Analyzing Human migration with Logistic Regression

Aviyan Khadka

December 8, 2022

Introduction

Humans have migrated from one place to the other for thousands of years and due to this migratory tendency, societies and civilizations have been established all over the world. From continental migration to city-wise migration, people move for a variety of reasons. In ancient times, people migrated for food security, security from predators, climate security, and for many other factors. However, in the modern day, the world has changed a lot. Many new aspects of the modern world and changes in human society have made migratory patterns very different from the ancient world. Modern society has been transformed into a nation-state structure and in this nation-structure, different metrics are used to measure the success or failure of that state. In this project, we aim to understand the human migratory trend through such metrics using statistical modeling.

Methodology

The Logistic model was chosen for this analysis. Logistic regression is a type of generalized linear model where the response is considered a binary variable. Moreover, generalized linear models are a generalization of ordinary linear regression. Unlike ordinary linear regression where the response is considered normally distributed, the GLM considers the response as a part of the exponential family. The response is taken as the linear combination of the explanatory variables, and the linear combination is formed through link functions that links these variables. The single parameter exponential family is given as,

$$f(y) = exp\left\{\frac{\theta y - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

where,
θ is the canonical parameter of interest
ϕ is a nuisance paramter
a(),b(),c() are known functions.

and,
μ=EY=b' (θ)
σ^2=var(Y)=b$''$(θ)a(ϕ)

Under the logistic model, a log link function(also called logit) is used to link the probability of the outcome to the explanatory variables.
From the link function,

$$\ln\left(\frac{p}{1-p}\right) = X\beta$$

$$\text{or, logit(p)} = X\beta$$

From the link, we can obtain the logistic function,

$$f(x) = \frac{1}{1 + e^{-x\beta}}$$

Where, $x\beta$ is the linear combination.

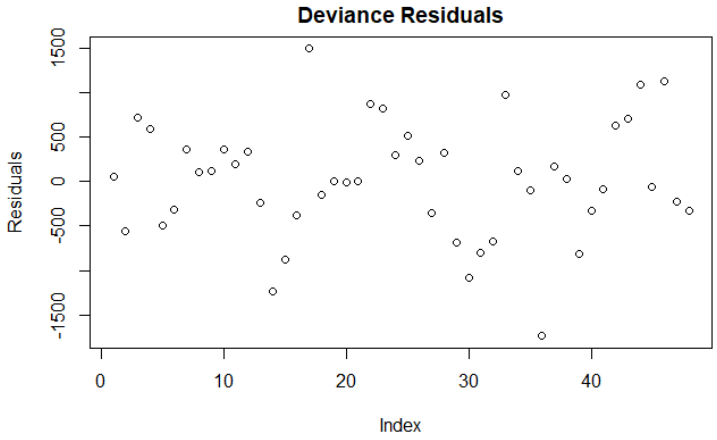
The dataset consists of population and economic data are taken from 48 randomly chosen countries from different geographic locations. There are 7 covariates that are taken, of which 6 are continuous and 1 is a categorical variable. The 6 continuous covariates are GDP growth rate, inflation, unemployment rate, crime rate, poverty rate, and terrorism index. The categorical variable is stability measured at 3 levels (Stable, Moderately Stable, and Unstable). The total number of people that have left the country and the people coming in are taken as the binomial count. The data has been taken from 2015. The goal is to understand if the logistic model is appropriate for the data and estimate the effects of these covariates on the migration trend.

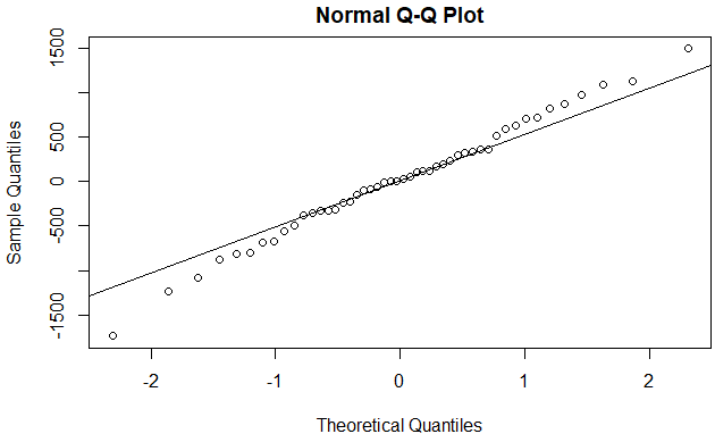
Analysis

The first step of our analysis was to find a suitable model where the relevant explanatory variables are included. The criteria used to select the model was the Akaike Information Criteria (AIC). AIC measures the quality of a model relative to other fitted models. Under this criteria, we found a model with 16 terms. The model consists of 8 main effects and 8 interaction terms.

	Estimate	**Std. Error**
Intercept	-1.392593811	7.920146e-04
GDP growth rate	0.027049698	5.277723e-05
inflation	-0.017608598	3.303821e-05
Stability(Stable)	-1.870229141	1.052089e-03
Stability(Unstable)	1.146201591	2.458166e-03
unemployment	-0.044552724	4.699206e-05
crime rate	0.008009467	1.127354e-05
Poverty	-0.011891014	7.241488e-06
Terrorism	-0.260452328	1.204001e-04
Stability(Stable)*unemployment	0.011733210	1.234633e-04
Stability(Unstable)*unemployment	-0.081272844	8.686715e-05
Stability(Stable)*crime rate	0.629760807	4.282251e-04
Stability(Unstable)*crime rate	-0.063271681	2.347948e-04
Stability(Stable)*Poverty	0.059645872	6.175409e-05
Stability(Unstable)*Poverty	0.007561753	1.856161e-05
Stability(Stable)*Terrorism	-0.036865829	3.108233e-04
Stability(Unstable)*Terrorism	0.127057471	3.361436e-04

The AIC for this model was 19983765. Moreover, from the residual deviance plots and the normal quantile plot, the model does not show any wrongdoings. Even though we could not obtain much information from the deviance residual plot, the normal quantile plot shows us that the residuals are approximately normally distributed without any major outliers.



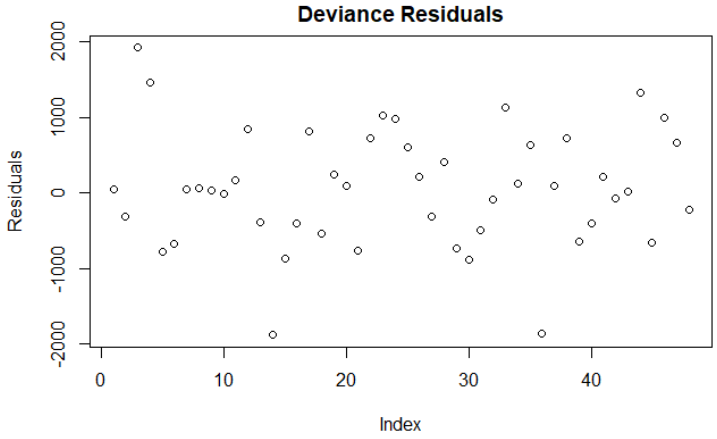


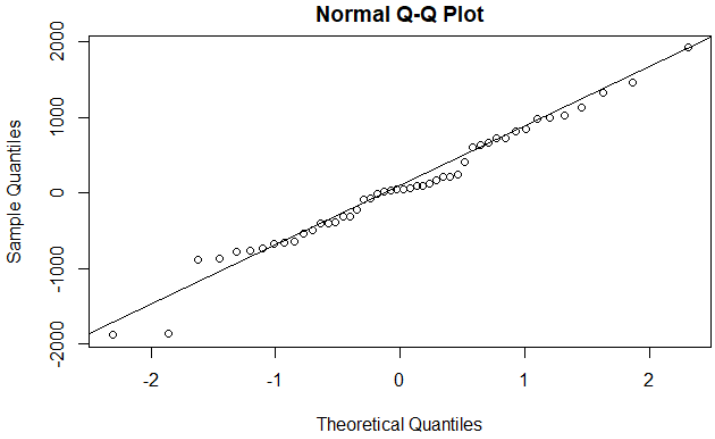
Furthermore, the deviance residuals for our model was 19983019 with 31 degrees of freedom. Such high deviance values leads us to suspect possible overdispersion in the model. Therefore, the quasibinomial model was fit to account for the overdispersion. This quasibinomial model has a higher standard error and has resulted in fewer significant variables. In compared to the previous model, the quasi model has only 3 significant explanatory variables (GDP growth rate, stability(stable), and Terrorism). Now, a new quasimodel has been fitted with these significant variables. Additionally, three more variables with low p values, in which two interaction terms have also been included in the new model. So, our new model gives us the following estimates,

	Estimate	**Std. Error**
Intercept	-2.55076544	0.330711877
Stability(Stable)	-0.82065746	0.497827924
Stability(Unstable)	1.64127019	0.550341129
crime rate	0.01686648	0.008220074
Terrorism	-0.21474568	0.071521848
Stability(Stable)*crime rate	0.59499829	0.366086511
Stability(Unstable)*crime rate	-0.19899188	0.069993182

Finally to confirm that this reduced model is better than the larger quasi model, we have done an ANOVA F-test. The F-test results in a p values of 0.2671 and thus we may conclude that the reduced quasi model is better than the full quasi model.

Looking at the residuals for our reduced model, the quantile plot is approximately along the standard normal line.





Conclusion

From our final model, we can now interpret the coefficients and answer our scientific question. However, first, we shall look into the adequacy of our model. While fitting the logistic model, we got an extremely high deviance residual value. Even after accounting for overdispersion, this may indicate that the logistic fit might not have been an appropriate model to select for this problem. Moreover, the obtained coefficients may not have been accurate estimates. They oppose the general intuition regarding general migratory patterns. For example, looking at the stability(stable) coefficient, there is a 56% decrease in migration when a country changes from fully stable to moderately stable. This discrepancy suggests there might be a problem with the model. Therefore, a more complex model may be needed to answer our question of interest.

References

1. Venkatesan, G., amp; Sasikala, V. (1970, January 1). A statistical analysis of migration using logistic regression model: Semantic scholar. undefined. Retrieved December 6, 2022, from https://www.semanticscholar.org/paper/A-Statistical-Analysis-Of-Migration-Using-Logistic-Venkatesan-Sasikala/7c537b00a975b846cfffdefe652dc7b8ce84d477
2. Our World in Data
3. World Bank
4. Global Terrorism Index