Linear regression analysis of the number of points scored at IMO 2015

1 Introduction

In my research, I examine the relationships between the total number of points scored by countries in the International Mathematical Olympiad (IMO) in 2015 and a set of 10 explanatory variables. I analyze various economic, social, and educational indicators for individual countries to identify factors that may influence mathematical performance in the IMO

2 Basic data characteristics

- *imo2015_tasks_points*: Dependent variable. Describes the number of points scored by countries in the IMO 2015, which is a measure of their success in the competition.
- *GDP*: Gross Domestic Product of the country. This economic indicator reflects the total value of goods and services produced in the country during a year. It serves as a measure of the country's size and economic development.
- *population*: This variable represents the number of inhabitants in a given country.
- *high_tech_exports*: High-tech exports ($). This variable refers to the value of advanced technology exports by a given country. It is an indicator that can represent the level of innovation and engagement of the country in the high-tech sector.
- *migration*: Number of migrants. This variable indicates the number of people migrating to a given country.
- *gov_educ_expenditure*: Percentage of public expenditure on education (% of GDP).
- *area*: Country area. This variable refers to the land area occupied by a given country.
- *prob_of_death*: Probability of dying. This indicator represents the level of safety and social stability in a given country.
- *internet_user_percentage*: Internet access rate. This indicator reflects the percentage of the population that has access to the Internet.
- *gross_enrollment_ratio*: Ratio of total enrollment, to the population of the age group that officially corresponds to the level of education shown
- *unemployment_rate*: This indicator measures the percentage of unemployed individuals in relation to the active labor force.

The data for explanatory variables were obtained from World Bank Open Data for the year 2015. The data related to the dependent variable, which is the points scored by countries, was acquired from the website IMO 2015.
The presented summary of data provides information about a sample consisting of 59 observations. The minimum, maximum, median, mean, and quartile values indicate that the data have a diverse range. For instance, *GDP* ranges from 1.055×10^{10} to 4.445×10^{12}, while *population* ranges from 3.308×10^{5} to 1.323×10^{9}. Our dataset contains both variables expressed as numbers, such as *GDP*, representing specific economic values, and variables expressed as proportions, such as *prob_of_death*, which are scaled within the range of 0 to 1. There are no missing data in the analyzed dataset.

3 Model Evaluation and Performance Metrics

Residual Standart Error(RSE) is 30.32. For our explanatory variable, which takes values from 0 to 151, this RSE value is relatively high, indicating some degree of mismatch between the model and the data.
Multiple R-squared is 0.5513, which means that 55.13% of Y variation is explained by the independent variables in this model. The remaining 44.87% of Y's variability is not accounted for in this model and may be caused by other factors, random errors or unaccounted variables.
Adjusted R-squared is 0.4579, which means that 45.79% of Y's variability is explained by the independent variables in this model, taking into account the number of independent variables and degrees of freedom
F-statistic is 5.899, and p-value is 9.57e-06. Based on these results, we can reject the null hypothesis at the 0.05 level of significance. The null hypothesis assumes that all regression coefficients are zero, which suggests that there is no effect of the independent variables on the explanatory variable. However, due to the low p-value, we have a strong belief that at least one of the independent variables has a significant effect on the explanatory variable.

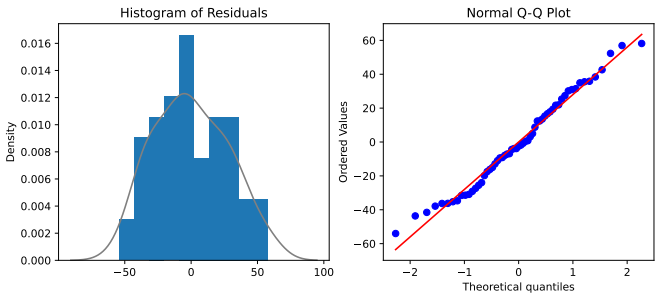
4 Analysis of the model coefficients

$$\begin{aligned} imo2015\_tasks\_points = (\mathbf{119.6}) + (\mathbf{-7.099 \cdot 10^{-12}}) \cdot GDP + (\mathbf{3.050 \cdot 10^{-8}}) \cdot population + (\mathbf{7.167 \cdot 10^{-10}}) \cdot high\_tech\_exports \\ + (\mathbf{1.301 \cdot 10^{-5}}) \cdot migration + (\mathbf{-490.9}) \cdot gov\_educ\_expenditure + (\mathbf{7.179 \cdot 10^{-6}}) \cdot area + (\mathbf{-796.1}) \cdot prob\_of\_death + \\ (\mathbf{-44.02}) \cdot internet\_user\_percentage + (\mathbf{32.02}) \cdot gross\_enrollment\_ratio + (\mathbf{-21.39}) \cdot unemployment\_rate \end{aligned}$$

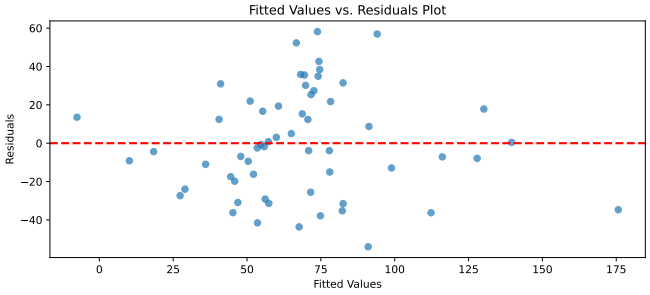
- *Intercept*: The value of the dependent variable when all independent variables are equal to zero. In this case, it is 119.6. It is a statistically significant coefficient (p-value < 0.05), which means it differs significantly from zero.
- *GDP*: The impact of GDP on achievements during IMO 2015. It is not a statistically significant variable because the p-value is greater than 0.05.
- *population*: The impact of population on achievements. It is not a statistically significant variable because the p-value is greater than 0.05.
- *high_tech_exports*: The impact of high-tech exports on achievements. It is a statistically significant variable with a p-value less than 0.05. In this case, the value is 7.167×10^{-10}. Non-mathematical interpretation: If we increase high-tech exports by one unit in a country, we expect the number of points to increase by 7.167×10^{-10}.
- *migration*: The impact of migration on achievements. It is not a statistically significant variable because the p-value is greater than 0.05, suggesting that it does not have a statistically significant effect on the dependent variable
- *gov_educ_expenditure*: The impact of public expenditure (% of GDP) on school development on achievements. It is not a statistically significant variable because the p-value is greater than 0.05
- *area*: The impact of the country's area on achievements. It is a statistically significant variable with a p-value less than 0.05. In this case, the value is 7.179×10^{-6}. Non-mathematical interpretation: If we increase the country's area by 1km^2, we expect the number of points to increase by 7.179×10^{-6}
- *prob_of_death*: The impact of the probability of death in the country on achievements. It is a statistically significant variable with a p-value less than 0.05. In this case, the value is -796.1. Non-mathematical interpretation: If we increase the probability of death in the country by 1%, we expect the number of points to decrease by 7.961
- *internet_user_percentage*: The impact of internet accessibility on achievements. It is not a statistically significant variable because the p-value is greater than 0.05
- *gross_enrollment_ratio*: The impact of the gross enrollment ratio on achievements. It is not a statistically significant variable because the p-value is greater than 0.05
- *unemployment_rate*: The impact of the unemployment rate on achievements. It is not a statistically significant variable because the p-value is greater than 0.05.

5 Further analysis of the model

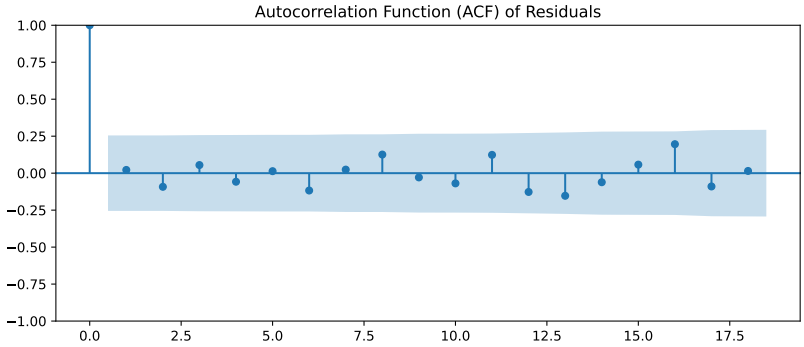
- Multicollinearity: The VIF values for all variables are below 5, indicating either the absence or a minor presence of multicollinearity among the independent variables. However, the correlation matrix reveals a high correlation (absolute values above 0.7) between pairs of variables: "*GDP*" - "*high_tech_exports*" and "*internet_user_percentage*" - "*gross_enrollment_ratio*". To avoid potential collinearity, it might be considered to remove the variables "*GDP*" and "*gross_enrollment_ratio*" (taking into account the correlation between "*gross_enrollment_ratio*" and "*prob_of_death*" as well, which is -0.59) from the model



- Normality of residuals: After checking the histogram and QQ-Plot, it can be observed that the distribution of residuals in the model is very close to a normal distribution. Additionally, the conducted tests for normality (Anderson-Darling Test, Shapiro-Wilk Test) show that the p-values are greater than 0.05. Based on these results, we can conclude that the assumption of normality of residuals is met.



- Homoscedasticity: After performing the Breusch-Pagan test and analyzing the residuals vs. predicted values plot, there is evidence of heteroskedasticity in the model. The p-value of the Breusch-Pagan test is equal to 0.03, leading us to reject the null hypothesis of homoskedasticity of errors in the model. Additionally, the residuals plot does not show a uniform spread around the y=0 line, suggesting the presence of heteroskedasticity.



- Autocorrelation: After performing the Durbin-Watson test, Ljung-Box test, and analyzing the autocorrelation function (ACF) plot of the model residuals, there is no significant evidence of autocorrelation. The p-values of the Ljung-Box test are greater than 0.05, and the Durbin-Watson statistic value of 1.85 (which is very close to 2) suggests no statistically significant autocorrelation. Additionally, the ACF plot shows that most of the model residuals' values fall within the blue shaded region, indicating the absence of significant autocorrelation.

- Linear relationship: After conducting the Rainbow test, we obtain a p-value greater than 0.05, which suggests no significant evidence against the linearity assumption in the model. However, Ramsey's RESET test returns a p-value of 0.017, indicating that there are statistically significant evidence against linearity in the model Based on these results, it can be concluded that there is some evidence of nonlinearity in the model.

  In conclusion, there is some evidence of nonlinearity in the relationships between the explanatory variables and the dependent variable. There is no significant evidence of autocorrelation in the model's errors. There is evidence of heteroskedasticity of residuals. The distribution of errors is close to a normal distribution. VIF values are less than 5, indicating no significant multicollinearity. This analysis suggests the need for further modification or improvement of the model

6 Identification of atypical observations

During the identification of atypical observations in the model, I calculated leverage and Cook's statistics for all observations. The results indicate that observations 27, 32, 46, and 50 have high leverage (exceeding the average leverage value plus two times the standard deviation). Additionally, observations 27 and 46 have high Cook's statistics. However, upon closer examination, there is no clear evidence that these observations are entirely unrealistic. Therefore, I do not see the necessity of removing these observations

7 Reducing insignificant variables

By conducting backward elimination, we obtain a model with the following variables: **high_tech_exports, area, prob_of_death i internet_user_percentage**.
From that we can conclude that the variables *high_tech_exports, area, prob_of_death,* and *internet_user_percentage* are significant in explaining the variability of the dependent variable (*imo2015_tasks_points*). On the other hand, the remaining variables (*gov_educ_expenditure, gross_enrollment_ratio, migration, population, unemployment_rate, GDP*) do not make a significant contribution to the model and can be omitted.

8 Upgrade to the second model

After eliminating insignificant variables, we proceed to the second model:

$$\begin{aligned} imo2015\_tasks\_points = (\mathbf{116.7}) + (\mathbf{5.798 \cdot 10^{-10}}) \cdot high\_tech\_exports + (\mathbf{7.880 \cdot 10^{-6}}) \cdot area + \\ + (\mathbf{-908.9}) \cdot prob\_of\_death + (\mathbf{-45.37}) \cdot internet\_user\_percentage \end{aligned}$$

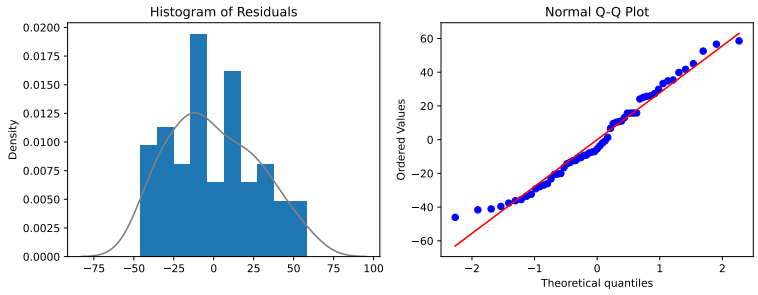
The coefficient values in the second model have not significantly changed compared to the previous model. The main issue with the second model is a slight violation of the assumption of normality of residuals (p-values for the Anderson-Darling test and the Shapiro-Wilk test are 0.025 and 0.049, respectively). To improve this aspect, we proceed to the third model.

9 Upgrade to the third model

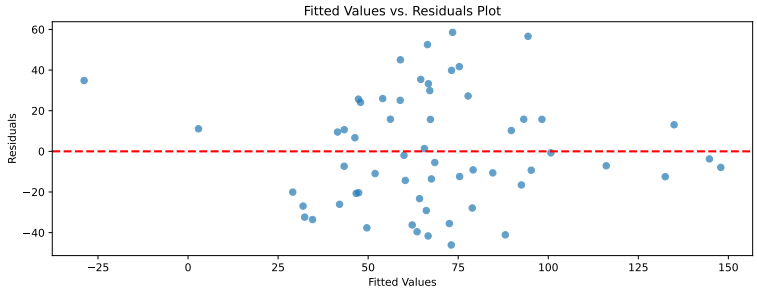
$$\begin{aligned} imo2015\_tasks\_points = (\mathbf{136.4}) + (\mathbf{5.207 \cdot 10^{-10}}) \cdot high\_tech\_exports + (\mathbf{0.02994}) \cdot \sqrt{area} + \\ + (\mathbf{-1024}) \cdot prob\_of\_death + (\mathbf{-28.74}) \cdot e^{internet\_user\_percentage} \end{aligned}$$

All variables are statistically significant; therefore, we proceed to check the assumptions of the model:

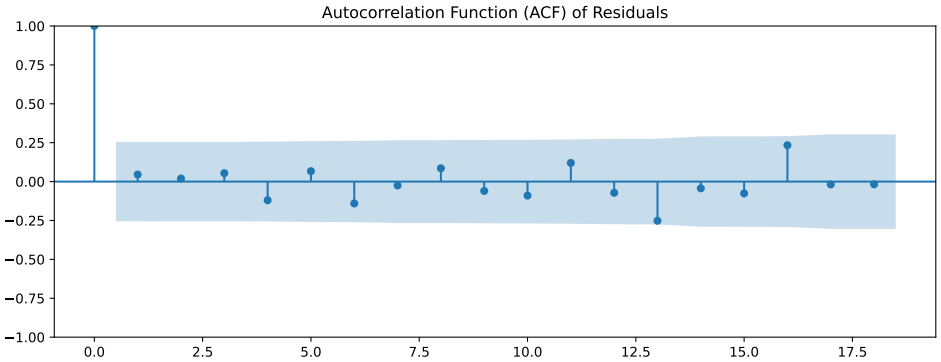
- Multicollinearity: Previously, we did not encounter significant issues with multicollinearity, and the situation remains unchanged in the case of the third model. The VIF values for the explanatory variables in the third model are also less than 5, indicating no significant multicollinearity between the variables



- Normality of residuals: Previously, we did not encounter any issues with the assumption of normality of residuals, and the situation remains favorable now as well. The tests for normality return p-values greater than 0.05(e.g., Anderson-Darling test - 0.15, Shapiro-Wilk test - 0.12). Additionally, the histogram of residuals and the QQ-plot depict a distribution of residuals that aligns with the normal distribution. Based on these results, we can infer that the assumption of normality of residuals is met.



- Homoscedasticity: Previously, we had an issue with heteroskedasticity, indicated by a p-value of 0.03 in the Breusch-Pagan test. However, now the p-value is greater than 0.05 (the p-value of the Breusch-Pagan test is now 0.08), which is a slight but positive difference compared to the value of 0.05. Nevertheless, it is worth noting that the plot of residuals does not exhibit a uniform spread around the y=0 line, suggesting the presence of heteroskedasticity. However, since the p-value of the test is greater than 0.05, we can assume the assumption of homoskedasticity.



- Autocorrelation: Regarding autocorrelation, we did not encounter any problems before, and the situation remains unchanged. The p-values from the Ljung-Box test for autocorrelation are greater than 0.05, indicating no autocorrelation. The analysis using the ACF plot confirms these results, as all autocorrelation values fall within the confidence intervals. Therefore, we can conclude that there is no issue of autocorrelation in the third model.

- Linear relationship: Previously, we observed an issue with the linear structure of the model based on the p-value of the Ramsey's RESET test, which was 0.017, suggesting a nonlinear relationship. However, now the p-value of this test is 0.15. Additionally, the result of the Rainbow test has a significantly higher p-value than 0.05. Therefore, it can be inferred that the model meets the assumption of a linear structure.

  In summary: In some cases, assumptions have improved; for example, the p-value of the Breusch-Pagan test increased from 0.03 to 0.08, which improves homoskedasticity, and the value of the Ramsey's RESET test increased from 0.017 to 0.15, suggesting better adherence to the assumption of a linear structure. In other cases, assumptions did not significantly change, such as in the case of autocorrelation, multicollinearity, and normality of residuals

10 Model comparison

In this section, we will primarily compare the third and first models. We will focus on evaluating several important measures of model performance, such as R-squared, adjusted R-squared, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC).

- We observe that the values of R^2, AIC, and BIC have remained virtually unchanged. For example, the AIC was 579 in the first model and is now 567, and similarly, the BIC has changed from 602 to 577. However, the adjusted R^2 in the third model (0.52) is higher than in the first model (0.45). This indicates that the third model better accounts for the number of independent variables and provides a more adjusted measure of fit to the data.
- For a more comprehensive analysis, I also conducted a comparison with the second model to have a broader understanding of the differences between the three models. However, after examining the AIC, BIC, and adjusted R^2 values, we notice that these results fall between the first and third models.

11 Summary

To summarize this project, I conducted an analysis using a linear regression model to understand how various factors describing the countries participating in the 2015 International Mathematical Olympiad affect the number of points scored. The variables studied included factors such as GDP, population, high-tech exports, number of migrants, percentage of public spending on school development, area of the country, probability of death rate, internet access

rate, enrollment rate and unemployment rate.

Most of the explanatory variables had no significant impact on the outcomes in the Olympiad. By conducting the reduction of non-significant variables, the model was optimized and included variables such as high-tech exports, country area, probability of death, and internet access rate. It turned out, that after applying the square root transformation on the country area and the exponential transformation on the internet access rate, the model performed better in statistical tests and showed improvement in the measure of various evaluation indices, such as AIC, BIC, adjusted R-square. This means that the transformations contributed to a better fit of the model to the data.

Increasing high-tech exports and the country's area is expected to lead to an improvement in the Olympiad results. It is intuitive that greater high-technology exports can have a positive effect on the score, as it indicates a developed science and technology sector, which can influence better mathematics education. In turn, the larger size of the country, as an indicator of economic development, can also positively affect the score, due to greater resources and educational opportunities

On the other hand, when decreasing the probability of death rate, one can expect an increase in the number of tasks solved at the Olympiad. It is worth noting that the coefficient at the exp(internet_user_percentage) variable is negative, which may seem counter-intuitive. This may suggest that, in the case of this particular model, a higher percentage of internet users is not directly related to higher scores on the math Olympiad. However, it is important to be aware that the results described are only part of a larger picture. There are many other factors that may play an important role in participants' achievements, but they were not included in the analyzed model