# Excess Mortality Estimation Considering Demographic Structure in Puerto Rico

Jintong Hong

## Abstract

This study aimed to construct a baseline model of mortality to estimate excess mortality in Puerto Rico for the years 2017 and 2018 following Hurricane Maria, using data from 1997 to 2016. The baseline mortality models were constructed using Linear Regression and Generalized Additive Model (GAM) methods and incorporated demographic and temporal covariates. Both models successfully identified an increase in mortality following Hurricane Maria. The GAM, with its ability to model non-linear relationships, demonstrated better predictive performance compared to the linear regression model, as it estimated a larger magnitude and longer duration of excess deaths after the hurricane. External validation using New York Times mortality data also supported the results observed in our model predictions. These findings help reveal the impact of Hurricane Maria on mortality in Puerto Rico and highlight the utility of GAM, while providing knowledge for public health measures and future policymaking.

## Introduction

The consequences of natural disasters usually extend far beyond acute, one-time damage. They often result in significant fluctuations in mortality within the affected regions. The quantification of excess mortality, which is the deviation from the normal observed deaths at the baseline level, is crucial for assessing the public health cost of such events, providing evidence-based public health advice, and helping construct strategies for similar crises. This study aims to build and assess models that aid in estimating excess mortality in Puerto Rico during the years 2017 and 2018, a post-hurricane period following Hurricane Maria, which hit the island in September 2017. This Category 5 hurricane inflicted widespread devastation across the island, disrupting essential services and damaging basic infrastructure. Understanding the specific magnitude and temporal dynamics of excess mortality during this post-hurricane period may be critical for future healthcare resource allocation, evaluating the current effectiveness of disaster response efforts, and identifying sub-populations that require targeted support in future crises.

Establishing a reliable baseline of expected mortality is a fundamental step for accurate excess mortality estimation. This project utilizes a comprehensive historical mortality dataset for Puerto Rico, covering 20 years from 1997 to 2016, which is provided in the `excessmort` R package. It is expected that this extensive temporal coverage can provide sufficient information for constructing statistical models that capture underlying mortality trends by age and gender, regular seasonal variations due to flu season, and the impact of public healthcare improvements. By analyzing this data, we aim to develop models capable of predicting expected mortality for the years 2017 and 2018 and demonstrating the impact of Hurricane Maria.

Recognizing that previous events in historic data may also cause abnormal mortality, this analysis excludes periods with elevated mortality associated with Hurricane Georges from the training set. Furthermore, to reduce model complexity and strengthen generalizability, age groups were reconstructed based on pre-2017 mortality patterns.

This study explores two distinct statistical modeling approaches for constructing baseline mortality: Linear Regression and Generalized Additive Models (GAMs). Linear regression models offer a relatively simple framework that assumes a linear relationship between predictors and mortality rates. In contrast, GAMs provide a more flexible, non-parametric approach capable of capturing non-linear associations between the predictors and directly estimating baseline mortality adjusted for population. By comparing the two methods, this research seeks to provide a comprehensive assessment of excess mortality in Puerto Rico following Hurricane Maria. Specifically, this study aims to achieve the following goals: 1) Develop and evaluate baseline mortality models for Puerto Rico using 20 years of historical data; 2) Estimate excess mortality for the years 2017 and 2018 using both linear regression and GAM frameworks; 3) Compare model prediction performances using metrics and visualizations; 4) Provide insights into the post-hurricane response to guide future policy adjustments and support studies estimating excess mortality from similar natural crisis.

## Methodology

This study aims to estimate the excess mortality in Puerto Rico for the years 2017 and 2018 using the dataset `puerto rico mortality` from the `excessmort` package, which dates back to 1997. Two statistical modeling approaches are employed: Linear Regression and Generalized Additive Model (GAM).

### 1. Dataset and Predictors

Mortality data spanning 20 years (1997-2016) were used to build the baseline mortality models for Puerto Rico. The dataset comprises 289,260 valid observations, with predictors including age group, date of the event, gender, and corresponding population size. To improve the baseline mortality estimation, a period of unusually high mortality (September 21st to November 30th, 1998) associated with Hurricane Georges was excluded from the training data. Age

groups were reconstructed based on the similarity of their weekly average mortality before 2017 to mitigate potential overfitting.

Three additional covariates were introduced to capture temporal trends: 1) week of the year (to account for seasonality); 2) number of years since 1997 (to model gradual long-term changes); and 3) number of days since January 1st, 1997 (as an alternative continuous time variable). These predictors aim to account for seasonal fluctuations (e.g., winter mortality spikes) and potential long-term shifts in mortality rates, possibly due to changes in the age structure of the Puerto Rican population (Figure 2.1).

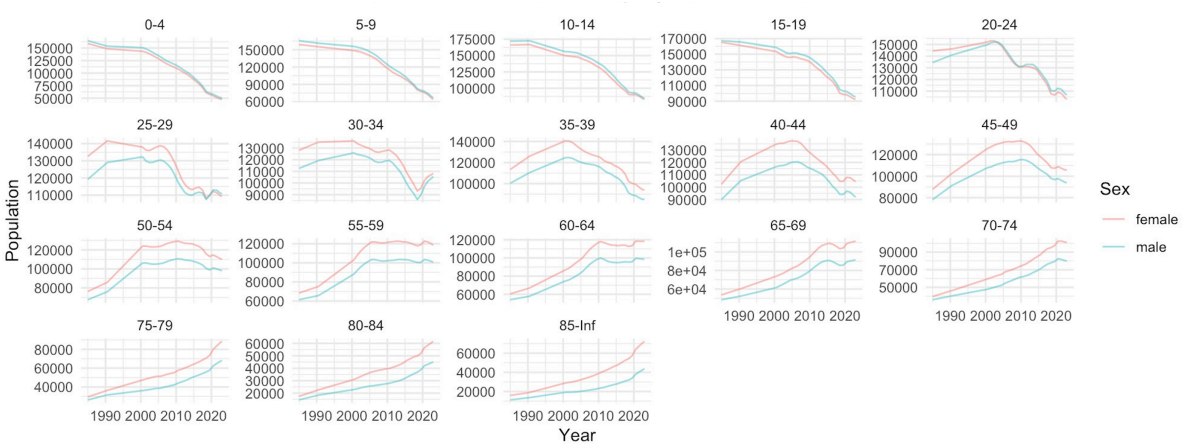


Figure 1: Figure 2.1: Population Trends Over Time by Age Group and Sex

## 2.1 Linear Regression Model

A simple linear regression framework was first used to predict the expected baseline mortality rate (deaths per population), considering sex, age group, week of the year, and either year or day since the start of the observation period. The number of death events was modeled indirectly, with population size used as an offset to account for varying population sizes across different age and gender groups.

Three specific linear regression models were fitted:

**Model 1:**

$$\mu_i = \beta_0 + \beta_{sex} \cdot x_{sex,i} + \beta_{agegroup} \cdot x_{agegroup,i} + \beta_{week} \cdot x_{week,i} + \beta_{year} \cdot x_{year,i}$$

(assumes a yearly constant change in mortality rate).

An interaction effect between sex and age group was visually identified (Figure 2.3), suggesting a disproportionate decrease in the male population compared to females after age 25. To account for this, the following model was fitted:

**Model 2:**

$$\mu_i = \beta_{sex} \cdot x_{sex,i} + \beta_{agegroup} \cdot x_{agegroup,i} + \beta_{week} \cdot x_{week,i} + \beta_{year} \cdot x_{year,i} + \beta_{sex \cdot agegroup} \cdot x_{sex \cdot agegroup,i}$$

(includes a yearly constant change and a sex-age group interaction).

Finally, a model assuming a daily constant change in mortality rate, also incorporating the sex-age group interaction, was fitted:

**Model 3:**

$$\mu_i = \beta_0 + \beta_{sex} \cdot x_{sex,i} + \beta_{agegroup} \cdot x_{agegroup,i} + \beta_{week} \cdot x_{week,i} + \beta_{day} \cdot x_{day,i} + \beta_{sex \cdot agegroup} \cdot x_{sex \cdot agegroup,i}$$

Model selection was based on the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) (Table 3.1, presented in the Results section). The expected number of deaths for week $i$ was then calculated using the chosen model (Model 2):

$$\text{Expected Deaths}_i = \hat{\mu}_i \times \text{population}_i$$

And the excess mortality was estimated as:

$$\text{Excess mortality}_i = \text{Observed Deaths}_i - \text{Expected Deaths}_i$$

### 2.2 Time Series Model (Generalized Additive Model - GAM)

A Generalized Additive Model (GAM) with a Poisson distribution and a log link function was employed to model the weekly number of mortality. This approach allows for non-linear relationships between the predictors and the expected number of deaths. The model incorporates a smooth function of time (varying by age group), a cyclic smooth function for seasonality, and the main effects of age group, sex, their interaction, and a linear effect of the year not fully captured by the smooth function.

The GAM was formulated as:

$$E(\text{Deaths}_i) = \text{Population}_i \times \exp\{\alpha(\text{day}_i, \text{age group}_i) + s(\text{week}_i) + w(\text{sex}_i, \text{age group}_i) + \gamma(\text{year}_i)\}$$

where $\text{Population}_i$ is the population at the time of observation $i$ (used as an offset). $\alpha(\text{day}_i, \text{age group}_i)$ represents a smooth, potentially non-linear trend over time (days since 1997.1.1) that varies by age group. $s(\text{week}_i)$ is a cyclic smooth function capturing the seasonal pattern across the weeks of the year. $w(\text{sex}_i, \text{age group}_i)$ models the interaction effect of sex

and age group on the baseline mortality rate. $\gamma(\text{year}_i)$ represents the linear effect of the numerical year on the mortality rate, accounting for longer-term trends not fully captured by the smooth term.

After fitting the GAM, the expected number of deaths for 2017 and 2018 was predicted by:

$$\widehat{\text{Mortality}}_i = \exp(\widehat{\log}(E(\text{Mortality}_i)))$$

Excess deaths were calculated similarly to the linear regression approach. The RMSE and MAE of both models (with the linear regression mortality rate recalculated using observed population) were computed for comparison. Additionally, publicly available mortality data from the New York Times for Puerto Rico (2015-2018) were used to further validate the baseline mortality models and assess for potential overfitting.

## Results

### 3.1 Linear Regression Model

The performance metrics (RMSE and MAE) for the three linear regression models are presented in Table 3.1. Model 2, which incorporates a constant yearly temporal trend and an interaction term between sex and age group for the mortality rate, exhibited the lowest RMSE (0.00022) and MAE (0.00015), indicating a better fit to the training data compared to Models 1 and 3.

Table 1: Table 3.1: Performance metrics for the linear regression models

| | Model 1 | Model 2 | Model 3 |
|---|---|---|---|
| RMSE | 0.00029 | 0.00023 | 0.00022 |
| MAE | 0.00021 | 0.00016 | 0.00015 |
| Adjusted R | 0.959 | 0.946 | 0.959 |

The coefficients for the age group covariates were all statistically significant ($p < 0.05$), demonstrating a strong association with mortality rate. The interaction terms between age groups and sex were generally significant, particularly for age groups above 50. Weeks in the latter half of the year also showed significant effects, along with a significant main effect of the year ($p < 0.05$). Overall, Model 2 explained a substantial proportion of the variance in the mortality rate, with an adjusted R-squared of 0.952, suggesting a good model fit.

Applying the mortality rate predictions from Model 2, the excess mortality for 2017 and 2018 was estimated and compared to the observed mortality (Figure 3.1). In the three weeks immediately following Hurricane Maria (September 20th to October 4th, 2017), the linear regression model estimated a total of 398.3 excess deaths.

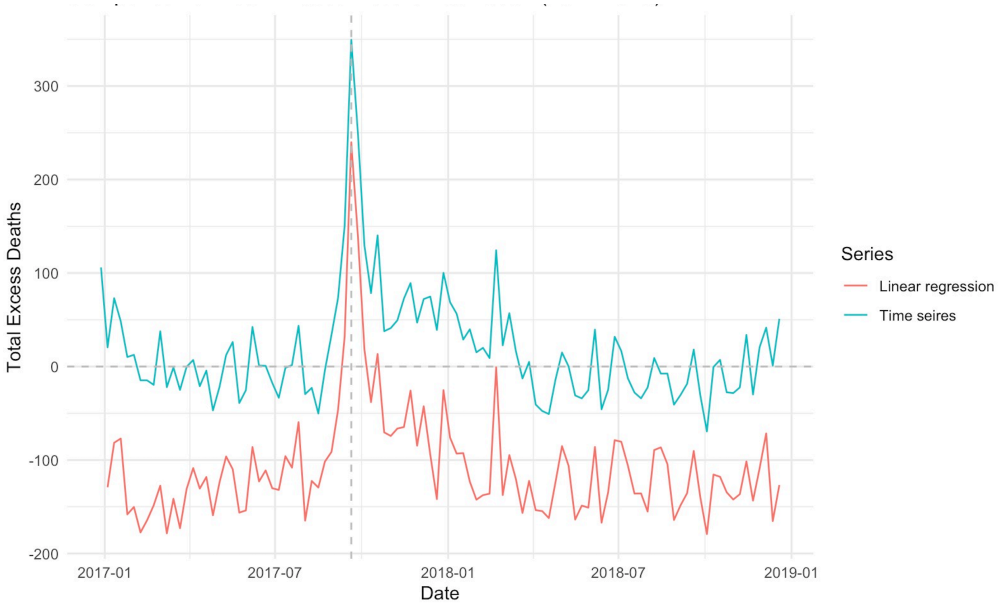


Figure 2: Figure 3.1: Comparison of Total Excess Mortality Estimation (2017-2018)

### 3.2 Time Series Model

The Generalized Additive Model (GAM) revealed that most age groups and their interactions with sex were statistically significant predictors of mortality, with higher estimated mortality rates observed for older age groups. The smooth terms for time (varying by age group) and the cyclic smooth term for the week of the year were also highly significant, indicating the presence of non-linear temporal trends and seasonal patterns in mortality. The GAM demonstrated a high proportion of explained variation with an adjusted R-squared of 0.952, suggesting a strong fit to the data.

The performance metrics for the GAM were an RMSE of 7.876 and an MAE of 4.4899. Estimating excess mortality for 2017 and 2018 using the GAM predictions and comparing it with observed mortality (Figure 3.1) showed a total of 598.7 excess deaths in the four weeks following Hurricane Maria (September 20th to October 25th, 2017).

### 3.3 Modeling Method Comparison

To directly compare the performance of the linear regression model with the GAM in predicting the number of deaths, the RMSE and MAE of the linear regression model were recalculated using the observed population to predict the number of deaths. The results, presented in Table 3.2, show that the GAM exhibited lower RMSE and MAE compared to the linear regression model, indicating a superior predictive performance.

Table 2: Table 3.2: Comparison of RMSE and MAE for the two modeling methods (predicting number of deaths)

| | LR | GAM |
|---|---|---|
| RMSE | 11.4469 | 7.876 |
| MAE | 7.2997 | 4.4899 |

The visualization of the excess mortality estimated by both models is shown in Figure 3.1. Furthermore, publicly available mortality data from the New York Times for Puerto Rico from 2015 to 2018 (Figure 3.2) were used to assess the validity of our baseline mortality models. The estimated excess mortality based on our models showed a similar pattern and magnitude to that derived from the NY Times data (Figure 3.3), providing evidence against overfitting or underfitting in our model construction.

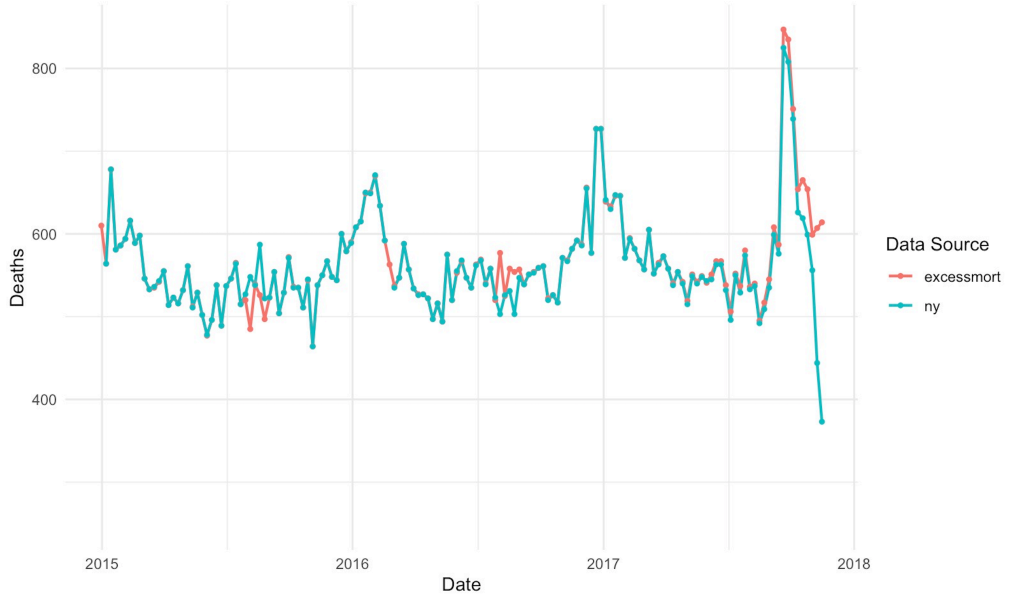


Figure 3: Figure 3.2: Comparison of Weekly Mortality: Excessmort vs NY Times

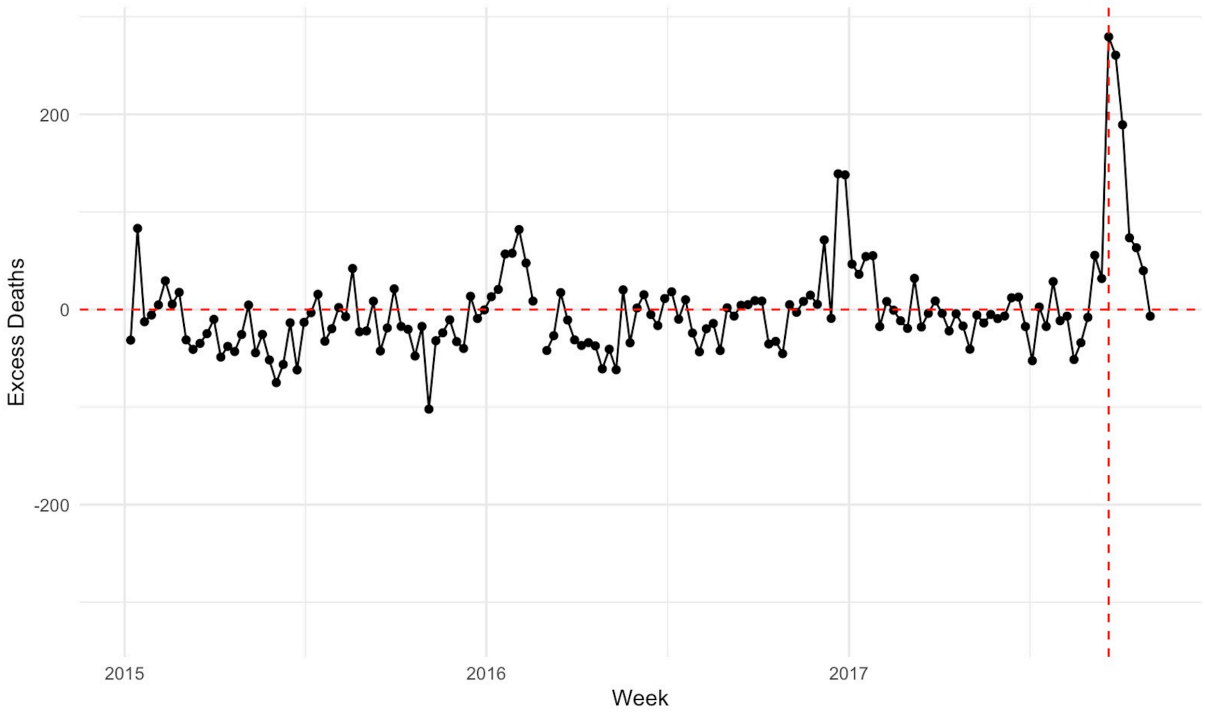


Figure 4: Figure 3.2: Weekly Excess Mortality in Puerto Rico by NY Times (2015-2017)

## Discussion

The results of this study reveals insights into the excess mortality in Puerto Rico during 2017 and 2018, particularly in the aftermath of Hurricane Maria, as well as a understanding of the baseline mortality of Puerto Rico.

Both the linear regression and the GAM identified a significant increase in mortality following the hurricane. However, the magnitude and duration of this increase differed between the two models. GAM outperformed the linear regression model, with lower RMSE and MAE values. The GAM, with its ability to capture non-linear relationships through smooth functions of consecutive time—varying across age groups and seasonal patterns such as winter flu—enhanced prediction accuracy. In addition, the GAM's adjusted R-sqaured was slightly higher than the linear regression model, indicating that it explains more variation of the dataset.

There was a substantial difference between the two models' estimates of excess deaths (398.3 vs. 598.7) and the duration of the immediate post-hurricane period (3 vs. 4 weeks). This emphasized the GAM's higher predictive power, which is crucial for accurate post-hurricane casualty assessment and for revising public health measures. The discrepancy suggests that the GAM accounted for non-linear temporal trends in baseline mortality and more complex

interactions. A linear model with simple interaction terms may not adequately capture these patterns, resulting in poorer estimates of excess mortality.

On the other hand, although the linear regression model may be limited in predictive performance due to its strict linearity assumption between predictors and mortality rates, it still provides valuable insights into Puerto Rico's mortality with a more interpretable framework.

The improvement observed when incorporating an interaction term between sex and age group (Model 2) highlights the importance of accounting for demographics in mortality patterns. This interaction effect was particularly significant for age groups above 50, suggesting that sex-based mortality differences become more pronounced with advancing age in Puerto Rico. The consistency between our model-based estimates and the publicly available New York Times mortality data provides essential external validation. This alignment suggests that our methodological approach successfully captured the underlying mortality patterns while avoiding both overfitting and underfitting. Such validation strengthens confidence in our estimates of excess mortality attributable to Hurricane Maria and its consequnces.

The statistically significant effects of age group, and the interaction between age group and sex in both models, highlight the difference of mortality risk across all demographics of population. The increased mortality rates observed in older age groups align with common epidemiological patterns. The significant week effects likely reflect the seasonality effects, such as flu outbreaks in winter. The significant yearly trend captured by the linear regression and the smooth trend in the GAM suggest a gradual increase of baseline mortality rates in Puerto Rico over last two decades, potentially influenced by factors such as dramatic changes in the population's age structure or improvements in healthcare.

Additionally, the consistency between the excess mortality estimates from our models and those reported by the independent New York Times mortality data from 2015 to 2018 provides strong validation for our GAM approach. This indicates that our baseline mortality model, constructed using 20 years of historical data up to 2016, did not overfit the training data and can reasonably predict expected mortality.

Our findings also have significant implications for public health planning and disaster response. The esimated excess mortality following Hurricane Maria in our model indicates the need for more robust emergency preparedness and response systems, particularly for vulnerable subjects in the population. The high mortality rates beyond the immediate hurricane hit suggests that secondary effects—such as disrupted healthcare services, compromised sanitation, damaged infrastructure— have contributed to disaster-related mortality that was not properly coped by the local authority. The identified interaction between sex and age group could help guide targeted interventions for specific demographic segments for future crisis. For example, elderly appeared to be vulnerable during the crisis with higher mortality rate, indicating a need for specialized support systems for this group during emergencies.

While our models accounted for seasonality, long-term trends, and demographic factors, they may not fully capture other potentially relevant covariates such as socioeconomic status, healthcare access, or geographic population distribution. Future studies could enhance the

baseline model's predictive performance by incorporating more detailed information and additional predictors, or use advanced machine learning approaches.

In conclusion, this study provides robust evidence of significant excess mortality in Puerto Rico following Hurricane Maria, along with methodological insights into the comparative strengths of different statistical approaches for estimating disaster-related mortality. These findings emphasize the severe impact of Hurricane Maria and the importance of comprehensive disaster response measures that can help mitigate loss in future crisis.