MODELING INFECTIOUS DISEASE: A CASE STUDY OF EBOLA VIRUS

Ayomide A. Ajiboye Princess Nadia Sarpong Saviour Avorgah

Purdue University Fort Wayne, Department of Mathematics

Abstract

In this work, we analyzed by means of linear regression models the relationship between the number of Ebola cases and their total deaths in some African countries (Guinea, Liberia and Sierra Leone). The models for each country is observed to differ from the other. We also investigated whether or not we can use the linear models to establish these relationships.

Keywords: modeling, ebola virus, linear regression, cases, deaths.

Introduction & Background

Ebola is a severe, infectious and frequently fatal disease that affect humans and other non-human primates. It is caused by ebolavirus. Its origin is unknown. However, Scientists believe that it is animal-borne and most likely comes from bats, which transmit the Ebola virus to other animals and humans. Ebola causes fever, pain, diarrhea and bleeding. Since its discovery in 1976, the majority of cases and outbreaks of Ebola Virus Disease (EVD) occurred in 2014-2016 in West Africa mostly in Guinea, Liberia and Sierra Leone.

Symptoms of Ebola can start two to twenty-one days after being infected by the virus. They most often start about eight to ten days after being exposed to the virus. The first symptoms are similar to the common flu. Early symptoms include; fever, chills, weakness, severe headaches, muscle aches. This makes it difficult to diagnose Ebola in someone who has been infected for only a few days. Healthcare providers use lab tests like Polymerase Chain Reaction (PCR), Antigen- Capture Detection Tests, etc. to help diagnose Ebola. It may take up to three days for the Ebola Virus to reach levels that lab tests can detect. There are currently two treatments approved by the Food and Drugs Administration (FDA) to treat Ebola Virus Disease. They are Inmazeb and Ebanga

Data Sources and Methods

The case counts by country were taken from the CDC (Center for Disease Control and Prevention), Ebola (Ebola Virus Disease), beginning with first outbreak which occurred in the Democratic Republic of Congo (formerly Zaire) in a village near the Ebola River, which gave the virus its name. The analysis will be performed using the software R and its packages. All data is publicly available and code is available and can be found here: `https://www.cdc.gov/vhf/ebola/history/2014-2016-outbreak/case-counts.html`.

In regression analysis, an attempt is made to account for the variation of the independent variables in the dependent variable synchronically (Unver and Gamgam, 1999). Regression analysis model is formulated as in the following;

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_n x_n + \epsilon \tag{1}$$

where y is the dependent variable, x_i represents the independent variables, β_i represents the parameters and ϵ is the Error.

References

Centers for Disease Control and Prevention, EBOLA Response. Africa Ebola cases and deaths by countries over time,`https://www.cdc.gov/vhf/ebola/resources/index.html`.

Minnesota Department of Health: Ebola Print Materials `https://healthstate.mn.us/diseases/ebola/basics.html`.

Cleveland Clinic: Ebola Virus Disease `https://https://my.clevelandclinic.org/health/diseases`

Analysis and Results

In this section, we will perform some analysis on our data by looking at correlation coefficients, scatter plots that shows the relationships between the variables, and also build linear regression models to better understand the relationship between the variables.

Here, we study the relationship between the total deaths (response variable), the total number of cases (predictor variable) and the year (predictor variable) for each country under study. We consider the variable "year" to be a factor which takes on 0 or 1 instead of considering it as a numerical variable.

Figure 1 shows the relationship between the total number of deaths and the total number of cases for each country and as we can see, there appears to be a strong positive correlation between these variables and this can be confirmed by looking at the correlation coefficients.

The correlation coefficient between the total number of deaths versus the total number of cases for Guinea is 0.9989, the correlation coefficient between the total number of deaths versus the total number of cases for Liberia is 0.9975, and the correlation coefficient between the total number of deaths versus the total number of cases for Sierra Leone is 0.9932

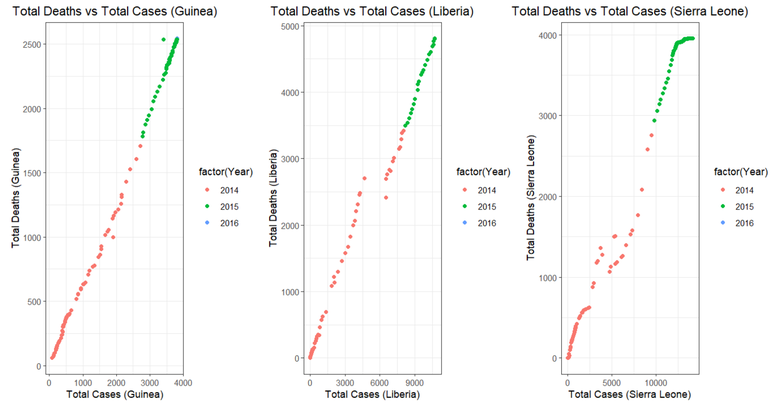

Figure 1. Plot of the total deaths versus total cases for each country

Future Research

Given that the data used for this exercise was relatively small, it will be good to have a larger data set so that we may be able to split it into training and testing sets. The training data set would be used to build the models and the test data set would be used to validate the models. It is recommended that as more data become available, a fraction of the data that was not used in building the model is used to validate the regression models.

Moving forward, we would consider additional predictive variables and observe how this affects our model generally. Note that the model only works for the specified years (2014 -2016) and thus will fail outside these boundaries. Thus, an attempt to make predictions for later years fails.

Discussions & Conclusions

	Model 1
(Intercept)	17.23^{**}
	(5.40)
Total.Cases..Guinea	0.61^{***}
	(0.00)
Year2015	195.44^{***}
	(13.48)
Year2016	208.17^{***}
	(16.66)
R^2	1.00
Adj. R^2	1.00
Num. obs.	265

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 1. Statistical models

Thus, the regression model equations for the three countries are:
For Guinea;

$$Y_1 = 17.23 + 0.61X_{1G} + 195.44X_2 + 208.17X_3$$

For Liberia;

$$Y_2 = 73.97 + 0.43X_{1L} + 83.87X_2 + 102.13X_3$$

For Sierra Leone;

$$Y_3 = 119.32 + 0.22X_{1S} + 882.07X_2 + 742.37X_3$$

where Y_1, Y_2, Y_3 are the predicted total deaths in Guinea, Liberia and Sierra Leone respectively. X_{1G}, X_{1L}, X_{1S} are the total cases in Guinea, Liberia and Sierra Leone respectively. X_2 is factor for Year 2015 (which takes the value 1 when year 2015 is being considered else it takes 0), X_3 is factor for Year 2016 (which takes the value 1 when year 2016 is being considered else it takes 0). Thus, to consider the year 2014 for study, X_2 and X_3 both take on values of 0.

To better understand the regression models presented above, let us consider the regression model for Guinea. Suppose we wanted to predict how many deaths could have occurred in the year 2015, if there were 10000 cases reported in that year, the prediction becomes:

$$Y_1 = 17.23 + 0.61(10000) + 195.44(1) = 6312$$

We found evidence that there is indeed a relationship between the number of Ebola cases and the total number of deaths. This relationship is linear. As the number of Ebola cases increases, the total death cases increases and as the former decreases, the later decreases. The regression model analysis described is both descriptive and predictive.

From the results we obtained in Section 1 above, the R^2 value for the regression model of Guinea is 0.998, that of Liberia is 0.9952, and that of Sierra Leone is 0.9930. These values tells us the proportion of variance in the dependent variable (number of cases and year) that can be explained by the independent variable (number of deaths). In other words, the R^2 values tells us how well the data fits our regression models. This means that it is possible to use a linear regression model to predict the total number of deaths given the total number of cases.

Ayomide A. Ajiboye, ajibaa02@pfw.edu Princess Nadia Sarpong, sarpp01@pfw.edu Saviour Avorgah, avorsk01@pfw.edu