Title : An Analysis of Global CO_2 Emissions and the Impact of GDP and Working-Age Population in 2020

Author : Pratik Ganguli (pgan501)

The ***wdi_2020*** dataset, sourced from the World Bank's **World Development Indicators(WDI)**, contains 217 observations covering various global development factors for the year 2020. This analysis focuses on predicting CO_2(CO_2 emissions per capita, in metric tons) as the dependent variable, using GDP(GDP per capita, in USD) and the WAP (working-age population, as the percentage of the total population aged 15 to 64 years) as independent variables. To ensure data consistency and analysis quality, 33 missing entries in key columns were removed, making the dataset suitable for accurate modeling.

To improve data interpretability and model performance, a logarithmic transformation was applied to these variables, addressing skewness of CO_2 and GDP caused by extreme values. This transformation enhanced distribution normality, reduced the influence of outliers, and standardized the variables, thereby strengthening the model's reliability. As illustrated in [Figure 1.1 & 1.2], the ggpairs plot compares relationships before and after the transformation. The correlation analysis of the transformed variables showed a strong relationship between GDP and CO_2(r = 0.850), highlighting the impact of economic growth on emissions. The WAP also correlated positively with CO_2(r = 0.789), suggesting that a larger workforce drives industrialization and urbanization, leading to higher emissions.

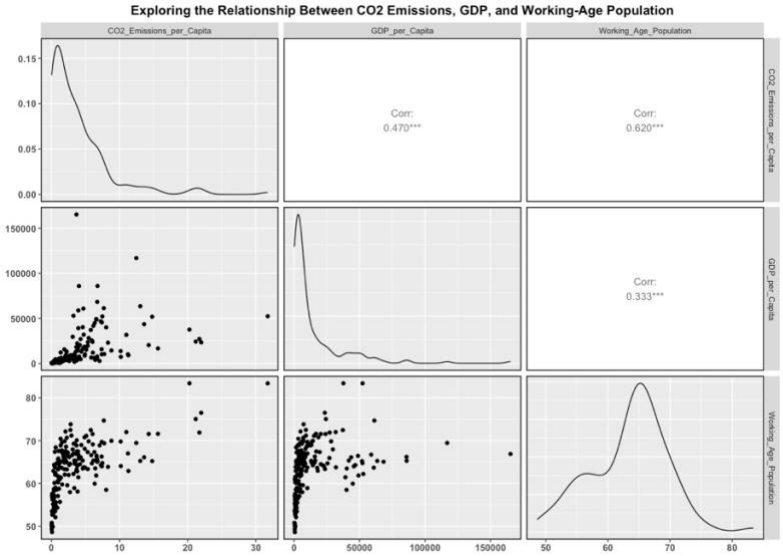

Figure 1.1 *Pairwise Correlation of All the Key Factors*

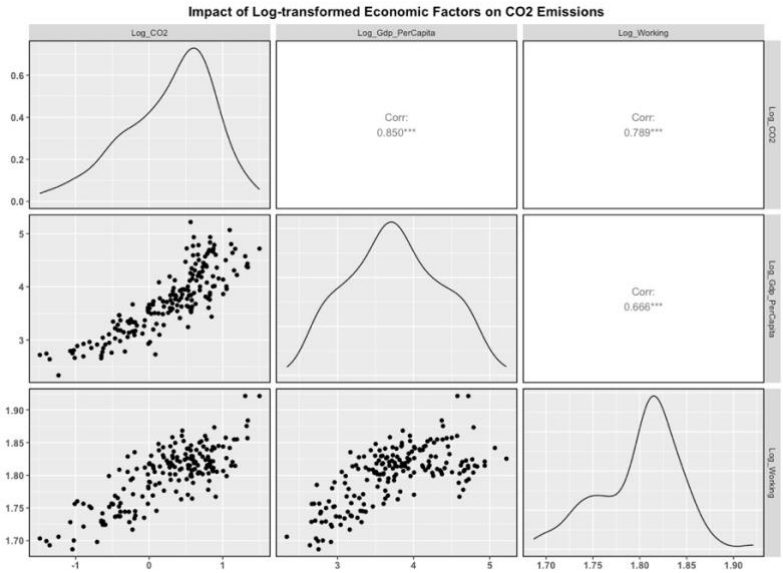

Figure 1.2 *Pairwise Correlation for Logarithmic Variables*

The [Table 1] summarizes a multiple linear regression model for predicting CO_2 with the equation $\hat{y} = -12.09 + 0.58x_1 + 5.65x_2$, where x_1 and x_2 represent GDP and WAP, respectively. Holding WAP constant, a 1% increase in GDP results in a 0.58% rise in CO_2, while holding GDP constant, a 1% increase in WAP leads to a 5.65% increase in CO_2, suggesting that a larger workforce contributes to economic expansion, and industrial activity, thereby increasing emissions. With a significance of $p < 2.2 \times 10^{-16}$ and an adjusted R^2 of 0.80, the model exhibits strong predictive power, explaining 80.94% of the variability in CO_2, driven by the combined influence of x_1 and x_2 as predictors. Additionally, the low residual standard error(0.26) supports the model's reliability in capturing the key economic factors driving CO_2 emissions.

The assumptions for linear regression appear to be satisfied, as shown in[Figure1.3]. The residual scattered plots exhibit no clear patterns and are randomly scattered around zero, indicating homoscedasticity and the absence of systematic bias. The histogram demonstrates that the residuals follow an approximately normal distribution, supporting the assumption of normality for valid statistical inference. Additionally, the scatterplots confirm the linear relationship between the independent variables, reinforcing the suitability of a linear regression model.

In summary, the analysis underscores the substantial impact of GDP and WAP on CO_2, emphasizing the relationship between economic development and environmental challenges. The regression model delivers reliable predictions, offering valuable insights into the factors driving carbon emissions.

Coefficient	**Estimate**
Intercept	-12.09
Log_Gdp_PerCapita	0.58
Log_Working	5.65
Adjusted R-squared (R^2)	**80.94%**
p-value	$< 2.2 \times 10^{-16}$
Residual standard error	0.26

Table 1 *Summary Statistics of the Linear Regression Model*

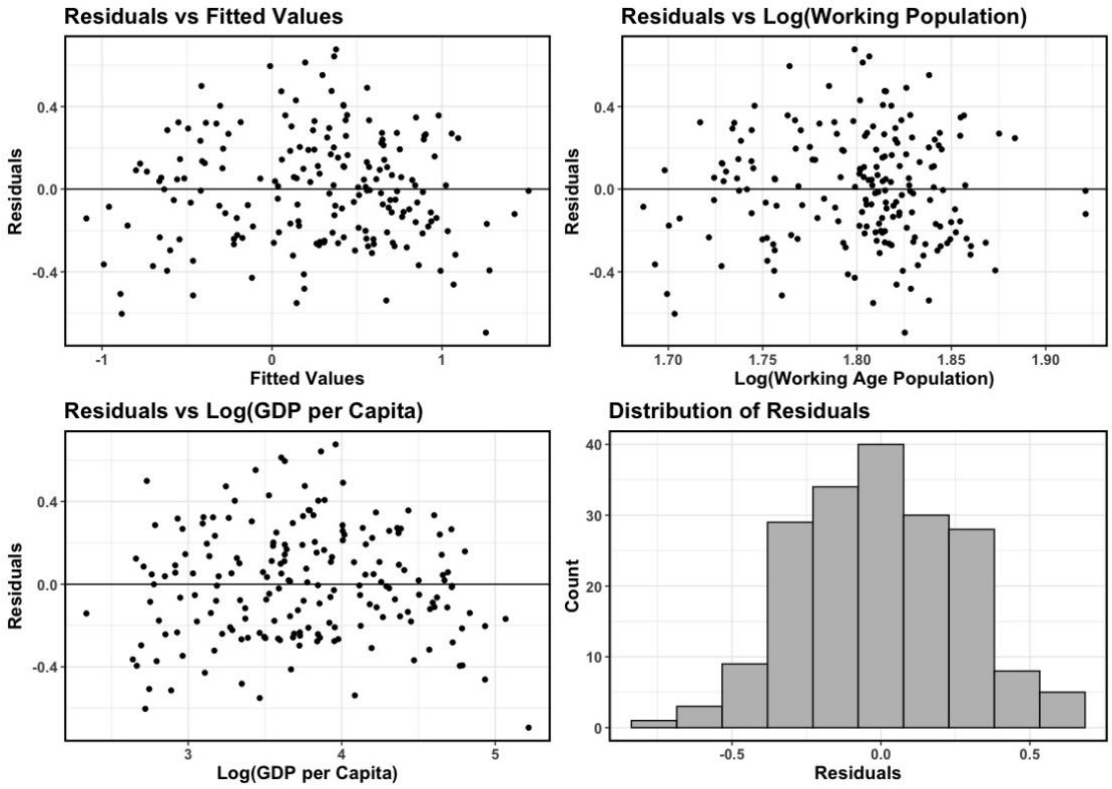

Figure 1.3 *Residual Plot Assessing the Impact of Economic Factors on CO_2 Emissions*