## sec_enroll   1.953e-02  6.756e-03   2.891 0.003946 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.996 on 757 degrees of freedom
## Multiple R-squared:  0.418,  Adjusted R-squared:  0.4157
## F-statistic: 181.2 on 3 and 757 DF,  p-value: < 2.2e-16
```

5.1.1 Interpretation of Model 1

- All predictors show statistically significant relationships with CO2 emissions.
- GDP per capita has the largest coefficient, suggesting it has a strong association with CO2 emissions. However, caution should be taken about claiming "strongest effect" without standardizing the variables.
- The positive coefficients for all variables indicate that increases in GDP, electricity access, and school enrollment are associated with higher CO2 emissions, highlighting the environmental challenges of development.
- The adjusted R-squared value of 0.4157 suggests that the model explains about 41.57% of the variance in CO2 emissions, indicating that there are likely other important factors not included in this model.

5.1.2 Assumption Checking for Model 1

```r
par(mfrow = c(2, 2))
plot(model1)
```

The diagnostic plots reveal several issues:

1. Residuals vs Fitted: Shows a clear non-linear pattern, indicating that the linear model may not be appropriate.
2. Normal Q-Q: Deviations from the line, particularly in the tails, suggest non-normality of residuals.
3. Scale-Location: The spread of residuals is not constant, indicating heteroscedasticity.
4. Residuals vs Leverage: Some points have high leverage, but none appear to have excessive influence on the model.

To address these issues, a log transformation is applied to the CO2 emissions variable:

```r
model1_log <- lm(log(CO2) ~ GDP_pc + elec_access + sec_enroll, data = data)
summary(model1_log)
```

```
##
## Call:
```

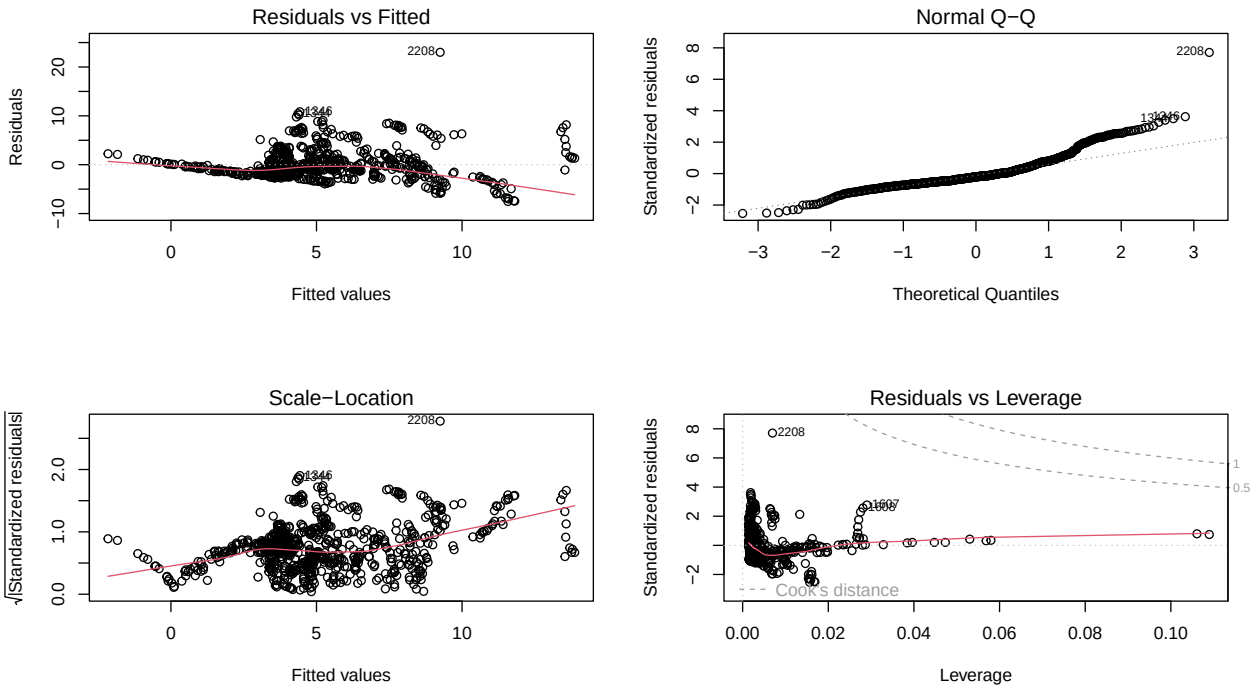

Figure 3: Figure 3: Diagnostic Plots for Model 1

```
## lm(formula = log(CO2) ~ GDP_pc + elec_access + sec_enroll, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.17545 -0.40676 -0.03019  0.42270  2.45585
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.010e+00  2.241e-01  -22.35   <2e-16 ***
## GDP_pc       1.385e-05  1.171e-06   11.82   <2e-16 ***
## elec_access  4.774e-02  2.860e-03   16.69   <2e-16 ***
## sec_enroll   1.389e-02  1.360e-03   10.21   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.603 on 757 degrees of freedom
## Multiple R-squared:  0.6811, Adjusted R-squared:  0.6798
## F-statistic: 538.9 on 3 and 757 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(model1_log)
```

The log transformation improves the model fit and addresses some of the assumption violations:

- The residuals vs fitted plot now shows a more random scatter.
- The Q-Q plot is closer to a straight line, indicating better normality.
- The scale-location plot shows more constant variance.

In this transformed model:

- A one-unit increase in GDP per capita is associated with a 1.385e-05 * 100 = 0.001385% increase in

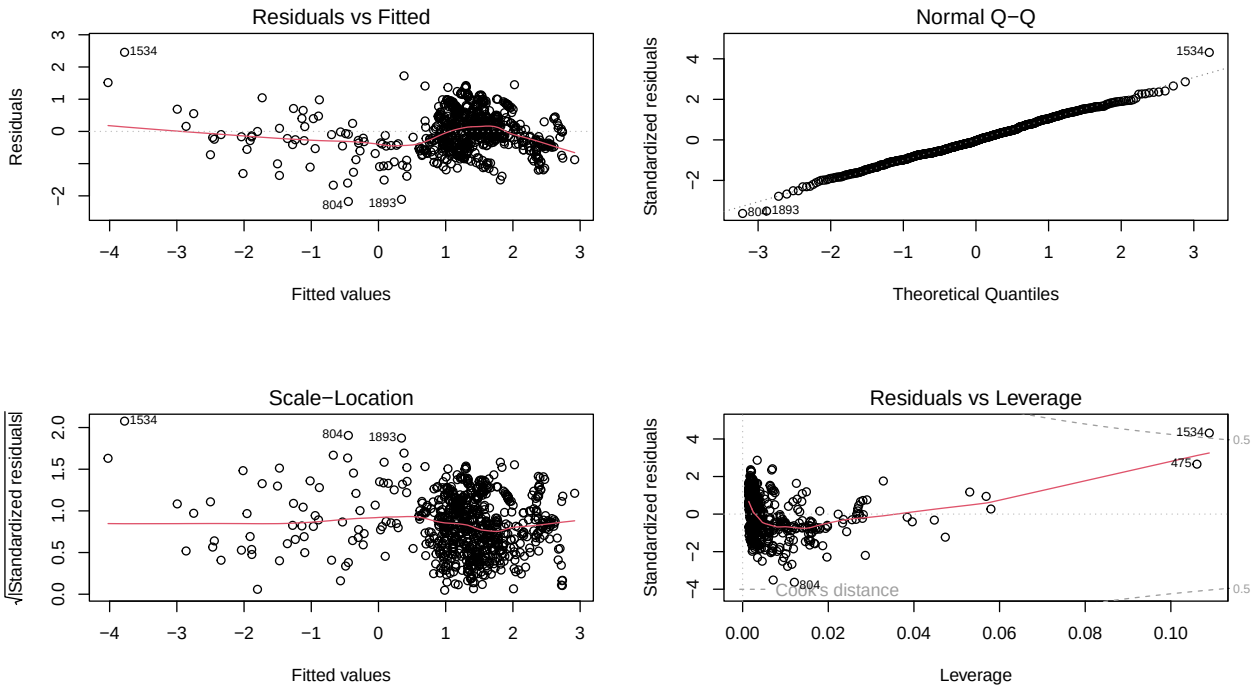

Figure 4: Figure 4: Diagnostic Plots for Log-Transformed Model 1

CO2 emissions, holding other variables constant.

- A one-percentage point increase in access to electricity is associated with a 4.774% increase in CO2 emissions.
- A one-unit increase in secondary school enrollment is associated with a 1.389% increase in CO2 emissions.

The adjusted R-squared has improved to 0.6798, indicating that the log-transformed model explains about 67.98% of the variance in log(CO2) emissions.

5.2 Model 2: Life Expectancy

The relationship between life expectancy and agriculture's share of GDP, GINI index, and CO2 emissions is examined:

```
model2 <- lm(life_exp ~ agri_gdp + gini + CO2, data = data)
summary(model2)
```

```
##
## Call:
## lm(formula = life_exp ~ agri_gdp + gini + CO2, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.196  -2.120   0.656   2.372  16.909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 83.79372    0.88143  95.066  < 2e-16 ***
## agri_gdp    -0.61928    0.02238 -27.674  < 2e-16 ***
## gini        -0.12385    0.02041  -6.067 2.05e-09 ***
```

```
## CO2          0.09148    0.04457   2.053   0.0404 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.768 on 757 degrees of freedom
## Multiple R-squared:  0.6399, Adjusted R-squared:  0.6385
## F-statistic: 448.5 on 3 and 757 DF,  p-value: < 2.2e-16
```

5.2.1 Interpretation of Model 2

- All predictors show statistically significant relationships with life expectancy.
- Agriculture's share of GDP and the GINI index have negative coefficients, suggesting that higher values in these variables are associated with lower life expectancy.
- Surprisingly, CO2 emissions show a positive association with life expectancy. This counterintuitive result may be due to confounding factors such as overall development level.
- The adjusted R-squared of 0.6385 indicates that the model explains about 63.85% of the variance in life expectancy.

5.2.2 Assumption Checking for Model 2

```
par(mfrow = c(2, 2))
plot(model2)
```

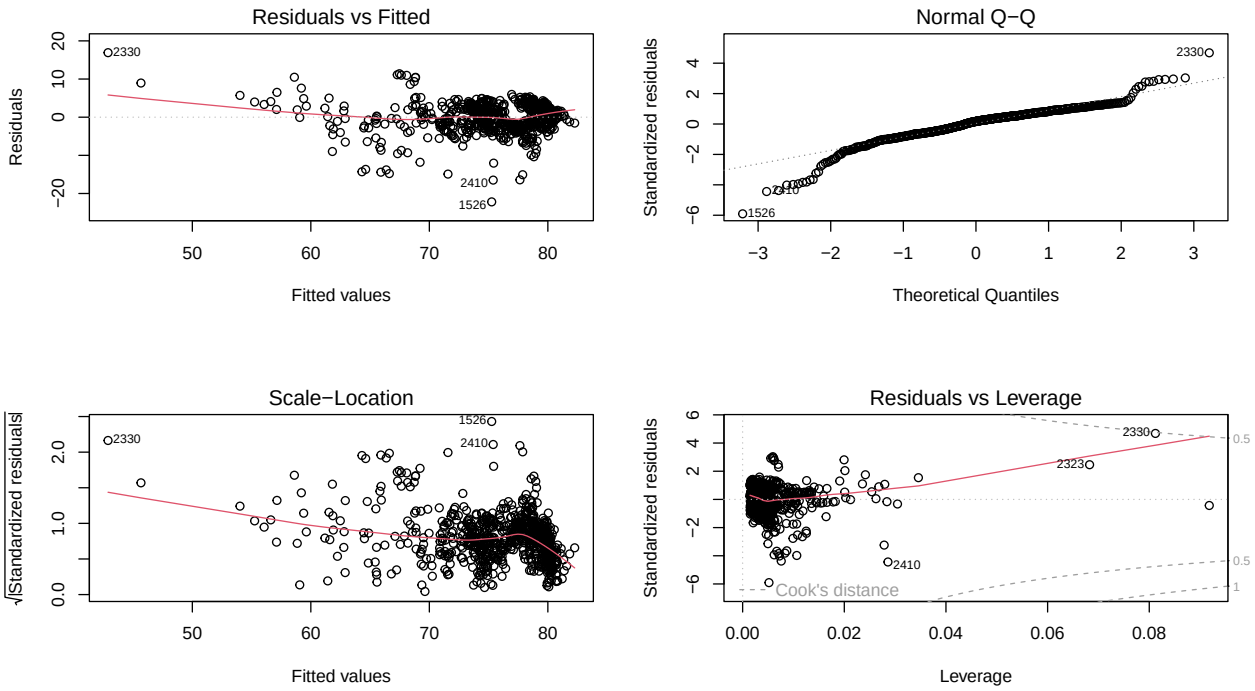

Figure 5: Figure 5: Diagnostic Plots for Model 2

The diagnostic plots for Model 2 show:

1. Residuals vs Fitted: A slight curve is visible, suggesting some non-linearity.

2. Normal Q-Q: The residuals follow the line quite well, indicating normality.
3. Scale-Location: The spread of residuals is relatively constant, suggesting homoscedasticity.
4. Residuals vs Leverage: No points appear to have excessive influence on the model.

While the assumptions for Model 2 are reasonably met, a Box-Cox transformation can be explored to potentially improve the model:

```
boxcox_res <- boxcox(life_exp ~ agri_gdp + gini + CO2, data = data)
```

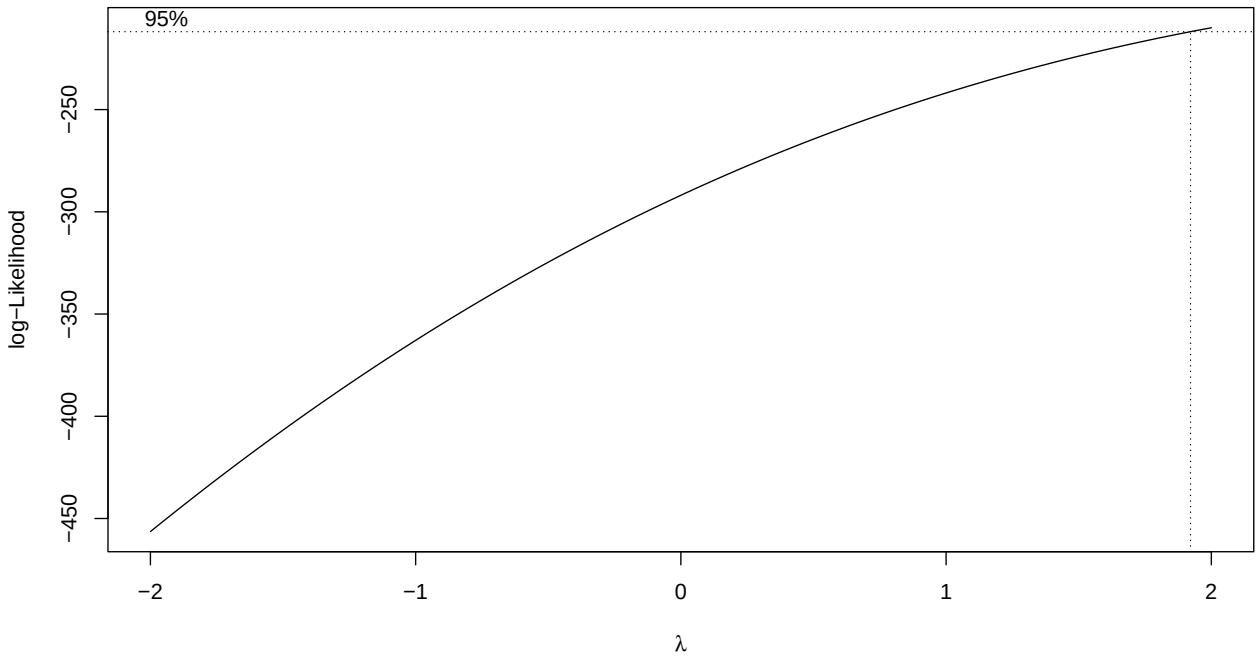

Figure 6: Figure 6: Diagnostic Plots for Box-Cox Transformed Model 2

```
lambda <- boxcox_res$x[which.max(boxcox_res$y)]
data$boxcox_life_exp <- (data$life_exp^lambda - 1) / lambda
model2_boxcox <- lm(boxcox_life_exp ~ agri_gdp + gini + CO2, data = data)
summary(model2_boxcox)
```

```
##
## Call:
## lm(formula = boxcox_life_exp ~ agri_gdp + gini + CO2, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1435.62  -171.62    47.34   176.55  1224.50
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3465.008     63.762  54.343  < 2e-16 ***
## agri_gdp     -43.678      1.619 -26.982  < 2e-16 ***
## gini          -9.219      1.477  -6.243 7.13e-10 ***
## CO2            7.893      3.224   2.448   0.0146 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 272.5 on 757 degrees of freedom
## Multiple R-squared:  0.6347, Adjusted R-squared:  0.6332
## F-statistic: 438.3 on 3 and 757 DF,  p-value: < 2.2e-16
```

```
par(mfrow = c(2, 2))
plot(model2_boxcox)
```

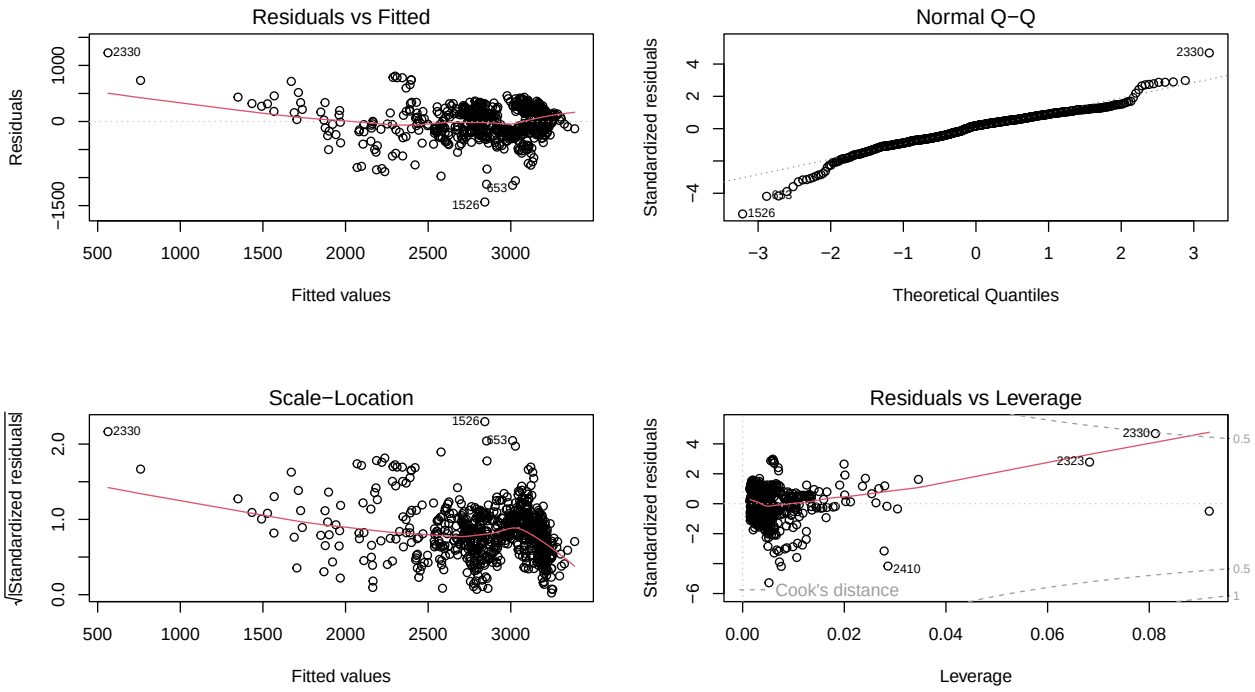

Figure 7: Figure 6: Diagnostic Plots for Box-Cox Transformed Model 2

The Box-Cox transformation slightly improves the model fit, but the interpretation becomes more complex. Given that the original model reasonably met assumptions, it may be preferable to retain the untransformed model for ease of interpretation.

6 Discussion

The analysis reveals complex relationships between socioeconomic factors, CO2 emissions, and life expectancy:

1. Economic development, as measured by GDP per capita, is associated with both increased CO2 emissions and higher life expectancy. This highlights the tension between development goals and environmental sustainability.
2. Access to electricity and secondary education are positively associated with CO2 emissions, suggesting that as countries develop and improve these indicators, they face increased environmental challenges.
3. Higher agricultural GDP share and income inequality are associated with lower life expectancy, indicating potential health challenges in agricultural economies and unequal societies.
4. The positive association between CO2 emissions and life expectancy is surprising and warrants careful interpretation. This relationship likely reflects the confounding influence of overall development rather than a direct causal link between emissions and health outcomes.

6.1 Practical Significance

While the models show statistically significant relationships, it's important to consider their practical implications:

- The log-transformed CO2 emissions model suggests that even small increases in GDP per capita or access to electricity are associated with substantial increases in emissions. This underscores the need for sustainable development strategies that can improve living standards without proportional increases in environmental impact.
- The life expectancy model highlights the importance of addressing income inequality and diversifying economies heavily dependent on agriculture. Even small reductions in the GINI index or agriculture's share of GDP are associated with meaningful increases in life expectancy.

6.2 Limitations

Several limitations of this study should be noted:

1. The models assume mostly linear relationships, which may not fully capture the complexity of these interactions.
2. Causality cannot be inferred from these observational data.
3. There may be important confounding variables not included in the models, such as healthcare spending or environmental regulations.
4. The analysis uses country-level data, which may mask important within-country variations.
5. The time period of 2010-2020 may not capture long-term trends or recent global events that could impact these relationships.

6.3 Future Research

Future research could address these limitations and expand on the findings:

1. Explore non-linear relationships, such as using polynomial terms for GDP in the CO2 emissions model.
2. Incorporate additional variables like healthcare expenditure, urban population percentage, or renewable energy usage.
3. Employ more advanced statistical techniques, such as mixed-effects models to account for country-specific effects over time.
4. Conduct regional analyses to understand how these relationships vary across different parts of the world.
5. Investigate the potential for interaction effects between variables.

7 Conclusion

This study provides insights into the complex interplay between economic development, environmental impact, and human wellbeing. While economic growth and improved access to education and electricity are associated with better life expectancy, they also come with environmental costs in terms of increased CO2 emissions. The challenge for policymakers is to find ways to promote development and improve living standards while mitigating environmental impacts.

The findings highlight the need for nuanced, context-specific approaches to sustainable development that consider the intricate relationships between economic, social, and environmental factors. Future research should focus on identifying strategies that can help countries navigate the trade-offs between development goals and environmental sustainability.

8 References

World Bank. (2021). World Development Indicators. Retrieved from https://databank.worldbank.org/source/world-development-indicators

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Wickham, H. et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686