Redefining Country Influence Through Research Output

Research project for 30549 - “Mathematical Statistics”

Edoardo Ghirardo, Elisa Tofanelli

January 2024

1 Introduction and motivation

How can we measure the impact that a country has on the world? Traditional metrics such as GDP, military strength and political influence seem too limitative for the 21st century. This is why we propose a new measure: the scientific output of a country. Specifically, we are interested in the number of annual articles published every year in scientific and technical journals, adjusted for population.

2 Research question and formulation of the hypothesis

Our aim is to verify whether or not a country’s wellbeing, lifestyle and effort in education can be used to predict its own scientific output, measured in scientific articles per million people. Additionally, we want to check whether countries traditionally regarded as “influential” actually do score higher than the others with our new metric.

3 Datasets

We chose two datasets for each of the three metrics we wanted to use: for wellbeing, we used a country’s life expectancy, as well as self reported life satisfaction; for lifestyle, we used the proportion of people living in urban areas, and of people using the internet; for effort in education, we used the education expenditure as share of GDP and the PISA scores in the Math test.
We conducted our analysis on the most recent year for which all of these metrics were available, which at the time of writing is 2018.
After downloading the datasets, we imported them in R and joined the interesting columns in a single data frame. For some, we had to slightly manipulate the data (see the R code for details).
We decided to focus our analysis on the countries for which all seven datasets were available. Our final dataset consisted of sixty-eight countries, each with one target value and six explanatory values.

4 Visualising the data

We plot separately the relationship between each explanatory value and the target.
Please refer to Figures 1, 2 and 3, for such plots.
From a qualitative analysis of these plots, we noticed that the type of relationship between our target variable and the predictor variables representing PISA Math scores and internet usage appears exponential in nature rather than linear, while the life expectancy appears to have a quadratic relationship. We took this into account in the regression model.

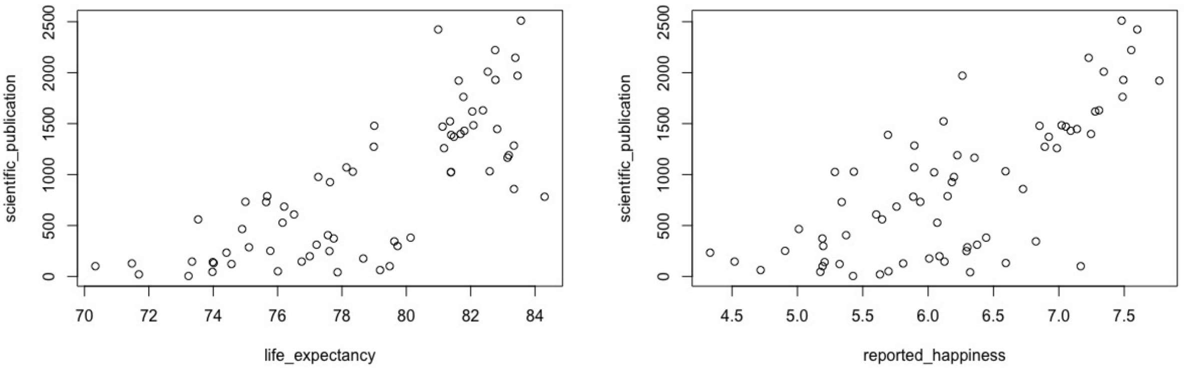

Figure 1: The two *wellbeing* metrics plotted against the target

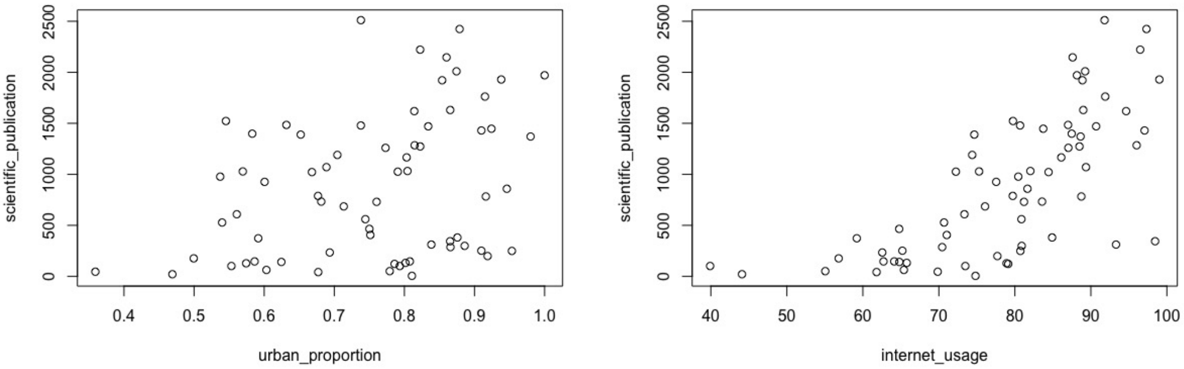

Figure 2: The two *lifestyle* metrics plotted against the target

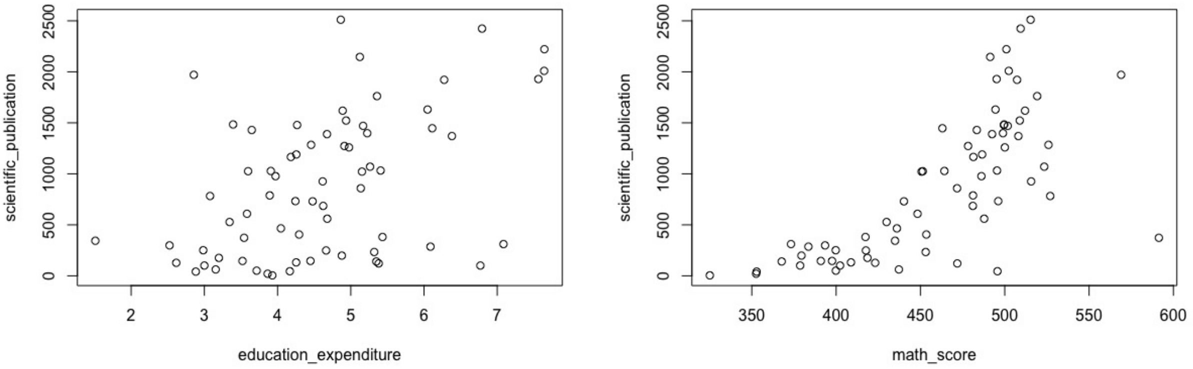

Figure 3: The two *effort in education* metrics plotted against the target

5 Multivariate linear regression

5.1 Mathematical background and fitting the model

We used a multivariate linear regression model to try to predict the scientific output of a country. Based on an observation of our explanatory variables (i.e. life expectancy, self reported life satisfaction, proportion of people living in urban areas, proportion of people using the internet, education expenditure and PISA Math scores), we tried to predict our target variable (i.e. the number of scientific publications per million inhabitants).

Mathematically, we have 68 independent[1] variables $Y_1, \ldots, Y_{68}$, each with a corresponding 10-dimensional[2] predictor vector: $(x_{1,1}, \ldots, x_{1,10}), \ldots, (x_{68,1}, \ldots, x_{68,10})$. Based on our model we can describe

$$Y_i = \sum_{j=1}^{10} \beta_j x_{i,j} + e_i, \quad i = 1, \ldots, 68$$

where $e_i, \ldots, e_{68}$ are independent normally distributed random variables with expectation 0 and variance σ^2.

Assuming that the design matrix X has full rank, the MLE for β is $\hat{\beta} = \left(X^T X\right)^{-1} X^T Y$.

5.2 Model selection

Our goal then was to check if there are models other than the one we had just fit that could suit our data better. In order to do this, we did not use cross-validation, because of the small sample size of our dataset, instead, we performed a test-based step down selection to see if a simpler model could have similar or better results. First, using a t-test, we tested the hypotheses: $H_0 : \beta_i = 0$ versus $H_1 : \beta_i \neq 0$ for all of the parameters β_i. Then, we checked if any of the p-values exceeded $\alpha = 0.05$: if this was the case, we repeated the procedure after having eliminated the covariate with the highest p-value. We repeated this process until all of the covariates had a p-value lower than α.

The result of this procedure is the following: we removed half of the covariates (we are only left with the PISA Math score, its exponential, the life expectancy and the reported life satisfaction), and ended up with an adjusted R^2 which was only marginally smaller: the initial value was 0.7975 while the final value with the simpler model was 0.7898.

6 Testing the hypothesis

We used hypothesis testing to check whether countries traditionally considered as "*influential*", or "*First-world*"[3] actually do score higher than others in our new metric. In order to do so we searched for a reliable division between developed and developing countries. Then we divided the countries of our data according to this division into the two categories. Ultimately, we performed a two sample t-test with:

- H_0: the scientific output is larger or equal in the developing countries
- H_1: the scientific output is larger in the developed countries

Mathematically, we have $X_1, \ldots, X_{36} \overset{i.i.d.}{\sim} N(\mu, \sigma^2)$ and $Y_1, \ldots, Y_{162} \overset{i.i.d.}{\sim} N(\nu, \tau^2)$ independent observations of the scientific output developed and developing countries, respectively. We test the null hypothesis $H_0 : \mu \leq \nu$ against the alternative hypothesis $H_1 : \mu > \nu$.

The test statistic we use is the asymptotic t-test:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{S_X^2/m + S_Y^2/n}}$$

The result of the test was to reject the H_0 which confirms our assumption that countries traditionally seen as developed actually do have a higher scientific output on average.

[1] This is one of the assumptions we make, for details on this and other assumptions we use throughout the study, please refer to the "Limitations of the study" section.

[2] The reason behind the vector being 10-dimensional is the following: we have the six variables from the six datasets, the exponential for two of them, the square for one of them, and the intercept (i.e. 1 as the first predictor variable for each observation).

[3] Please refer to the conclusion and the bibliography to see which definition of *first-world* we used.

7 Limitations of the study

In this section, we conducted tests to assess the assumptions and limitations of our analysis:

- **Assumption of normality of residuals:** we initially examined whether there is significant statistical evidence against the normality of the residuals from our multiple linear regression model. While the results did not provide enough evidence to reject the assumption of normality, it's crucial to note that this doesn't conclusively confirm the normal distribution of residuals. Therefore, our model relies on the assumption of normally distributed residuals, representing the first limitation. See Figure 4, for a normal QQ-plot of this data.

- **Assumption of constant variance of residuals:** continuing our examination of the residuals, it's crucial to assess whether they exhibit consistent variability. This can be determined by examining the QQ plot, which compares the fitted values against the residuals (refer to Figure 5). In our analysis, we don't discover clear evidence in favor of homoscedasticity. Instead, we observe a subtle deviation from the assumption, suggesting a mild violation. Therefore, we highlight the issue of constant variance as the second assumption behind our model.

- **Normality Assumption for Scientific Outputs:** a key aspect of our study involves assuming that the scientific outputs of countries are normally distributed, enabling us to conduct hypothesis testing using paired t-tests. However, the Kolmogorov-Smirnov Test revealed compelling evidence against the normality of the `developing_data` and insufficient evidence to support the normality assumption for `developed_data`. Consequently, relying on the normality assumption for these datasets represents the third limitation of our model. See Figure 6 for normal QQ-plots of this data.

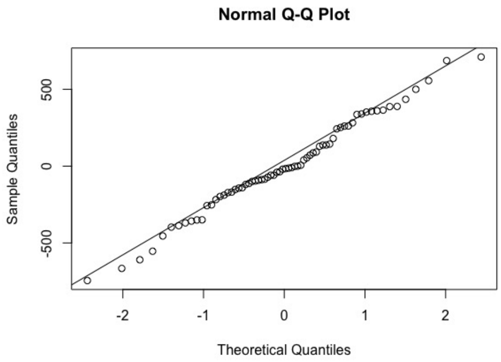

Figure 4: Normal QQ-plot of the e_i

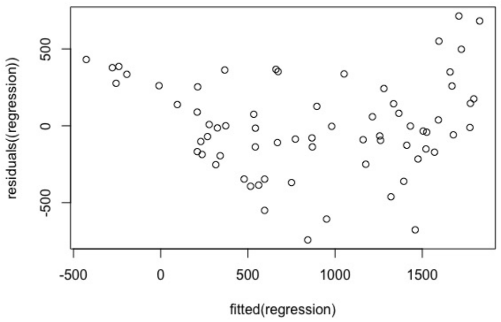

Figure 5: Plot of the e_i

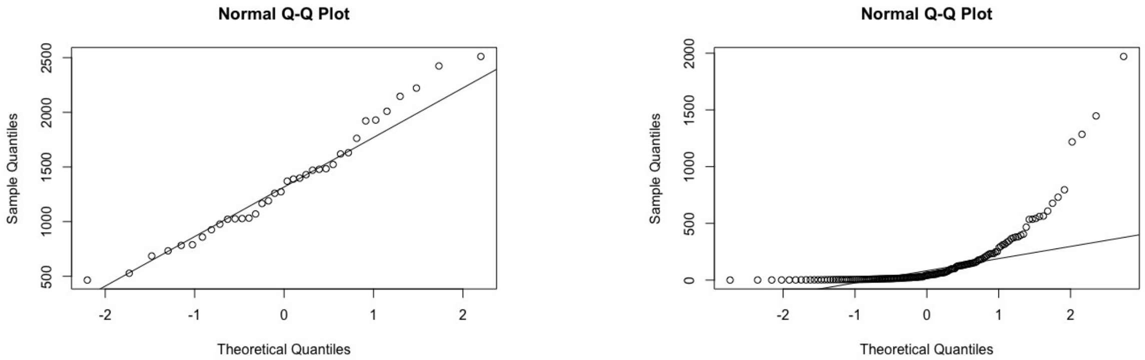

Figure 6: Normal QQ-plots of the developed (left) and developing (right) countries' scientific output

8 Conclusion

The results of our exploratory statistical analysis seem promising and naturally spawn further questions. However, they should not be taken as anything more than they are: a statistical analysis.
One of the questions that naturally comes to mind after our conclusions is *which factors have the strongest impact on scientific output?*. A quantitative analysis capable of providing significant evidence that a certain factor has a higher impact on the total scientific output than others can be extremely useful to policy makers who have to decide how to allocate the government spending of their country.
It is important, however, to remember once again that a statistical analysis is not necessarily due to causation: two factors sharing a high correlation does not necessarily imply that artificially increasing one will cause a significant increase in the other: the correlation might be otherwise explained, for instance because they are caused by a common third factor.
In our findings, we provide sufficient statistical evidence to say that countries traditionally considered as first-world do have a higher mean scientific output per million inhabitants, however this does not necessarily tell us that in order for a country to be "first-world" (which we have also decided not to discuss the definition of, because it would only be the source of a potentially endless political debate), it needs to have a high scientific output: more likely this is just due to countries in the first-world being historically more developed and wealthier, which is associated with being able to afford a higher level of research in all areas of academia.

9 Bibliography

We used https://ourworldindata.org to download the datasets. This website collects verified data from reliable sources and uploads it in easy to work with formats. Specifically, each dataset can be found at:

- Life expectancy: https://ourworldindata.org/life-expectancy
- Self reported life satisfaction: https://ourworldindata.org/happiness-and-life-satisfaction
- Share of urban population: https://ourworldindata.org/urbanization
- Share of people using the internet: https://ourworldindata.org/internet
- Education spending: https://ourworldindata.org/financing-education
- PISA Math score: https://ourworldindata.org/global-education

We used the list of *first-world* countries provided by the United Nations in their *World Economic Situation and Prospects 2014*, available at https://www.un.org/en/development/desa/policy/wesp/wesp_current/2014wesp_country_classification.pdf
The mathematical background we used can be found on *Fetsje Bijma, Marianne Jonker, Aad van der Vaart. An Introduction to Mathematical Statistics. Amsterdam University Press (2018)*