Analyzing Excess Mortality in Puerto Rico After Hurricane Maria

Abstract

This report investigates excess mortality in Puerto Rico in the aftermath of Hurricane Maria by comparing statistical models on mortality data from the excessmort package. Additionally, a comparison of data from New York Times with the excessmort package was performed to ensure data quality. Excess mortality, defined as the observed minus predicted mortality rates for a given time period, provides a more accurate measure of the lasting impact of a natural disaster than direct deaths alone. The analysis includes data cleaning, exploratory data analysis, and model building using linear, log-linear , and negative binomial regression techniques to address overdispersion in mortality counts. The negative binomial model demonstrated superior fit, as indicated by lower AIC and improved residual diagnostics. It produced more stable estimates across age groups and sexes and better captured the effects of Hurricane Maria in terms of true mortality attributable to the event. Comparisons between the two data sources showed discrepancies, particularly due to missing NYT data in late 2017, reinforcing the importance of data completeness. Overall, this project emphasizes the importance of robust statistical modeling and data validation in assessing the public health impact of natural disasters and provides recommendations for improving mortality prediction for future adverse events.

Introduction

Hurricane Maria struck Puerto Rico on September 20th of 2017, with the category four storm leading to widespread infrastructure damage and 64 official deaths. However, the true storm-related death count is suspected to be much higher, with the Puerto Rican government eventually acknowledging that the count may have been more than 20 times the originally reported figure (Robles, 2018). Excess mortality is an established metric in epidemiology, with excess deaths defined as the difference in the observed versus expected number of deaths, both direct and indirect, for a given event based on historical trends (National Center for Health Statistics, n.d.). Excess mortality is key to the discrepancy seen between the original and updated death counts. Comparing the observed death rates in the last four months of 2017 to those in the same time frame in previous years showed that the Puerto Rican community continued to experience adverse effects after the storm had passed.

Combined with a general downward trend in the size of Puerto Rico's population, the drastic increase in excess mortality after Hurricane Maria indicated that the death count initially publicized did not accurately assess the storms' impact. Without reliable methods to assess the impact of major events like Hurricane Maria, affected areas may not receive the appropriate aid and recovery funds needed to rebuild. Therefore, it's critical to establish methods to identify periods of excess mortality when they occur and to assess the lingering impacts of natural disasters on infrastructure, tourism, and access to medical care for the impacted communities.

Several methods were used to assess excess mortality after Hurricane Maria beyond death certificate identification. A survey of 3299 households across Puerto Rico yielded a total of 4645 estimated excess deaths from September to December of 2017 (Kishore et. al., 2018). Other analyses were more conservative, with an independent assessment by George Washington University reporting 1191 excess deaths between September 2017 and February of 2018 (Santos-Burgoa et. al., 2018) and 1139 excess deaths reported by researchers from Penn State and University of Texas from September to November (Santos-Lozada & Howard, 2018). Contrary to other analyses, the authors reported that the death count in December of 2017 had returned to a normal level considering historical variation. The differences in reported counts between these reports shows the subjectivity in statistical approaches to quantifying excess deaths, indicating a need for comparison among them.

The extreme results from Kishore et. al.'s survey-based methodology came with large confidence intervals, indicating uncertainty about their estimate. They had also underestimated pre-hurricane mortality, demonstrating the importance of data sources used in the statistical approach and the high error rates that prediction from survey methods can have (Santos-Lozada and Howard, 2018). Other approaches used generalized linear and time series models successfully to predict excess mortality rates (Santos-Burgoa et. al., 2018; Acosta and Irizarry, 2023).

There have been many models proposed to predict periods of excess mortality for natural disasters like Hurricane Maria, but few analyses have explicitly contrasted multiple statistical approaches on the same dataset. Given the importance of the data source in the estimates and precision of excess mortality, a comparison of the information contained in multiple data sources is also warranted. This analysis aims to explore the trends in the Puerto Rican population contained in the excessesmort package from Acosta and Irizarry, propose a series of statistical models to estimate excess mortality, and finally to compare the data contained within the excessmort R package to the data from the government of Puerto Rico's Health Department published by the New York Times (Irizarry et. al., 2025).

Methods

All subsequent analyses were performed using R version 4.4.1. Data used in this analysis were sourced from the Irizarry and Acosta's excessmort package and from the demographic registry of the Government of Puerto Rico's Department of Health via the New York Times. The puerto_rico_counts dataset from excessmort contains information on daily mortality in Puerto Rico from 1985 to 2022 with covariates for age group, date, sex, population size, and death count.

The New York Times data was stored in pdf format. Text was extracted from the pages using the pdftools package, and formatted using regular expressions to identify the columns containing daily death counts for the days of each month in 2015-2017 (Ooms, 2025). All charts were dropped for simplicity. The excessmort data was aggregated from daily into weekly death counts by taking the sum across days, and a death rate variable was computed by dividing the total reported deaths by the mean of the population for the given week. Week and year were converted to their corresponding MMWR (Morbidity and Mortality Weekly Report) values for consistency Centers for Disease Control and Prevention, n.d.). The data was then further grouped by age group and sex after collapsing age group into 3 categories based on similarity in weekly death count trends over time (See Supplemental Figure 1).

After data was aggregated, exploratory data analysis was performed to assess the composition of Puerto Rico's population and the trends in mortality rates over time. Trends over time were stratified both by sex and age to see how each covariate interacted with death rate separately as well as jointly. Lastly, we aggregated death rates across years to examine the effect of MMWR week on reported death counts (See Supplemental Figure 2).

A training dataset was then created for the models by limiting the data to only 2017 and earlier. A linear regression model was first fit with death rate as the outcome and covariates for population, week, age, and sex as motivated by our exploratory data analysis. It is worth noting that this model assumes normality in the outcome and therefore doesn't place a restriction on the allowable predicted values for rate (Mordkov, 2000).

Next, a generalized linear model was fit using a log link to account for the count-based nature of our outcome using the same covariates as the linear model. An offset term for population size was used to convert the interpretation of the systematic component of the model from a count to a rate for a given group. This allows us to compare our outcome easily between models and groups with different population sizes. Overdispersion was assessed for by taking the ratio of the residual deviance to the degrees of freedom (Penn State Eberly College of Science, n.d.).

Finally, a negative binomial model was fit using the MASS package (Venables & Ripley, 2002). This can be thought of as an overdispersed version of the poisson model, which is useful in this case to produce similar estimates to the previous model but with larger standard errors given that the variation of death count may be greater than the mean (Santos-Burgoa, 2018). The covariates included in this model included all those previously fit plus an additional interaction term between sex and age group. This was motivated by the plot of observed versus predicted mortality rates, where two distinct bands can be seen within each age stratum, showing that the effect of age group on death rate may change with sex (Figure 2). The fit of the poisson and negative binomial models was evaluated on AIC.

These models were then used to generate an expected death rate based on data before 2017 to use as a historical baseline for excess mortality calculations. Excess mortality was calculated as the difference in predicted model value versus observed death rate. Trends in excess mortality were graphed from 1985 to 2017 to examine the dataset for any periods with higher expected mortality rates than expected, and predictions were recomputed without these periods. Finally, the models trained on the historical data were used to predict death rates for 2017 and 2018 to calculate excess mortality relevant to Hurricane Maria.

Lastly, we performed a comparison of the data in the excessmort package and the New York Times PDF using different scores plotted over time. Excessmort data was reaggregated to remove the sex covariate which was not captured in the Department of Health's data. The difference variable for each data source was calculated by subtracting the death count for each day of 2017 minus the death count of 2016. The differences between the two data sources in each of these scores was plotted over time to allow for convenient assessment of agreement between them.

Results

Exploratory data analysis in Figure 1 showed that from 1985 to 2022, population size for young people declined while population size for older people grew. Additionally, there was an overall decrease in mortality rates over time which could be a possible explanation as to why population sizes increased for older people.

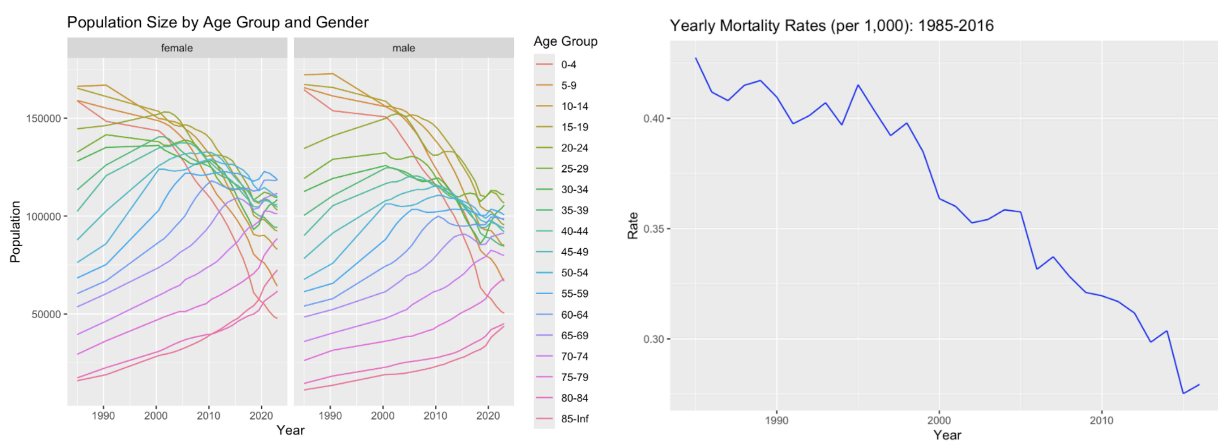

Figure 1: Population and Mortality Over Time by Age and Sex

To find the model that best predicts mortality rate and compare excess mortality, linear, log-linear, and negative binomial models were fitted. In Table 1, when fitting the linear model, the estimated mean mortality rate from 1985 to 2016 was 0.0149 with a standard deviation of 0.0000491. However, when checking for model diagnostics, the normality assumption and constant variance assumptions were very clearly violated (See Supplemental Figure 3).

Given that the linear model violated assumptions, a log-linear model with mortality rates following a poisson distribution was fitted. The log linear model yielded an AIC of 125,630 and had a residual deviance of 49,066 on 13,354 degrees of freedom. The ratio of residual deviance to degrees of freedom was 3.67 indicating overdispersion and a violation of the model's assumption of having the mean and variance be equal. Because of this, a negative binomial model was fitted which yielded an AIC of 101,244 and a residual deviance of 13,764 on 13,351 degrees of freedom. The ratio of residual deviance produced by this model was much closer to one. The negative binomial model had a mean mortality rate of 0.00148 with standard deviation 6.014e-19.

Table 1: Diagnostics, Overall Mean, and Standard Deviation of Mortality Rate by Model

Model	AIC	Residual Deviance	Degrees of Freedom	Mean (Mortality Rate)	SD (Mortality Rate)
Linear	-	-	-	0.00149	0.0000491
Log-Linear	125630	49066	13354	-	-
Negative Binomial	101244	13764	13351	0.00148	6.014e-19

When looking at the estimated expected mortality, both the linear and negative model overpredicted mortality rates especially for older age groups. Figure 2 also shows that considerable variation in the data was left unexplained. However, from Figure 2, it is evident that the negative binomial model had an improvement in estimated expected mortality compared to the linear model, particularly seen in younger age groups. .

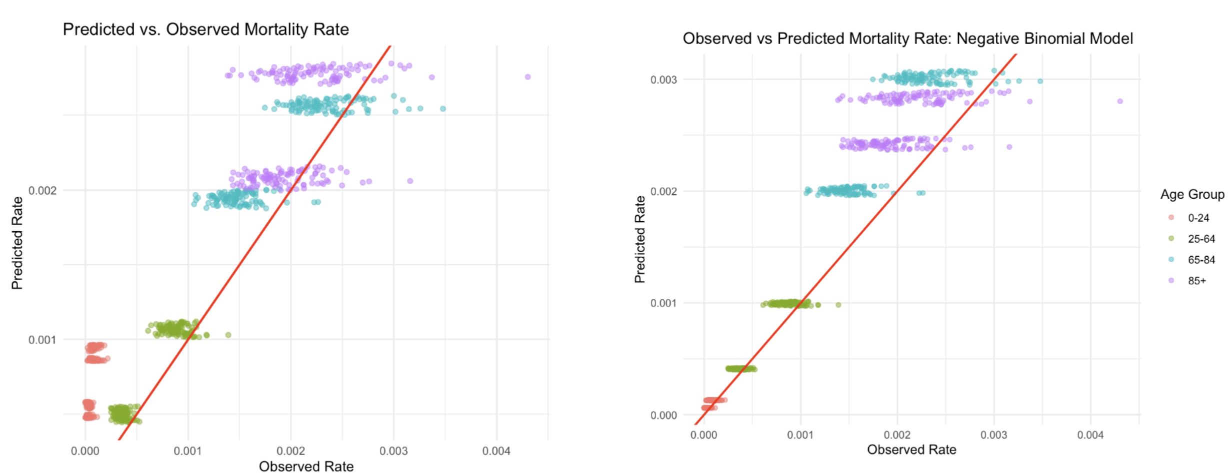

Figure 2: Estimated Expected Mortality by Age For Linear and Negative Binomial Model

To check for periods that might have had an unusual amount of excess mortality, a plot of excess mortality across weeks from 1985 to 2018 stratified by age and sex was created (See supplemental figure 4). From Supplemental Figure 4, there was an unusually high number of excess mortality rates around 1985 potentially due to the Puerto Rico Floods of 1985 (Jibson, 1985). However, further analysis showed that taking out this time period with higher levels of excess mortality did not affect the fitted models much and thus, models were kept as is fitted on all data from 1985-2016.

Using the fitted models to predict mortality rates and compare excess mortality from 2017-2018, Figure 3 shows that both the linear and negative binomial consistently overestimated the mortality rates for both sexs and age groups. For the linear model, both sexes and older age groups, besides the 0-24 age group, had a spike in change in excess mortality around the time that Hurricane Maria hit, showing how the linear model was not able to account for unforeseen disasters. Similarly, for the negative binomial model, although the predictions were still overestimated, the differences between excess mortalities for males and females were smaller (Figure 3). The model improved predictions for younger age groups as well and the model was able to consistently predict even when Hurricane Maria hit.

Figure 3: Linear vs. Negative Binomial - Excess Mortality 2017-2018

To compare data from the excessmort package and New York Times Data, a plot of the difference of 2016-2017 daily differences was plotted. Figure 4 showed that from the beginning of 2016 to roughly August of September the data between the two sources was very similar. However, data discrepancies arose around August 2017 and spiked in December 2017. This was primarily due to the New York Times did not have any data reported in December 2017.

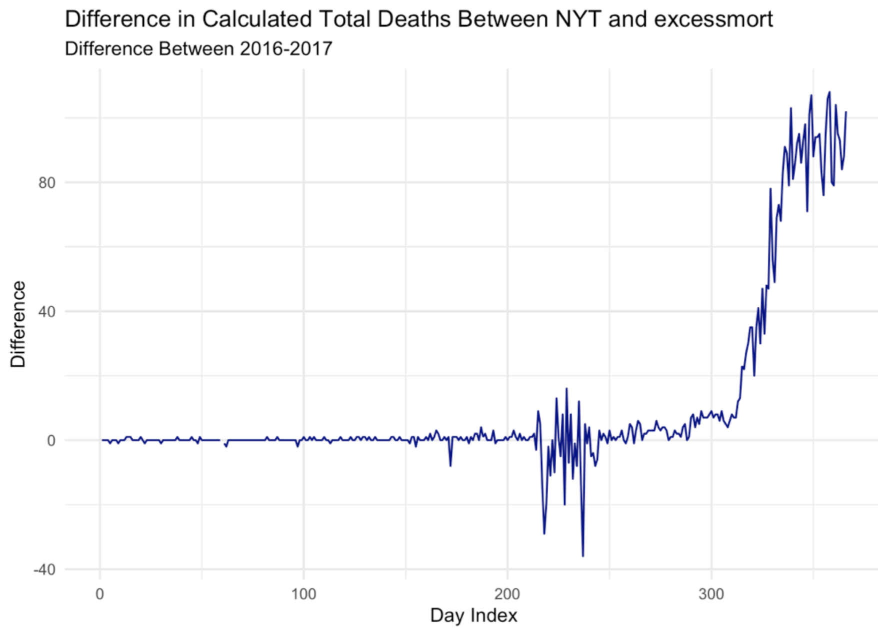

Figure 4: Comparison of excessmort Data and New York Times Data

Discussion

From our analysis, it was evident that the negative binomial model predicted mortality rates more accurately than the linear model. As shown in Plot X, the linear model consistently overestimated excess mortality across both sex and age groups. In particular, the magnitude of excess mortality was notably greater for males and the 65–84 and 85+ age groups, suggesting strong sex and age effects that the linear model may have exaggerated. A pronounced spike in excess mortality around weeks 35–40 of 2017, coinciding with Hurricane María, was observed across both sexes and for age groups 25 and older. This spike highlights the linear model's lack of robustness in handling periods of sharp mortality increases, likely due to its assumption of constant variance. Interestingly, although the linear model overestimated mortality for the 0–24 age group, this group did not show the same sharp increase in excess mortality during the hurricane period. This may reflect greater resilience among younger populations, while older individuals—particularly those with underlying health conditions—were more vulnerable due to infrastructure disruptions, such as loss of electricity.

In contrast, the negative binomial model produced more stable and realistic estimates, as seen in Figure 3. While overestimation was still present, the model clearly reduced variance in predictions and better aligned with expected mortality patterns. For example, the excess mortality gap between males and females was narrower, indicating that the model more appropriately captured the sex effect. Similarly, the flattened and smoother trends for the 0–24 and 25–64 age groups suggest that the negative binomial model provided more reliable predictions for these younger populations.

When cross checking the data used in 2016 and 2017 from the “excessmort” package in R with the New York Times reported data, it was found that both sources reported very similar data all throughout 2016 and for the most part of 2017. Figure 4 shows that while the two data sources reported similar data for the most part, there were some discrepancies that arose around August and September of 2017 with a spike at the very end of 2017. This was mainly due to the New York Times not having any data reported at all in December of 2017 and further highlights the importance of understanding where data used for analyses is taken from. Had the analyses been conducted using the New York Times data, the results would have been vastly different.

By implementing and comparing two statistical models to analyze excess mortality in Puerto Rico, this study provides a more reliable method to accurately assess the impacts of natural disasters such as Hurricane Maria. However, although the analyses showed that the negative binomial model predicted better than the linear model, there are still additional improvements that are worth mentioning to be implemented in the future. Given the strong age effect discussed earlier, it could be beneficial to implement a hierarchical model that would better account for the age variation left unexplained by the negative binomial model. Additionally, because the data spans time and exhibits seasonal and abrupt shifts, it is of possible interest to try and implement a time series model with a spline. By adding in a spline to the model, this would allow death counts to be a smooth function of time, effectively helping with predicting mortality rates more accurately. Lastly, although our model accounts for some of the main variables related to mortality, there are still many confounding variables that are likely to be strongly associated with mortality rates and corresponding predictions. Examples of these are a person’s socioeconomic status, geographic location, and access to healthcare as these are major variables tied to mortality rates especially in times of natural disasters.

In conclusion, this report aimed to compare multiple statistical methods and cross check alternative data sources in order to most accurately predict periods of excess mortality for natural disasters like Hurricane Maria. Developing robust predictive models that accurately predicts excess mortality is crucial for public health preparedness and for understanding the broader impacts of natural disasters especially on vulnerable communities.

References

Centers for Disease Control and Prevention. (n.d.). Morbidity and Mortality Weekly Report (MMWR) series. https://www.cdc.gov/mmwr/about.html

Irizarry, R. A., Acosta, R., Robles-Fontán, M., & Tamás, F. (2025). excessmort: Excess mortality (Version 0.8.1) [R package]. CRAN. https://CRAN.R-project.org/package=excessmort

Jibson, R. W. (1985). *Catastrophic landslide: The Mameyes, Puerto Rico, landslide* [Photograph]. U.S. Geological Survey. https://www.usgs.gov/media/images/catastrophic-landslide-0

Kishore, N., Marqués, D., Mahmud, A., Kiang, M. V., Rodriguez, I., Fuller, A., ... & Balsari, S. (2018). Mortality in Puerto Rico after Hurricane Maria. The New England Journal of Medicine, 379(2), 162–170. https://doi.org/10.1056/NEJMsa1803972

Mordkoff, J. T. (n.d.). The normal distribution. https://www2.psychology.uiowa.edu/faculty/mordkoff/GradStats/part%201/I.07%20normal.pdf

National Center for Health Statistics. (n.d.). Excess deaths associated with COVID-19. CDC Web Archive. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm

Ooms, J. (2025). pdftools: Text extraction, rendering and converting of PDF documents (Version 3.5.0) [R package]. CRAN. https://cran.r-project.org/package=pdftools

Penn State Eberly College of Science. (n.d.). Lesson 7: Multiple Linear Regression – 7.3 Extra Sums of Squares. STAT 504: Analysis of Discrete Data. https://online.stat.psu.edu/stat504/lesson/7/7.3

R Core Team. (2024). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/

Robles, F., Davis, K., & Almukhtar, S. (2018, August 9). Puerto Rico’s hurricane death toll was higher than reported, study says. The New York Times. https://www.nytimes.com/2018/08/09/us/puerto-rico-death-toll-maria.html

Sandberg, J. K., Acosta, R. J., Irizarry, R. A., & Zeger, S. L. (2023). A retrospective observational study of excess all-cause mortality in Puerto Rico after Hurricane Maria.

Frontiers in Public Health, 11, 10200579.
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10200579/

Santos-Burgoa, C., Sandberg, J., Suárez, E., Goldman, A., Pérez, C. M., Rivera, A., ... & Zeger, S. L. (2018). Differential and persistent risk of excess mortality from Hurricane Maria in Puerto Rico: A time-series analysis. The Lancet Planetary Health, 2(11), e478–e488. https://www.thelancet.com/journals/lanplh/article/PIIS2542-5196(18)30209-2/fulltext

Santos-Lozada, A. R., & Howard, J. T. (2018). Use of Death Counts From Vital Statistics to Calculate Excess Deaths in Puerto Rico Following Hurricane Maria. JAMA, 320(14), 1491–1493. https://doi.org/10.1001/jama.2018.10929

Venables, W. N., & Ripley, B. D. (2002). Modern applied statistics with S (4th ed.). Springer. https://CRAN.R-project.org/package=MASS