Analyzing Potential Economic Dependencies of the COVID-19 Mortality Rate

Chenxu (Robin) Mao

March 2024

Introduction

The Coronavirus Disease 2019 (COVID-19) is an acute respiratory disease that sparked a recent worldwide pandemic that culminated in the loss of 7 million lives [1] and inflicted an economic downturn, with losses surpassing $8.5 trillion globally [2].

Oppose the existing literature such as *In The Role of Economic Structural Factors in Determining Pandemic Mortality Rate. Goutte et al. (2020)* [3], which uses principal component analysis and analyzes the COVID-19 mortality rate dependencies in different French regional economies with data provided by government institutes; and *Social, Economic, and Regional Determinants of Mortality in Hospitalized Patients With COVID-19 in Brazil. Rodrigues et al. (2022)* [4], which applies logistic regressions to analyze the economic dependencies in the two major Brazilian subregions using private data, this paper will have a broader approach with a research goal of finding the generalized economic dependencies and their quantified correlation w.r.p. to the COVID-19 mortality rate among nations.

Methods

Methods: Variable Selection

Poisson regression is a generalized linear model form of regression designed to predict the count of events in a fixed time interval given the influence of explanatory variables, which in our case are nations' macroeconomic characteristics.

The response variable is the COVID-19 mortality rate of a country, measured in cases per day. The macroeconomic variables that are predictors of a nation's economic performance are population, GDP, GDP per capita, unemployment rate and inflation rate.

We will be using the rate variation of the Poisson regression. Ideally, the final model will possess the following form,

$$\ln(\lambda) = \vec{X}\vec{\beta}$$

Where λ is the mortality rate in cases per day, $\vec{\beta}$ is the coefficient matrix.

A Poisson regression with five variables will be produced after the initial model fitting using all the above variables, referred to as the full model throughout this paper. We will validate the model assumptions using deviance residual vs fitted value plots, DFBETAS plots, etc. and reduce the model for accuracy and interpretability improvement, using BIC as the selection standard as it imposes more punishment on extra parameters. We will validate the assumptions once we have a probable optimal reduced model. If the final model contains only statistically significant parameters, we conclude the correlations between the COVID-19 mortality rate and the reduced parameters.

Methods: Model Diagnostics and Validations

Poisson regression introduces the following assumptions:

1. Independent y_i's
2. Independent errors
3. $\ln(\mu) = \vec{X}\vec{\beta} + \ln(T)$
4. $y_i \geq 0, \forall y_i \in Y$
5. $Y \sim Poisson(\lambda)$
6. $E(Y|x) = Var(Y|x) = \lambda$
7. Lack of influential points
8. Lack of multicollinearity

We will use a scatterplot of deviance residual against covariates to check for assumptions 1 and 2. Once any is violated, we will transform the data. We will use a deviance residuals vs fitted plot to check for assumption 3, upon violation we would have to change the variable selection. Assumption 4 holds since we are using the COVID-19 count dataset. We will use a QQ plot to check for assumptions 5 and 6, upon violation we will commit data transformations. We will use DFBETAS to check for assumption 7 using the ordinary cutoff (DFBETAS $\geq$ 1 or DFBETAS $\leq$ -1) and the Variance Inflation Factor (VIF) for assumption 8 using the $VIF = 5$ cutoff. We will drop any violating observations.

Once all assumptions are met, we will use BIC backward variable selection to reduce the model, stabilize the parameter variance and increase the model readability. Once a final model is proposed, we will use cross-validation to quantify the model performance.

Results

Results: Data Description

The full model is evaluated as below,

Table 1: Full Model

	Intercept	InflationRate	UnemploymentRate	GDPperCapita	GDP	Population
estimate	18.00	-0.005322	0.02404	0.005588	0.00005453	-0.001599
std.err.	0.0000232	0.0000007882	0.000001756	0.0000003740	0.000000003	0.000000134
p-value	insignificant	insignificant	insignificant	insignificant	insignificant	insignificant
VIF	-	1.037776	1.116394	1.299066	1.949170	1.770155

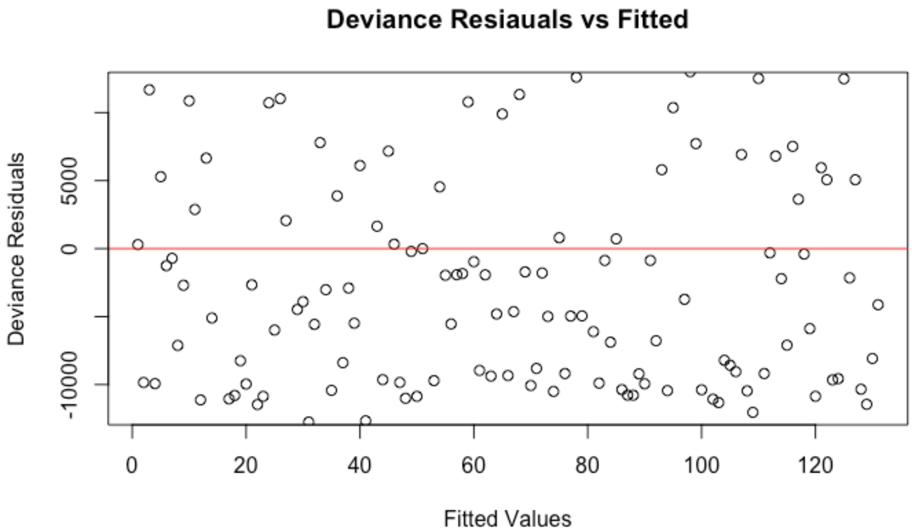


Figure 1: deviance residuals vs fitted values

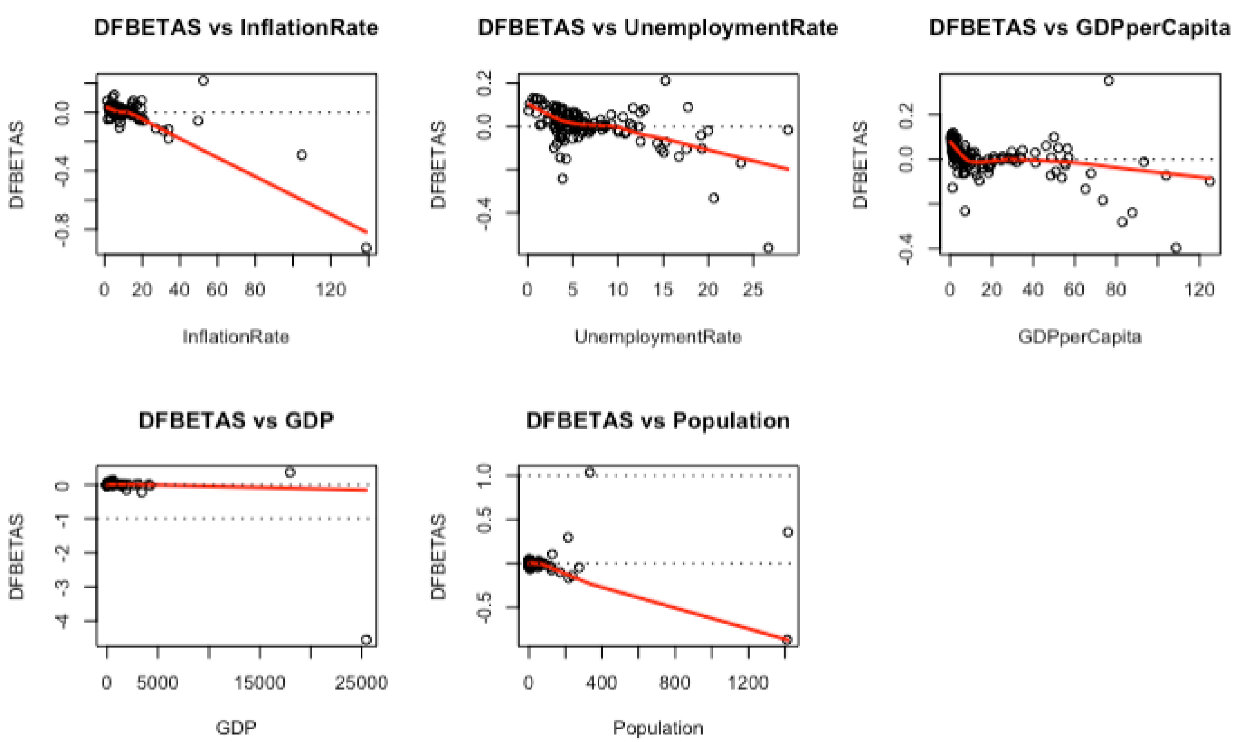


Figure 2: DFBETAS vs predictors in full model

Where Population is measured in millions, GDP per capita is measured in thousands and GDP is measured in billions $USD.

The small p-values and standard error show that all full model parameters are significant. The VIF (Variance Inflation Factor) for each predictor is below 5, which concludes the absence of multicollinearity in the full model.

The deviance residuals vs fitted plot (Figure 1) possess a random spread around 0 without any systematic pattern or structure, which suggests a good model fit that indirectly supports the linearity assumption. The DFBETAS vs predictors plots (Figure 2) suggest the existence of influential points on high values for all predictors using the general cutoff (DFBETAS $\geq$ 1 or DFBETAS $\leq$ -1), which violates the assumption of no influential points. These influential observations are removed.

Moreover, computation shows a significant inequality between the expected value and the variance, which suggests an overdispersion in the model. We will try switching to a negative binomial or data transformation to fix this issue.

Results: Analysis Presentation

To solve the overdispersion problem, both trials of negative binomial fitting yield infeasible models with significant p-values on all parameters (see Appendix 1 and 2, where the Appendix 1 model has only two significant parameters and the Appendix 2 model has none). Then we tried applying cox-box transformation with optimized lambda on the mortality rate, which fixed the severe overdispersion issue. The Coxbox optimal lambda is selected to be 0.1818182 from the log-likelihood of the Coxbox function on the Poisson model (see Appendix 3) without the influential points. Note that the Coxbox transformation is applied as,

$$y_i = \frac{y_i^\lambda - 1}{\lambda}$$

Here y_i denotes the mortality rate and $\lambda = 0.1818182$.

After further research, we found that R has problems computing likelihood, AIC and BIC for Poisson regressions [5]. Thus we apply cross-validation of $k = 4$, randomly divide the original dataset into four subsets, commit model fitting on the first subset and calculate the aggregate MSE (Mean Square Errors) on the other three subsets. The final model is selected with the lowest aggregate MSE and without parameter redundancy.

Discussion

Discussion: Final Model Interpretation

The above algorithm drops Population as a predictor and yields the following result:

Table 2: Final Model

	Intercept	InflationRate	UnemploymentRate	GDP	GDPperCapita
estimate	4.650	-0.00287	0.01449	0.00009237	0.004426
std.err.	0.03529	0.000824	0.00326	0.0000241	0.000696
p-value	insignificant	insignificant	insignificant	insignificant	insignificant
VIF	-	1.082148	1.149510	1.267688	1.335291

With an aggregate MSE of 49394.06 over the three fitting subsets, an average MSE of 16131.35. The model is given as,

$$\ln\left(\frac{\lambda_i^{0.1818182}-1}{0.1818182}\right) = 4.650 - 0.00287\text{InflationRate} + 0.01449\text{UnemploymentRate} + 0.00009237\text{GDP} + 0.004426\text{GDPperCapita}$$

Where GDP per capita is measured in thousands \$USD and GDP is measured in billions \$USD. The mortality rate is measured in cases over the entire pandemic duration. To improve the readability, we rearrange the final model as,

$$\begin{aligned}
\frac{\lambda_i^{0.1818182}-1}{0.1818182} &= \exp(4.650 - 0.00287\text{InflationRate} + 0.01449\text{UnemploymentRate} \\
&\quad + 0.0000923\text{GDP} + 0.004426\text{GDPperCapita}) \\
\lambda_i^{0.1818182} &= 0.1818182\exp(4.650 - 0.00287\text{InflationRate} + 0.01449\text{UnemploymentRate} \\
&\quad + 0.0000923\text{GDP} + 0.004426\text{GDPperCapita}) + 1 \\
\lambda_i &= \log_{0.1818182}[0.1818182\exp(4.650 - 0.00287\text{InflationRate} + 0.01449\text{UnemploymentRate} \\
&\quad + 0.0000923\text{GDP} + 0.004426\text{GDPperCapita}) + 1]
\end{aligned}$$

A model performance visualization is provided (Figure 3). The parameters show for each unit increase in the inflation rate, the nation's COVID-19 mortality rate tends to have a 1.00168 increase on average. Each unit increase in the unemployment rate tends to yield an increase of 0.9915 cases in the mortality rate over the pandemic. Note that assumptions and parameter redundancies are checked throughout the selection process as we select only valid models.

The final model concludes positive correlations between the COVID-19 mortality rate and the inflation rate, unemployment rate, GDP, and GDP per capita. The final model answers the research question as there exist generalized, statistically significant, positive correlations between the COVID-19 mortality rate and inflation rate, unemployment rate, GDP and GDP per capita.

Discussion: Limitations

Figure 3 visualizes the final model's prediction accuracy over the entire dataset of both fitting and training. We can see that the model's lower bound does not match the actual spread, which is a negative result of the Boxcox transformation. This mismatch results in a decrease in the model accuracy, especially over a certain input domain. We failed to fix this as we cannot use any transformation beyond this course.

Moreover, the model interpretability could be improved. The final model takes a complicated form because we used six datasets from different sources, and the initial merged dataset is intended for us to perform time-series analysis as offset terms. This loss in readability makes the final model harder to understand and use. We failed to adjust for this issue as the corresponding contents cannot be used.

Moreover, we failed to quantify the model performance using course-focused content (we used cross-validation instead) due to R's inconsistency on AIC and BIC computation on Poisson regression models [5]. R's inconsistency in the negative binomial model, such as different libraries (e.g. MASS vs. glmmTMB, see Appendix 1 and 2) produced significantly different parameters, resulting in additional model complexity.

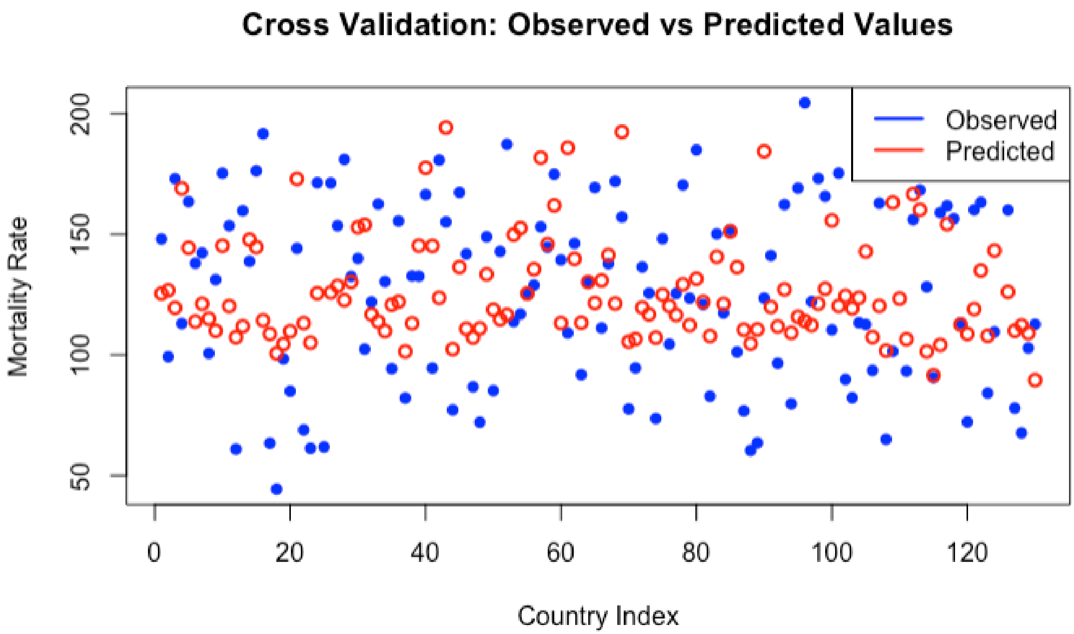


Figure 3: Final Model Cross Validation Result

Appendix

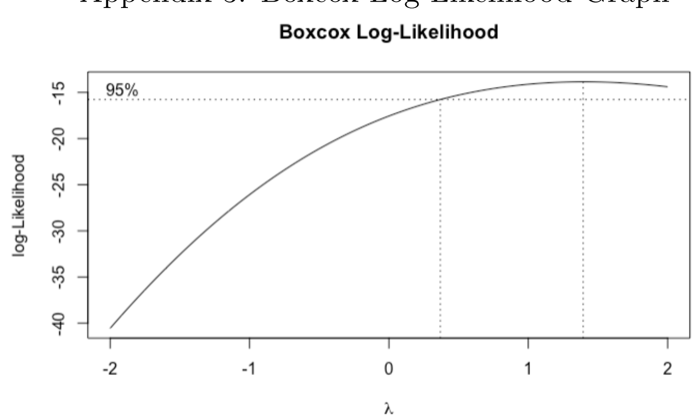
Appendix 1: Negative binomial model using MASS

	Intercept	InflationRate	UnemploymentRate	GDPperCapita	GDP	Population
estimate	4.684	-0.001859	0.01536	0.004288	-0.0000072	0.0001437
std.err.	0.05251	0.0016	0.0048	0.0011	0.000024	0.00023
p-value	insignificant	significant	insignificant	insignificant	significant	significant
VIF	-	1.055313	1.080359	1.184659	2.621268	2.608819

Appendix 2: Negative binomial using glmmTMB

	Intercept	InflationRate	UnemploymentRate	GDPperCapita	GDP	Population
estimate	0.7522	0.2394	0.2516	1.9113	0.9703	0.6106
std.err.	8450	376.77	132.9587	1071.4095	2946.0952	477.1791
p-value	significant	significant	significant	significant	significant	significant

Appendix 3: Boxcox Log-Likelihood Graph



References

[1] World Health Organization. "COVID-19 Deaths — WHO COVID-19 Dashboard." World Health Organization, 2024, data.who.int/dashboards/covid19/deaths?n=c.

[2] United Nations. "COVID-19 to Slash Global Economic Output by $8.5 Trillion over next Two Years." United Nations, 2020, www.un.org/en/desa/covid-19-slash-global-economic-output-85-trillion-over-next-two-years.

[3] Goutte, S., Peran, T., & Porcher, T. (2020). The role of economic structural factors in determining pandemic mortality rates: Evidence from the COVID-19 outbreak in France. Research in International Business and Finance, 54, 101281. https://doi.org/10.1016/j.ribaf.2020.101281

[4] Rodrigues, W., da Costa Frizzera, H., Trevisan, D. M. de Q., & Prata, D. (2022, March 9). Social, economic, and regional determinants of mortality in hospitalized patients with COVID-19 in Brazil. Frontiers. https://www.frontiersin.org/journals/public-health/articles/10.3389/fpubh.2022.856137/full

[5] https://stackoverflow.com/questions/38859906/why-is-the-likelihood-aic-of-my-poisson-regression-infinite. This is normally not citable but is here to support the R AIC & BIC computation failure.