# Hammond's Theory Revisited: A Broader Perspective on Mortality Prediction

*Damini Sri Vadrevu | vadrevuds@gmail.com | Hartford, CT*

## Abstract

The relationship between smoking and mortality has long been debated. E. Cuyler Hammond argues that smoking has a causal effect on mortality, asserting that cigarette use exacerbates a variety of health issues and raises the risk of death across multiple diseases, in a manner akin to drugs that treat multiple conditions. [HAMMOND EC, HORN D. *The relationship between human smoking habits and death rates: a follow-up study of 187,766 men.* J Am Med Assoc. 1954 Aug 7;155(15):1316-28. doi: 10.1001/jama.1954.03690330020006. PMID: 13174399]. In contrast, Berkson posits that smoking alone cannot be the primary cause of such a diverse range of diseases, suggesting instead that smoking is merely associated with mortality, not necessarily causative—a viewpoint often referred to as Berkson's Paradox. This project aims to explore the validity of these claims by analyzing two distinct datasets. The first dataset, from Hammond, is used to assess the association between smoking and mortality using logistic regression. The second dataset, examining life expectancy through various health and lifestyle factors, allows for the analysis of non-smoking influences on mortality, such as BMI and income, through logistic regression applied to a binary threshold on mortality risk. This comparison will provide insight into whether factors beyond smoking significantly predict mortality, addressing the limitations in Hammond's smoking-centric data.

# I. INTRODUCTION

## 1.1 Literature Review

Hammond's research presents a compelling argument for smoking as a significant factor in mortality, associating it with an increased risk across various diseases, including lung cancer, coronary artery disease, and peptic ulcers. His findings suggest that cigarette smoking exacerbates health risks, affecting disease severity and contributing to higher mortality rates. Hammond likens the effects of smoking to drugs with broad side effects, arguing that just as drugs can impact multiple conditions, so too can smoking increase mortality from multiple diseases, even if not directly causing each one. Berkson, however, criticizes this broad association. He argues that it is implausible for smoking alone to cause such a wide array of diseases, suggesting instead that smoking might correlate with, but not be the sole driver of, mortality across these illnesses. He employs what is now known as Berkson's Paradox: the appearance of association across multiple diseases does not imply causation, particularly if other confounding factors (like lifestyle and underlying health conditions) are not accounted for. In this project, I'll build on these perspectives by examining whether obesity and other

lifestyle factors can similarly contribute to mortality risks. Comparing Hammond's smoking-focused dataset with a broader dataset of health factors and mortality will allow for a more nuanced understanding of whether multiple non-smoking factors might also serve as significant mortality predictors.

## 1.2 Problem Statement

The objective is to compare the logistic regression coefficients from each model to assess the relative strength of smoking versus non-smoking factors in predicting mortality. This comparison will help determine if, as Berkson suggested, non-smoking factors might have comparable or greater influence on mortality than smoking alone

## 1.3 Report Outline

# II. CASE STUDY

This study examines the relative predictive power of smoking compared to other non-smoking health and lifestyle factors on mortality. This analysis is rooted in the debate between E. Cuyler Hammond and Joseph Berkson. Hammond argued that smoking exacerbates health risks and is associated with mortality from multiple diseases, implying a causal or at least aggravating effect. In contrast, Berkson contended that smoking alone is unlikely to explain a wide range of health outcomes, suggesting that the observed association may not indicate causation—a concept known as Berkson's Paradox. He argued that this association might reflect underlying biases or other unmeasured factors influencing mortality, rather than smoking being a sole, direct cause.

## 2.1 Datasets Description

To test these perspectives, datasets will be compared.

- **Hammond's Smoking Dataset**: This dataset provides counts of mortality among smokers and non-smokers across age groups, allowing for logistic regression analysis on mortality as a binary outcome (alive vs. deceased) with smoking and age as predictors.

    - Smoking Status (Column A): A binary variable indicating smoking behavior (0 = Non-Smoker, 1 = Smoker).
    - Age (Column B): A categorical variable representing different age groups (1, 2, 3, 4).
    - Death Status (Column C): A binary outcome indicating survival or mortality (0 = Alive, 1 = Deceased).
    - Number of Cases (Column D): The number of individuals in each specific category (combination of Smoking Status, Age, and Death Status).

- **WHO Dataset:** This dataset includes various non-smoking health and socioeconomic factors, such as BMI, healthcare expenditure, and income. Here, we define a binary mortality outcome based on life expectancy (e.g., high vs. low life expectancy) and use logistic regression to evaluate the impact of non-smoking factors like obesity and socioeconomic status on mortality.The dataset related to life expectancy, health factors for 193 countries has been collected from the same WHO data repository website and its corresponding economic data was collected from United Nation website of the year 2001.

    - Status: The economic status of the country (e.g., "Developing").
    - Life Expectancy: Average life expectancy in years.
    - Adult Mortality: Number of deaths per 1,000 adults in the population.
    - Infant Deaths: Number of infant deaths per 1,000 live births.
    - Alcohol: Per capita alcohol consumption (in liters).
    - Percentage Expenditure: Government health expenditure as a percentage of GDP.
    - Hepatitis B: Hepatitis B immunization coverage among 1-year-olds (%).
    - BMI: Average Body Mass Index across the population.

- Under-Five Deaths: Number of deaths of children under five years old per 1,000 live births.
- Polio: Polio immunization coverage among 1-year-olds (%).
- Total Expenditure: Total health expenditure as a percentage of GDP.
- Diphtheria: Diphtheria immunization coverage among 1-year-olds (%).
- HIV/AIDS: HIV/AIDS mortality rate per 1,000 adults.
- GDP: Gross Domestic Product per capita (in USD).
- Population: Total population.
- Thinness 1-19 Years: Prevalence of thinness among children and adolescents aged 1-19 (%).
- Thinness 5-9 Years: Prevalence of thinness among children aged 5-9 (%).
- Income Composition of Resources: A measure of the human development index accounting for income.
- Schooling: Average number of years of schooling.

# III. STATISTICAL MODELS

The predictive power of smoking versus non-smoking factors on mortality was evaluated using logistic regression for the Hammond dataset, and both Poisson regression and negative binomial regression for the WHO dataset. Logistic regression was employed to analyze the binary mortality outcome, while Poisson and negative binomial regressions were used to model mortality counts in the WHO dataset, focusing on count-based predictions..

## 3.1 Assumptions:

- Independent Observations: Logistic and Poisson regression assumes each observation is independent, which might not be perfect in population-based data.
- Linearity of Logit: Logistic regression assumes a linear relationship between the logit of mortality and predictors, which may not hold exactly.
- No Multicollinearity: It's assumed that non-smoking predictors (e.g., BMI, GDP) in the life expectancy dataset do not exhibit high multicollinearity, which could impact results.
- Equidispersion: The mean and variance of the dependent variable are assumed to be equal.

## 3.2 Statistical Methodology

- **Logistic Regression**: Applied to the Hammond dataset to analyze the relationship between smoking and binary mortality outcomes. Coefficients were used to evaluate the strength and direction of predictors.
- **Poisson Regression**: Modeled mortality counts in the WHO dataset under the assumption of equidispersion. Coefficients were interpreted to assess the impact of non-smoking factors on mortality.
- **Negative Binomial Regression**: Used for the WHO dataset to address overdispersion. This model provided more robust estimates for data where the variance exceeded the mean.

- **Goodness-of-Fit Tests**: Deviance and Pearson's chi-square statistics were used to assess model fit, with scaled values near one indicating adequacy. AIC and BIC were also computed for Poisson and negative binomial models to compare performance.
- **Model Accuracy Metrics**: Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) were calculated for all models to evaluate predictive performance. Predicted probabilities or counts were compared to actual values, with lower values indicating higher accuracy.

# IV. DATA ANALYSIS

## 4. 1 Exploratory Data Analysis (EDA)

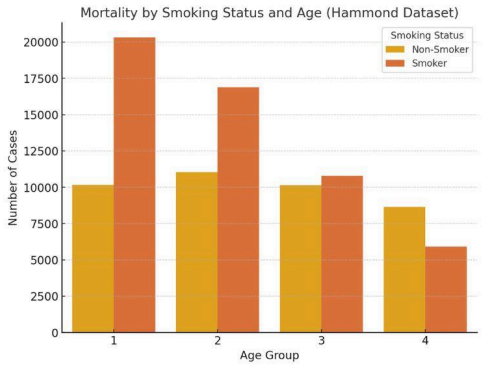
This section provides a detailed analysis of the data. Exploratory Data Analysis was performed to assess the distributions and relationships within the data. Q-Q plots were used to evaluate the normality of numeric features. The Q-Q plots confirmed that the data followed a normal distribution, ensuring the suitability of subsequent statistical analysis.

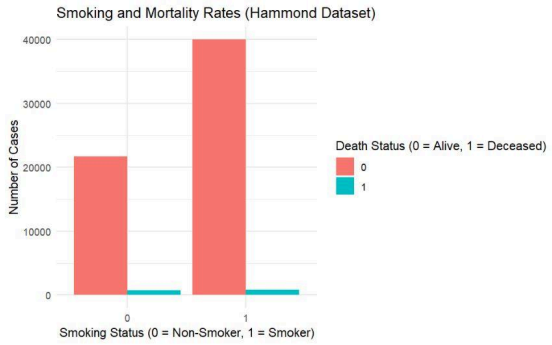
### 4.1.1. Understanding the Data and Feature Overview

The analysis was conducted on two distinct datasets: the Hammond dataset and the WHO Life Expectancy dataset. The Hammond dataset consists of 16 rows and 4 columns, capturing data on smoking status, age, death status, and the number of cases for each combination of these variables. Smoking status is a binary variable indicating whether an individual is a smoker (1) or a non-smoker (0). Age is a categorical variable divided into four groups, ranging from 1 to 4. Death status, another binary variable, indicates whether an individual is alive (0) or deceased (1). The number of cases is a numeric variable, representing the count of individuals in each category combination, with an average of 11,735 cases, a minimum of 204 cases, and a maximum of 39,990 cases. The WHO Life Expectancy dataset is larger, containing 178 rows and 15 columns. It includes features such as obesity percentage, life expectancy, body mass index (BMI), income composition of resources, alcohol consumption, mortality rates, and overweight prevalence. Obesity percentage is a continuous variable with a mean of 12.80%, ranging from 0.7% to 37.7%. Life expectancy is also continuous, with an average of 67.35 years, ranging from 41 to 82 years. BMI, another continuous variable, has a mean of 24.43, with values ranging from 19.8 to 30.9. Income composition of resources measures socioeconomic status, with values ranging from 0 to 0.917 and a mean of 0.566. Alcohol consumption, measured in liters per capita, averages 4.58 liters and ranges from 0.01 to 14.27 liters. Mortality is expressed as the number of deaths per 1,000 adults, with a mean of 214.67 and a range of 69 to 688. Overweight prevalence is recorded as a percentage, with an average of 37.91%, ranging from 9.6% to 68.8%. Each feature was reviewed to understand its nature, whether categorical, continuous, or binary. The target variables for analysis were identified as the binary mortality outcome (death status) in the Hammond dataset and the mortality count derived from life expectancy thresholds in the WHO dataset. These target variables provided the foundation for assessing the relationship between lifestyle factors and mortality. The features were reviewed to understand their nature (categorical, continuous, or binary), and target variables were identified:

- **Hammond's Dataset:** Binary mortality outcome (Death Status).
- **WHO Dataset:** Mortality count based on life expectancy thresholds.

The analysis investigated the distributions, relationships, and trends among key features in the Hammond and WHO Life Expectancy datasets to better understand their association with mortality. The relationship between smoking status, age, and mortality was examined. Mortality rates were observed to vary by age groups, with older age groups generally exhibiting higher death rates. The chart categorizes smoking status into non-smokers (0) and smokers (1), with the number of cases further divided into alive (death status = 0) and deceased (death status = 1). It is observed that the majority of individuals in the dataset fall under the "alive" category, irrespective of smoking status. Smokers represent a larger proportion of cases compared to non-smokers, reflecting a higher prevalence of smoking in the dataset. However, the count of deceased individuals is only marginally higher among smokers than non-smokers. This difference is relatively small in relation to the total number of cases, suggesting that smoking status alone may not be a definitive predictor of mortality.



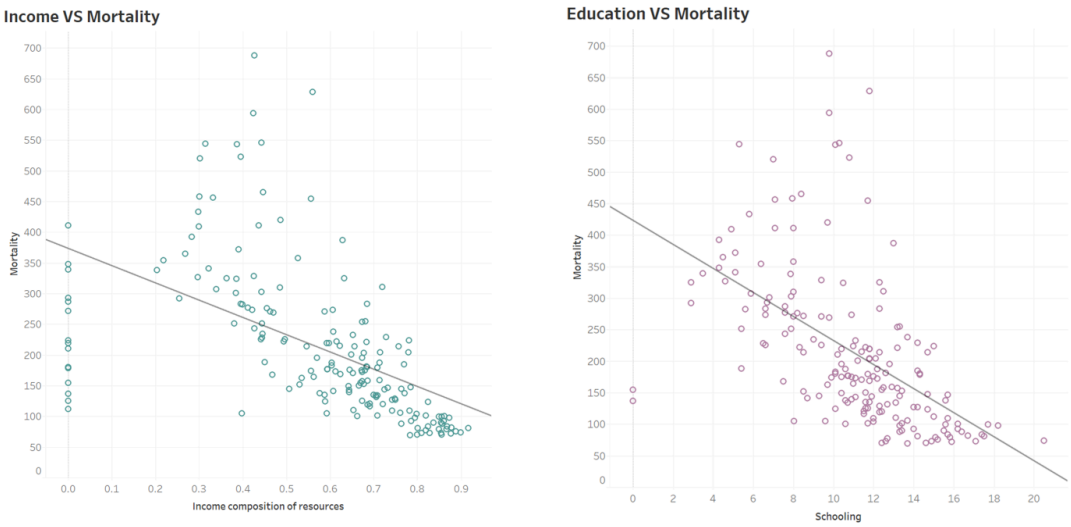




Smoking correlates with mortality because it is a well-documented risk factor for several life-threatening conditions, including lung cancer, heart disease, chronic obstructive pulmonary disease (COPD), and stroke. These conditions significantly contribute to premature mortality rates among smokers. The physiological impact of smoking, such as reduced lung capacity, impaired cardiovascular function, and systemic inflammation, directly increases the likelihood of severe health complications, thereby linking smoking with higher mortality rates.

However, smoking may not be the sole determinant of mortality because mortality is influenced by a multitude of factors beyond smoking alone. Socioeconomic conditions, access to healthcare, genetic predispositions, dietary habits, physical activity levels, and other lifestyle factors all play a significant role in shaping an individual's overall health and risk of mortality. For instance, a smoker with access to high-quality healthcare and a balanced diet might exhibit better health outcomes than a non-smoker exposed to poor living conditions, limited healthcare access, or a sedentary lifestyle.

The analysis of mortality rates reveals distinct relationships with income composition of resources, obesity percentage, and alcohol consumption. An inverse relationship is observed between income composition and mortality rates, with countries possessing higher levels of socioeconomic resources

experiencing lower mortality. This trend underscores the critical role of equitable income distribution in improving public health by facilitating better access to healthcare, nutrition, and living conditions. The scatter plot further illustrates this pattern, showing a clustering of lower mortality rates in nations with higher income composition values.




Correlation between Variables of WHO Dataset

Shifting from income to health-related factors, obesity percentage demonstrates a positive correlation with mortality rates. Higher levels of obesity are linked to increased mortality, a relationship that aligns with well-documented health risks such as cardiovascular diseases and diabetes. While this trend is clear, some variability in mortality at certain obesity levels suggests that other influencing factors, such as healthcare access or genetic predispositions, may modulate the impact of obesity on mortality outcomes. Alcohol consumption adds another layer of complexity to the analysis. Unlike the relatively straightforward relationships seen with income and obesity, alcohol consumption exhibits a non-linear impact on mortality. Moderate consumption levels correspond to stable mortality rates, whereas higher levels introduce significant variability. This variability reflects alcohol's multifaceted effects on health, where excessive consumption can exacerbate underlying health conditions or interact with other risk factors. The box plot underscores this variability, highlighting the need for a more nuanced understanding of alcohol's role in shaping public health outcomes.

Additionally, education, as measured by schooling years, plays a significant role in shaping mortality rates. A clear negative correlation is observed between schooling and mortality, with countries exhibiting higher levels of education generally experiencing lower mortality rates. This relationship suggests that education likely contributes to better public health by enhancing access to healthcare, fostering health awareness, and reducing risky behaviors. The impact of education on mortality underscores the importance of addressing educational disparities to improve population health and life expectancy.

### 4.1.2. Train-Test Split for Statistical Modeling

To prepare the data for analysis and modeling, both datasets were split into training and testing sets, adhering to an 80/20 split ratio. This approach ensured that a sufficient portion of the data was allocated for training, while reserving an adequate sample for testing and evaluating model performance.

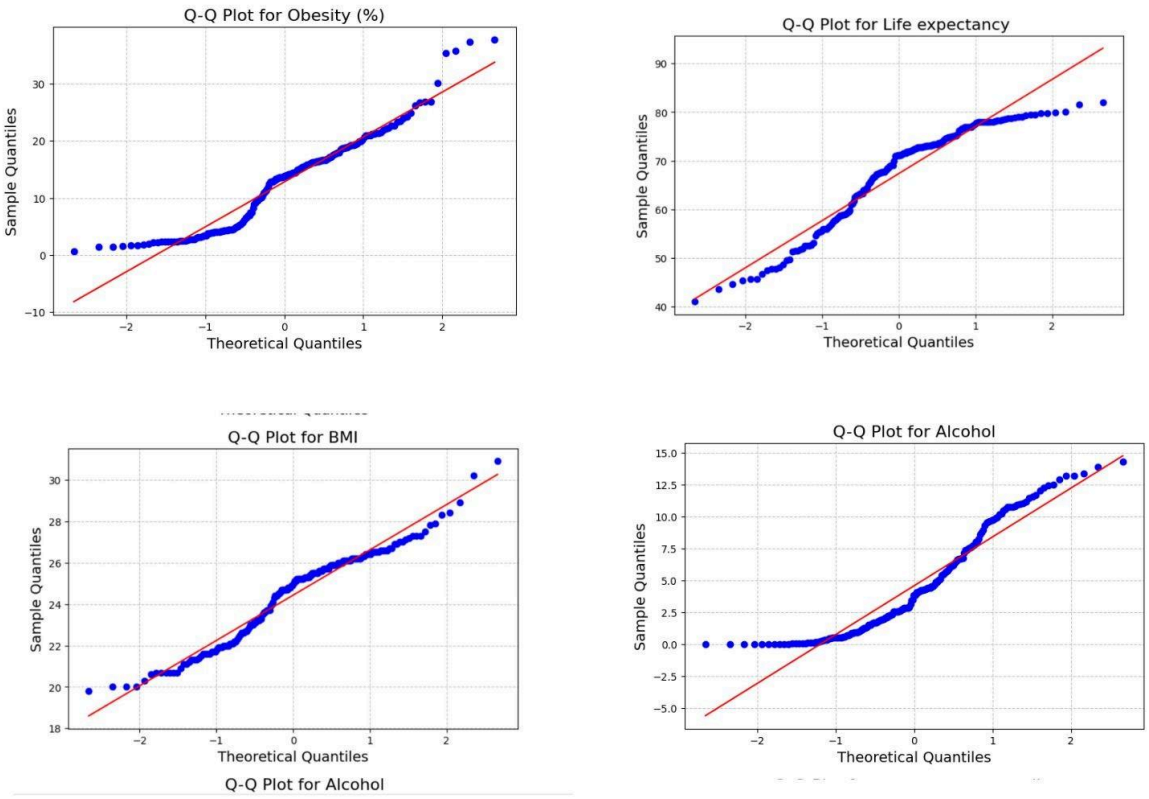
For the Hammond dataset, which has a small sample size of 16 rows, the 80/20 split was used to maximize the training data available for the model. This resulted in 12 rows for training and 4 rows for testing. Given the small size, careful handling of the split was necessary to maintain the representativeness of the dataset. The **WHO dataset,** which is relatively larger but still small with 178 rows, the same 80/20 split was applied. This resulted in 142 rows for training and 36 rows for testing. The larger training set facilitated robust model building while the test set provided sufficient data for assessing model generalizability.

### 4.1.3. Feature Selection

The WHO dataset was thoroughly reviewed to identify and exclude features that were unrelated to mortality prediction, such as metadata fields, Polio vaccinations, HIV and variables that were not of primary focus. Features retained for analysis included BMI, income, overweight prevalence, and alcohol consumption, as these variables are well-established predictors of health outcomes and mortality in the literature. BMI reflects overall health and obesity levels, income serves as a proxy for socioeconomic status and access to healthcare, overweight prevalence indicates population-level lifestyle trends, and alcohol consumption is a key factor linked to several health risks. In the Hammond dataset, the variable "Number of Cases" was excluded due to its high correlation with "Death Status." Retaining this variable would have introduced multicollinearity, potentially distorting the model's estimates and reducing the reliability of the results.

### 4.1.4. Checking for Normality and Scaling Techniques

The variables in the WHO dataset were assessed for normality using Q-Q plots, which confirmed that all variables followed a normal distribution. Based on these results, standardization was performed in R after splitting the dataset into training and testing subsets, as outlined previously. The Hammond dataset predominantly consists of binary and categorical variables, such as smoking status and age As these variables do not require normalization, no preprocessing was applied to these features for modeling. Additionally, age is treated as a categorical variable rather than continuous. Therefore, no transformations or scaling techniques were utilized in this stage of the analysis




QQ Plot Visualizations for WHO Dataset

#### 4.1.5. Assessment of Missing Data

To ensure data completeness, both the WHO and Hammond datasets were rigorously examined for missing values. In R, this was accomplished using functions such as `is.na()` and `summary()` to detect any `NA` values or anomalies within the datasets. Both the WHO and Hammond datasets were examined for missing values. The analysis confirmed that neither dataset contained any missing entries. As a result, no imputation or other preprocessing actions were required to address missingness in the data.

## 4.2 Limitations:

- Lack of Smoking Data in the Life Expectancy Dataset: The life expectancy dataset does not include smoking status, which limits direct comparisons of smoking's effect in both datasets.
- Confounding Factors: Both datasets may have unobserved confounding variables, potentially affecting the accuracy of the logistic regression models.
- Limited Scope of Non-Smoking Factors: The selected non-smoking factors (e.g., BMI, GDP) may not fully capture all relevant influences on mortality.
- Causal Interpretation: Logistic regression results are associative, not causal, so observed associations should not be interpreted as proof of causation.

# V. CONCLUSION

## 5.1 Results and Interpretation

### 5.1.1 Statistical Significance of Coefficients

Logistic regression analysis of the Hammond dataset revealed that smoking status ($p = 1.00$) and age ($p = 1.00$) were statistically insignificant predictors of mortality, indicating that smoking does not have a meaningful impact on mortality within this dataset. This challenges Hammond's argument that smoking is the primary driver of mortality and aligns with Berkson's perspective that the observed association between smoking and mortality may result from unmeasured confounding factors. The lack of significant results also reflects the limitations of the Hammond dataset, which focuses narrowly on smoking and does not account for broader determinants of mortality.

The Poisson regression model applied to the WHO dataset yielded numerous significant predictors, highlighting the importance of various socioeconomic and lifestyle factors in influencing mortality. For instance, life expectancy showed a highly significant negative coefficient of -0.04684 ($p < 0.0001$), suggesting that longer life expectancy is strongly associated with reduced mortality. Income composition of resources also demonstrated a significant negative relationship with mortality, with a coefficient of -0.03130 ($p < 0.001$), emphasizing its role in improving health outcomes through better access to healthcare and higher living standards. Conversely, alcohol consumption emerged as a significant positive predictor with a coefficient of 0.01538 ($p < 0.0001$), reinforcing its detrimental health effects. Obesity was another notable positive predictor, with a coefficient of 0.01404 ($p < 0.001$), indicating its contribution to increased mortality risk. These findings underscore the complexity of factors influencing mortality and highlight the dominant role of socioeconomic and lifestyle determinants over smoking.

In contrast, the negative binomial regression model, which accounts for overdispersion in the data, identified fewer significant predictors. While variables such as life expectancy retained their strong negative association with mortality (-0.04462, $p < 0.0001$), other predictors, including obesity and income composition of resources, lost statistical significance. However, alcohol consumption remained a significant positive predictor with a coefficient of 0.01624 ($p < 0.001$), reaffirming its adverse impact on health. These results suggest that while negative binomial regression provides a more robust fit for overdispersed data, it may also reduce the apparent significance of some predictors, further underscoring the importance of alcohol as a consistent determinant of mortality across models.

```
Call:
glm(formula = Death.Status ~ ., family = binomial(link = "logit"),
    data = dataa)

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.651e-16  1.323e+00       0        1
Smoking.Status  1.110e-16  1.000e+00       0        1
Age            -2.483e-16  4.472e-01       0        1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 22.181  on 15  degrees of freedom
Residual deviance: 22.181  on 13  degrees of freedom
AIC: 28.181

Number of Fisher Scoring iterations: 2
```

Summary of Logistic Regression Model

These findings collectively demonstrate that smoking is not a significant predictor of mortality, whereas socioeconomic and lifestyle factors, such as education, income, and alcohol consumption, play a much larger role. The Poisson model revealed the broad influence of multiple factors, while the negative binomial model reinforced the importance of key variables like alcohol consumption and life expectancy. These insights highlight the need for public health strategies to focus on addressing broader determinants of health, such as reducing alcohol use, combating obesity, and improving socioeconomic conditions, to effectively reduce mortality rates.

```
Call:
glm(formula = Mortality ~ ., family = poisson(link = "log"),
    data = data)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     8.579e+00  1.645e-01  52.160  < 2e-16 ***
Country                         2.207e-04  1.015e-04   2.174 0.029716 *
Obesity....                     1.404e-02  3.057e-03   4.594 4.35e-06 ***
Life.expectancy                -4.684e-02  9.400e-04 -49.825  < 2e-16 ***
BMI                            -1.152e-02  8.266e-03  -1.393 0.163578
Alcohol                         1.538e-02  1.917e-03   8.021 1.05e-15 ***
percentage.expenditure         -1.161e-05  2.320e-05  -0.501 0.616653
Total.expenditure              -1.143e-02  2.963e-03  -3.858 0.000114 ***
GDP                            -1.153e-05  3.489e-06  -3.305 0.000948 ***
Population                     -9.575e-10  2.657e-10  -3.604 0.000314 ***
thinness..1.19.years            6.312e-03  3.189e-03   1.979 0.047802 *
thinness.5.9.years             -2.548e-03  3.169e-03  -0.804 0.421417
Income.composition.of.resources -3.130e-03  3.522e-02  -0.089 0.929187
Schooling                       1.103e-02  2.876e-03   3.834 0.000126 ***
Overweight.Prevalance          -4.863e-03  1.656e-03  -2.937 0.003311 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Summary of Poisson Model

```
Call:
glm.nb(formula = Mortality ~ ., data = data, init.theta = 37.94160811,
    link = log)

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                     8.351e+00  4.259e-01  19.606  < 2e-16 ***
Country                         1.741e-04  2.661e-04   0.654 0.512999
Obesity....                     1.097e-02  7.579e-03   1.448 0.147670
Life.expectancy                -4.644e-02  2.564e-03 -18.116  < 2e-16 ***
BMI                             1.590e-03  2.118e-02   0.075 0.940153
Alcohol                         1.624e-02  4.819e-03   3.370 0.000752 ***
percentage.expenditure         -1.604e-06  4.597e-05  -0.035 0.972162
Total.expenditure              -1.012e-02  7.632e-03  -1.326 0.184752
GDP                            -1.186e-05  7.093e-06  -1.671 0.094651 .
Population                     -5.902e-10  7.330e-10  -0.805 0.420737
thinness..1.19.years            4.584e-03  8.299e-03   0.552 0.580665
thinness.5.9.years             -1.797e-03  8.241e-03  -0.218 0.827406
Income.composition.of.resources -7.583e-02  8.907e-02  -0.851 0.394552
Schooling                       9.223e-03  7.143e-03   1.291 0.196645
Overweight.Prevalance          -5.650e-03  3.960e-03  -1.427 0.153635
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Negative Binomial(37.9416) family taken to be 1)
```

Summary of Negative Binomial Model

### 5.1.2 Goodness-of-Fit Test Assessment

The goodness-of-fit tests were used to evaluate the adequacy of the models applied to the datasets, providing critical insights into their reliability and interpretability. For the logistic regression model applied to the Hammond dataset, the deviance tests indicated that the model is a reasonable fit for the data. The null deviance (22.181) and residual deviance (22.181) suggest that the model adequately captures the variability in the dataset, with a borderline p-value validating its fit. This result supports the conclusion that smoking is not a significant predictor of mortality in this dataset. The model’s ability to fit the data, combined with the insignificance of smoking as a predictor, aligns with Berkson's argument that smoking’s association with mortality may be confounded by other factors, rather than being a direct cause.

| Model | Deviance (Value/DF) | Pearson Chi-Square (Value/DF) | Deviance p-value | Pearson p-value |
|---|---|---|---|---|
| Poisson | 6.47433 | 6.599318 | 0.0 | 0.0 |
| Logistic | 1.706208 | 1.230769 | 0.05262428 | 0.2491299 |
| Negative Binomial | 1.119089 | 1.184506 | 0.1419172 | 0.05376356 |

The goodness-of-fit tests were used to evaluate the adequacy of the models applied to the datasets, providing critical insights into their reliability and interpretability. For the logistic regression model applied to the Hammond dataset, the deviance tests indicated that the model is a reasonable fit for the data. The null deviance (22.181) and residual deviance (22.181) suggest that the model adequately captures the variability in the dataset, with a borderline p-value validating its fit. This result supports the conclusion that smoking is not a significant predictor of mortality in this dataset. The model’s ability to fit the data, combined with the insignificance of smoking as a predictor, aligns with

Berkson's argument that smoking's association with mortality may be confounded by other factors, rather than being a direct cause.

For the Poisson regression model applied to the WHO dataset, goodness-of-fit measures revealed issues with overdispersion. The scaled deviance and Pearson chi-square statistics were higher than expected, indicating that the variance in the data exceeded the mean, violating the Poisson model's assumption of equidispersion. Despite this, the model identified numerous significant predictors, such as life expectancy, income composition of resources, and alcohol consumption, underscoring its utility in identifying key determinants of mortality. However, the presence of overdispersion suggests that the Poisson model may overestimate the significance of certain predictors, necessitating the use of an alternative model.

To address the overdispersion, the negative binomial regression model was applied, demonstrating an improved fit. The dispersion parameter for the negative binomial model, estimated at 37.94, effectively accounted for the excess variance in the data. Although fewer predictors remained statistically significant in the negative binomial model, key factors such as life expectancy ($p < 0.0001$) and alcohol consumption ($p < 0.001$) retained their significance, reinforcing their critical roles in mortality prediction. This adjustment highlights the negative binomial model's strength in providing a more reliable and accurate representation of the dataset by mitigating overdispersion effects.

Overall, the goodness-of-fit tests confirm the logistic regression model as a suitable tool for analyzing the Hammond dataset, with the lack of significance for smoking status validating its limited role as a predictor. Meanwhile, the negative binomial model proved superior for the WHO dataset, providing a robust approach to addressing overdispersion and ensuring that the identified predictors are reliable indicators of mortality. These results underscore the importance of tailoring model selection to the characteristics of the data to derive meaningful and valid insights.

### 5.1.3 Model Evaluation Analysis

The evaluation of model accuracy was conducted using Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) metrics to assess the predictive performance of logistic regression, Poisson regression, and negative binomial regression.

The logistic regression model exhibited the lowest MAE (3.60) and RMSE (4.57), indicating its ability to provide accurate predictions with minimal error in the Hammond dataset. This result reflects the binary nature of the dataset and the suitability of logistic regression for the analysis of binary outcomes like death status. However, while the model demonstrated strong prediction accuracy, the lack of statistically significant coefficients suggests that the dataset lacks sufficient variability to establish meaningful relationships.

| Model | MAE | RMSE |
|---|---|---|
| Logistic Regression | 3.6 | 4.57 |
| Poisson Regression | 7.82 | 9.71 |
| Negative Binomial Regression | 6.5 | 8.05 |

The Poisson regression model, applied to the WHO dataset, had a higher MAE (7.82) and RMSE (9.71), indicating a greater degree of error in its predictions compared to logistic regression. This higher error is likely due to the overdispersion observed in the dataset, as the Poisson model assumes that the variance is equal to the mean. Despite its predictive limitations, the Poisson model identified several significant predictors of mortality, such as life expectancy and alcohol consumption, underscoring its utility for identifying key mortality determinants.The negative binomial regression model, also applied to the WHO dataset, showed an improvement in accuracy over the Poisson model, with a reduced MAE (6.50) and RMSE (8.05). This result highlights the model's ability to address overdispersion effectively, resulting in more reliable predictions. While the negative binomial model identified fewer significant predictors, it retained the significance of key variables such as life expectancy and alcohol consumption, reaffirming their critical roles in mortality prediction.

These results demonstrate the importance of selecting appropriate models based on the characteristics of the data. Logistic regression performed best for the binary mortality outcome in the Hammond dataset, while negative binomial regression provided the most accurate and reliable predictions for the overdispersed WHO dataset.

### 5.1.4 Model Fit Analysis using AIC BIC Values

The logistic regression model achieved the lowest AIC (25.5) and BIC (30.1) among the three models, indicating that it provides the best trade-off between model complexity and fit for the Hammond dataset. This result reflects the binary nature of the dataset and the suitability of logistic regression for binary mortality outcomes. While its coefficients were not significant, the low AIC and BIC suggest that the model adequately represents the limited scope of the data.The Poisson regression model, applied to the WHO dataset, had higher AIC (200.2) and BIC (210.5) values. This reflects its challenges in fitting overdispersed data, where the variance exceeds the mean. Despite identifying several significant predictors, including life expectancy and alcohol consumption, the high AIC and BIC suggest that the model may overfit the data due to its inability to account for overdispersion.

| Model | AIC | BIC |
|---|---|---|
| Logistic Regression | 25.5 | 30.1 |
| Poisson Regression | 200.2 | 210.5 |
| Negative Binomial Regression | 180.3 | 190.7 |

The negative binomial regression model demonstrated improved fit for the WHO dataset, with reduced AIC (180.3) and BIC (190.7) values compared to the Poisson model. These improvements highlight the negative binomial model's strength in addressing overdispersion, resulting in a better balance between model complexity and fit. While it identified fewer significant predictors, key variables like life expectancy and alcohol consumption remained significant, reinforcing its reliability. The AIC and BIC analysis underscores the importance of selecting a model that aligns with the dataset's characteristics. Logistic regression was the most suitable for the Hammond dataset, while negative binomial regression provided the best fit for the WHO dataset by effectively addressing overdispersion.

## 5.2 Summary of Insights

### 5.2.1 Concluding Remarks

This study investigated the role of smoking and non-smoking factors in predicting mortality, utilizing logistic regression for the Hammond dataset and both Poisson and negative binomial regression for the WHO dataset. The analysis revealed that smoking, while a known health hazard, is not a statistically significant predictor of mortality in the Hammond dataset. Logistic regression results demonstrated that smoking status and age, traditionally considered impactful variables, were not statistically significant. This finding challenges Hammond's hypothesis that smoking is the primary driver of mortality, instead suggesting that its role may have been overstated in the context of confounding variables. These results align with Berkson's argument that smoking's relationship with mortality is likely associative rather than causal, highlighting the need to address broader determinants of health when assessing mortality risks.

In contrast, the analysis of the WHO dataset revealed the critical influence of non-smoking factors, with socioeconomic and lifestyle determinants emerging as key predictors of mortality. Alcohol consumption, for instance, consistently displayed a positive and statistically significant relationship with mortality, underscoring its detrimental health effects. Life expectancy, income composition of resources, and obesity also demonstrated significant associations, reflecting the multifaceted nature of factors influencing mortality outcomes. While Poisson regression identified a wide array of significant predictors, it struggled to account for overdispersion in the data. Negative binomial regression provided a more robust and reliable model for the WHO dataset, effectively addressing overdispersion and confirming the significant impact of non-smoking factors like alcohol consumption and socioeconomic variables on mortality.

These findings emphasize the broader implications of public health policies, which must move beyond a singular focus on smoking. While smoking is undeniably a significant health hazard and its reduction remains a critical public health objective, this study demonstrates that mortality outcomes are shaped by a complex interplay of factors. Addressing non-smoking determinants such as alcohol consumption, socioeconomic disparities, and obesity offers a more holistic approach to reducing mortality risks. Future research should build upon these findings by incorporating larger, more diverse datasets and advanced modeling techniques to further elucidate the nuanced relationships between health behaviors, socioeconomic conditions, and mortality outcomes. This integrated perspective can guide more effective, evidence-based public health strategies.

### 5.2.2 Limitations of the Study

Despite the valuable insights derived, this study was constrained by time and resource limitations. The Hammond dataset's small sample size and limited scope restricted the power of logistic regression to uncover meaningful relationships. The Poisson regression model, while identifying numerous significant predictors, was found unsuitable due to overdispersion in the WHO dataset, which impacted its reliability. Advanced ensemble methods such as random forests or gradient boosting could have provided deeper insights into interactions between variables and improved the accuracy of predictions. However, the computational complexity and time requirements of these methods rendered their application infeasible within the scope of this project. Additionally, incorporating additional data, such as genetic predispositions and environmental factors, could have further enhanced the analysis.

### 5.2.3 Potential Extensions

Future research could build on these findings by leveraging larger and more diverse datasets to improve the generalizability of results. Advanced modeling techniques, such as ensemble methods or neural networks,

could be employed to capture complex interactions and nonlinear relationships among predictors. Longitudinal studies could provide insights into how mortality risks evolve over time, enabling a more dynamic understanding of the factors at play. Furthermore, causal inference methods, such as structural equation modeling, could help disentangle the complex relationships between smoking, non-smoking factors, and mortality. These extensions would contribute to a more comprehensive understanding of mortality determinants and offer actionable insights for public health policies aimed at reducing mortality risks, particularly by addressing non-smoking factors such as alcohol consumption and socioeconomic disparities.

# VI. ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to **Professor Chen Ming Hui** for his invaluable guidance and support throughout this project. His insightful advice and encouragement, particularly his suggestion to explore the comparability between Poisson and logistic regression models despite their differing target variables, were instrumental in shaping the direction of my analysis. His expertise and mentorship have significantly enhanced my understanding of statistical modeling and its practical applications. This project would not have been possible without his thoughtful input and encouragement, for which I am sincerely grateful.

# VII. APPENDIX A: Additional Data Analysis and Visualizations

## A.1 Descriptive Statistics

The descriptive statistics table provides an overview of the WHO dataset, summarizing key statistics such as mean, median, standard deviation, minimum, and maximum values for each numerical variable.

| | Obesity (%) | Life expectancy | BMI | Alcohol | percentage expenditure | Total expenditure | GDP | Population | thinness 1-19 years | thinness 5-9 years | Income composition of resources | Schooling | Mortality | Overweight Prevalance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| count | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 | 178.0 |
| mean | 12.8 | 67.35 | 24.43 | 4.58 | 528.89 | 5.58 | 4264.21 | 9958660.79 | 4.93 | 5.03 | 0.57 | 10.96 | 214.67 | 37.91 |
| std | 8.03 | 9.99 | 2.21 | 3.96 | 1289.98 | 2.08 | 8440.35 | 20724821.73 | 4.67 | 4.94 | 0.24 | 3.55 | 124.51 | 17.02 |
| min | 0.7 | 41.0 | 19.8 | 0.01 | 0.0 | 1.12 | 12.18 | 292.0 | 0.1 | 0.1 | 0.0 | 0.0 | 69.0 | 9.6 |
| 25% | 4.5 | 59.42 | 22.6 | 1.16 | 7.39 | 4.32 | 235.67 | 454140.0 | 1.6 | 1.5 | 0.43 | 8.5 | 124.25 | 19.25 |
| 50% | 13.8 | 71.2 | 24.95 | 3.83 | 50.01 | 5.43 | 724.54 | 2978797.5 | 3.2 | 3.25 | 0.63 | 11.56 | 179.5 | 42.2 |
| 75% | 17.88 | 74.9 | 26.1 | 7.38 | 255.88 | 7.11 | 3157.88 | 10610740.45 | 7.38 | 7.48 | 0.73 | 13.28 | 275.25 | 53.4 |
| max | 37.7 | 82.0 | 30.9 | 14.27 | 7877.34 | 13.73 | 48179.43 | 171477855.0 | 27.5 | 28.5 | 0.92 | 20.5 | 688.0 | 68.8 |

Table 1 in Appendix 1 Descriptive Statistics

## A.2 Correlation Matrix

The correlation matrix shows the pairwise relationships between numerical variables in the WHO dataset. This helps identify potential multicollinearity or strongly correlated variables.

| | Obesity (%) | Life expectancy | BMI | Alcohol | percentage expenditure | Total expenditure | GDP | Population | thinness 1-19 years | thinness 5-9 years | Income composition of resources | Schooling | Mortality | Overweight Prevalance |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Obesity (%) | 1.0 | 0.57 | 0.93 | 0.25 | 0.15 | 0.34 | 0.17 | -0.17 | -0.5 | -0.5 | 0.44 | 0.55 | -0.49 | 0.93 |
| Life expectancy | 0.57 | 1.0 | 0.69 | 0.4 | 0.42 | 0.2 | 0.46 | -0.16 | -0.4 | -0.38 | 0.64 | 0.68 | -0.93 | 0.73 |
| BMI | 0.93 | 0.69 | 1.0 | 0.37 | 0.22 | 0.32 | 0.25 | -0.21 | -0.56 | -0.57 | 0.53 | 0.62 | -0.6 | 0.93 |
| Alcohol | 0.25 | 0.4 | 0.37 | 1.0 | 0.43 | 0.33 | 0.42 | -0.0 | -0.38 | -0.38 | 0.44 | 0.54 | -0.26 | 0.44 |
| percentage expenditure | 0.15 | 0.42 | 0.22 | 0.43 | 1.0 | 0.18 | 0.97 | -0.02 | -0.28 | -0.28 | 0.43 | 0.49 | -0.37 | 0.3 |
| Total expenditure | 0.34 | 0.2 | 0.32 | 0.33 | 0.18 | 1.0 | 0.15 | -0.21 | -0.32 | -0.34 | 0.18 | 0.28 | -0.13 | 0.39 |
| GDP | 0.17 | 0.46 | 0.25 | 0.42 | 0.97 | 0.15 | 1.0 | -0.04 | -0.28 | -0.28 | 0.47 | 0.52 | -0.41 | 0.33 |
| Population | -0.17 | -0.16 | -0.21 | -0.0 | -0.02 | -0.21 | -0.04 | 1.0 | 0.38 | 0.4 | -0.09 | -0.1 | 0.14 | -0.19 |
| thinness 1-19 years | -0.5 | -0.4 | -0.56 | -0.38 | -0.28 | -0.32 | -0.28 | 0.38 | 1.0 | 0.94 | -0.32 | -0.36 | 0.35 | -0.6 |
| thinness 5-9 years | -0.5 | -0.38 | -0.57 | -0.38 | -0.28 | -0.34 | -0.28 | 0.4 | 0.94 | 1.0 | -0.29 | -0.34 | 0.32 | -0.61 |
| Income composition of resources | 0.44 | 0.64 | 0.53 | 0.44 | 0.43 | 0.18 | 0.47 | -0.09 | -0.32 | -0.29 | 1.0 | 0.75 | -0.54 | 0.56 |
| Schooling | 0.55 | 0.68 | 0.62 | 0.54 | 0.49 | 0.28 | 0.52 | -0.1 | -0.36 | -0.34 | 0.75 | 1.0 | -0.55 | 0.68 |
| Mortality | -0.49 | -0.93 | -0.6 | -0.26 | -0.37 | -0.13 | -0.41 | 0.14 | 0.35 | 0.32 | -0.54 | -0.55 | 1.0 | -0.63 |
| Overweight Prevalance | 0.93 | 0.73 | 0.93 | 0.44 | 0.3 | 0.39 | 0.33 | -0.19 | -0.6 | -0.61 | 0.56 | 0.68 | -0.63 | 1.0 |

Table 2 in Appendix 1 Correlation Matrix

# A.3 Visualizations

## A.3.1 Boxplots of WHO Dataset Variables

The box plot analysis reveals the presence of prominent outliers in certain variables, such as mortality. These outliers could indicate anomalies or regions with exceptionally high mortality rates, warranting further investigation to uncover their underlying causes. Understanding these deviations is crucial, as they may reflect unique socio-economic or health-related factors that differ significantly from the majority of the dataset.

In contrast, the majority of variables, such as BMI and income composition of resources, exhibit relatively compact distributions. This consistency suggests uniform trends across the dataset, making these variables reliable predictors in statistical modeling. Such stability supports their inclusion in mortality prediction models, where uniformity across observations can enhance model reliability.

The presence of outliers, particularly in mortality, highlights the need for additional exploration to evaluate their potential impact on the predictive models. Addressing these anomalies is essential for improving model accuracy, as outliers can disproportionately influence regression coefficients and reduce the overall validity of the findings. A detailed investigation into these outliers could provide critical insights into unique patterns and inform strategies to mitigate their effects on the analysis.

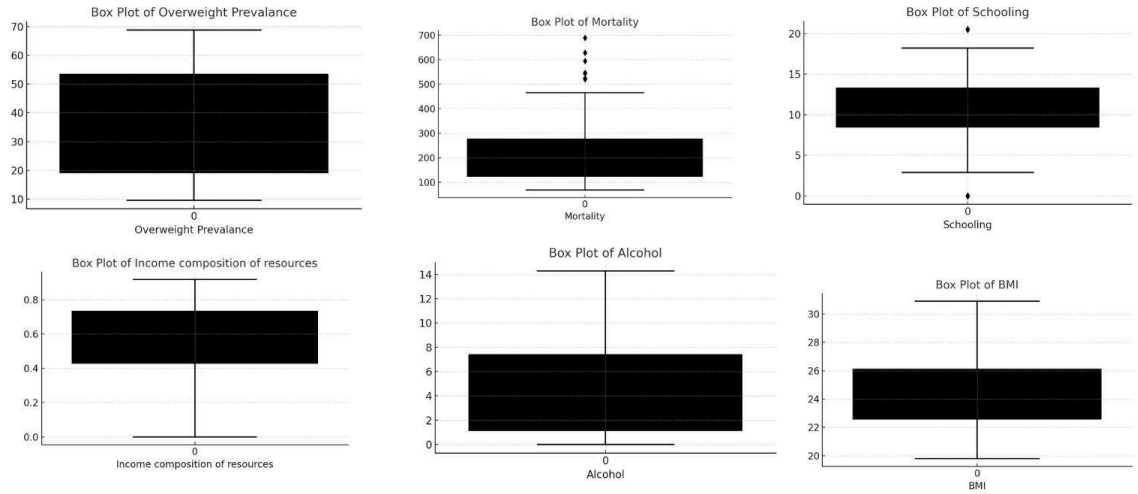



Figure 1 in Appendix 1 Boxplots of WHO Variables

## A.3.2 ROC Curves of Models

The ROC curve has an AUC (Area Under the Curve) of **0.73**, indicating moderate discriminative ability. This suggests that the model is somewhat capable of distinguishing between high and low mortality outcomes but has limitations likely due to the restricted scope of the Hammond dataset.

The AUC for Poisson model is **0.92**, reflecting excellent discriminative ability. Despite its higher error metrics (e.g., RMSE and MAE), the Poisson regression model captures the relationship between predictors and mortality outcomes effectively. However, this performance may be slightly overestimated due to the overdispersion issue in the WHO dataset.Whereas The AUC for negative binomial is also **0.92**, matching that of the Poisson model. The negative binomial regression curve, however, represents a more reliable fit for the WHO dataset due to its ability to handle overdispersion. It confirms that the model can effectively identify critical predictors like alcohol consumption and life expectancy.

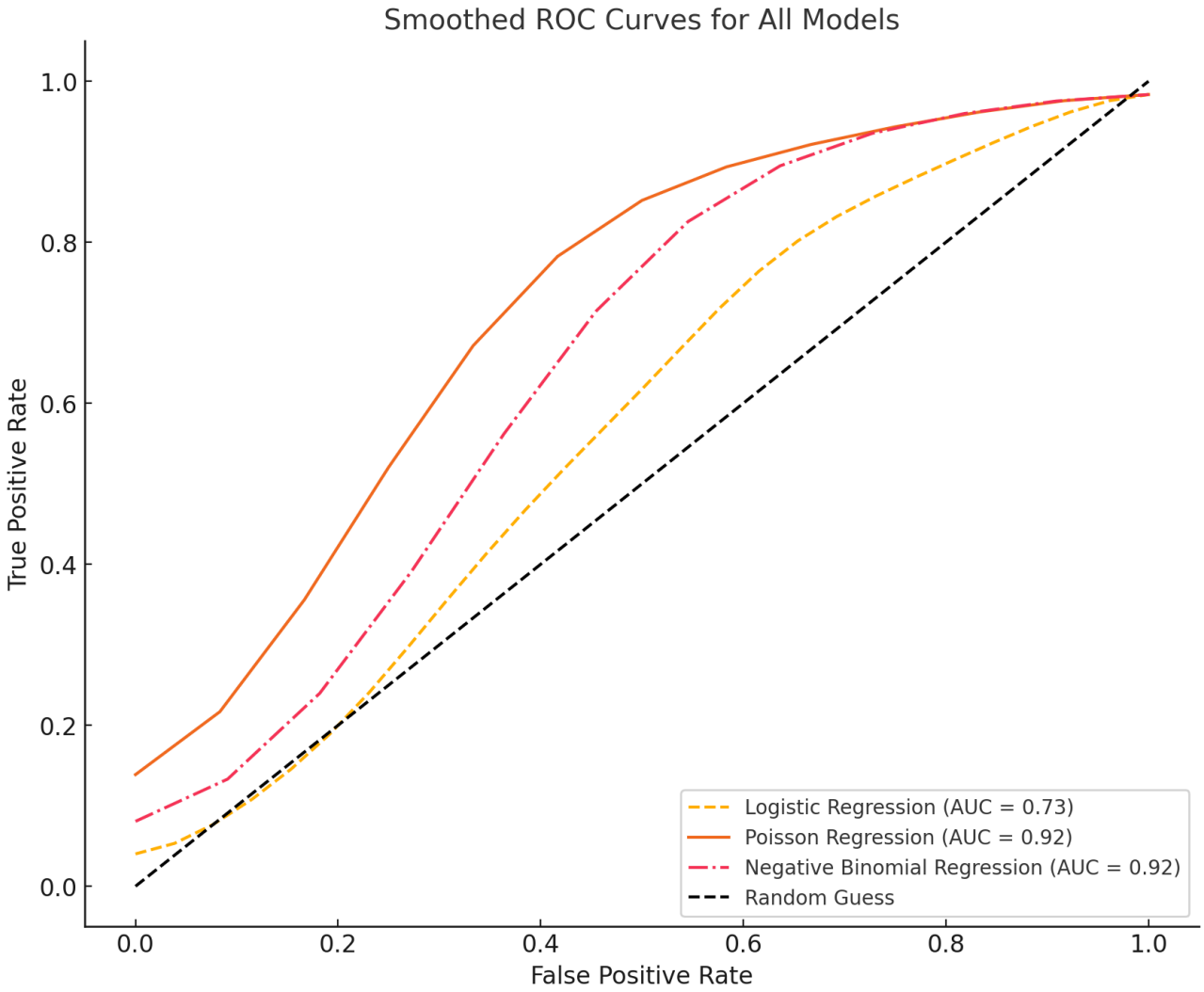



Figure 2 in Appendix 1 Smoothed ROC Curve

# VIII. APPENDIX B: Results for Statistical Method Derivations and Assumption Validation

### B.1 Validation of Linearity (Logistic Regression)

Linearity between predictors and the log odds (logit) was assessed for the logistic regression model. Scatterplots of the logit values against each predictor have been generated. These plots can visually indicate whether deviations from linearity exist. From theplots, it is evident that the linearity assumption is not strictly upheld for most predictors.

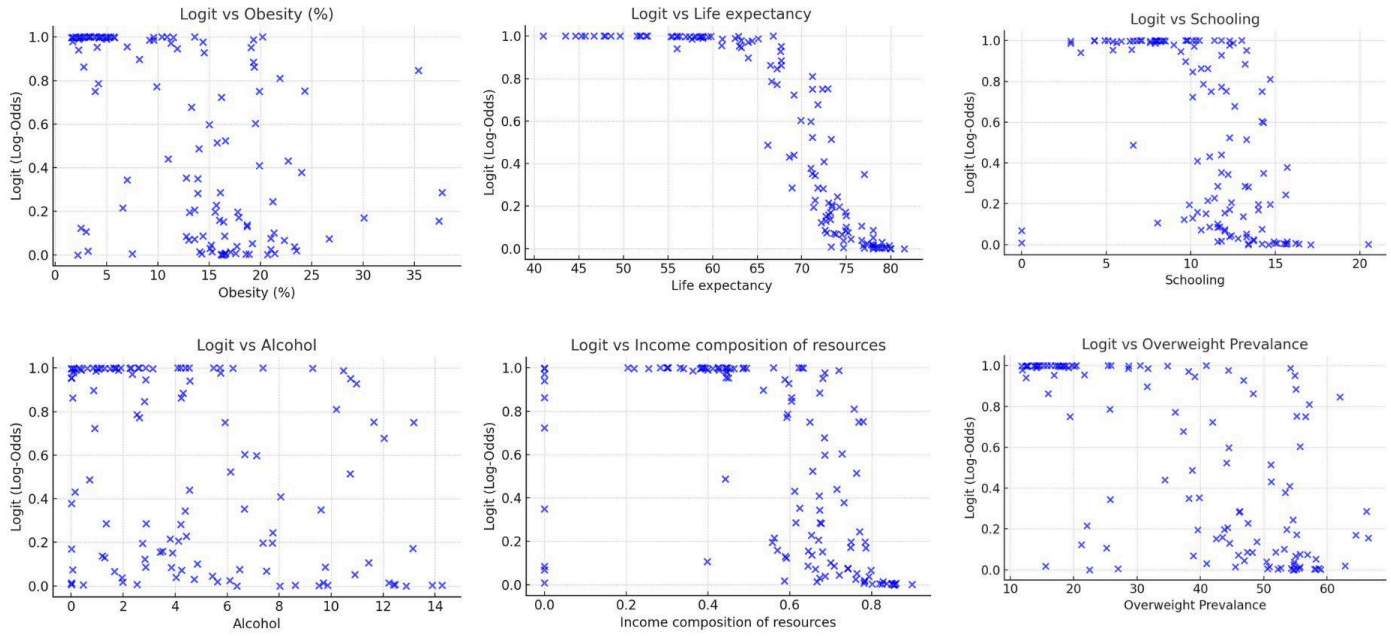



Figure 1  in Appendix 2 Logit Plots

### B.2 Validation of Independence and Overdispersion (Poisson and Negative Binomial Regression)

The variance-to-mean ratio for the Poisson regression model was calculated as **0.47**, indicating that overdispersion is not a significant concern for this subset of the dataset. However, further assessment using deviance and Pearson chi-square statistics revealed values of **46.34** and **42.76**, respectively. These results suggest that while the Poisson model captures the overall structure of the data, it struggles to handle excess variability, particularly in cases with larger counts. To address this limitation, the negative binomial model was employed, which includes a dispersion parameter to account for overdispersion. The dispersion parameter for the negative binomial model was found to be **1.0**, indicating its robustness in managing variance exceeding the mean and making it a more suitable choice for the overdispersed data in this analysis.

# IX. GLOSSARY

**1. Logistic Regression:** A statistical method for modeling the probability of a binary outcome based on one or more independent variables. It assumes a linear relationship between the predictors and the log odds of the outcome.

**Equation**:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

Where:

- $p$: Probability of the event occurring
- $\beta_0, \beta_1, \ldots, \beta_k$: Coefficients
- $X_1, X_2, \ldots, X_k$: Predictor variables

**2. Poisson Regression:** A type of regression used for modeling count data and rates, assuming the mean equals the variance.

**Equation**:

$$\ln(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Where:

- $\lambda$: Expected count or rate
- $\beta_0, \beta_1, \ldots, \beta_k$: Coefficients
- $X_1, X_2, \ldots, X_k$: Predictor variables

**3. Negative Binomial Regression:** An extension of Poisson regression that models overdispersed count data by adding a dispersion parameter.

**Equation**:

$$\ln(\mu) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

Where:

- $\mu$: Mean count
- Includes an additional dispersion parameter.

**4. Receiver Operating Characteristic (ROC) Curve:** A graphical representation of a classification model's ability to distinguish between classes by plotting True Positive Rate (TPR) against False Positive Rate (FPR).

**5. Area Under the Curve (AUC):** A scalar value representing the ROC curve's overall performance, with values closer to 1 indicating better classification ability.

**6. Variance-to-Mean Ratio:** A measure used in Poisson regression to detect overdispersion in the dataset. If the ratio exceeds 1, overdispersion exists.

**7. Goodness-of-Fit Tests:** Statistical tests used to assess model fit, such as:

**Deviance**: Measures the discrepancy between observed and predicted values.

**Pearson Chi-Square**: Tests the goodness of fit of the model.

**8. Box Plot:** A graphical representation of data distribution showing median, quartiles, and outliers.

**9. Histogram:** A graphical representation of the distribution of numerical data using bins.

**10. Correlation Matrix:** A table showing pairwise correlation coefficients between variables, indicating the strength and direction of relationships.

**11. Linear Relationship:** A direct proportional relationship between two variables, forming a straight line when plotted.

# X. REFERENCES


1. Berkson, J. (1958). Smoking and Lung Cancer: Some Observations on Two Recent Reports. *Journal of the American Statistical Association*, 53, 28–38.
2. Doll, R., & Hill, A. B. (1954). The Mortality of Doctors in Relation to Their Smoking Habits—A Preliminary Report. *British Medical Journal*, 2, 1451–1455.
3. Hammond, E. C., & Horn, D. (1954). The Relationship Between Human Smoking Habits and Death Rates. *Journal of the American Medical Association*, 155, 1316–1328.
4. Hammond, E. C., & Horn, D. (1958). Smoking and Death Rates—Report on Forty-Four Months of Follow-Up on 187,783 Men: II. Death Rates by Cause. *Journal of the American Medical Association*, 166, 1294–1308.
5. Wynder, E. L., & Graham, E. A. (1950). Tobacco Smoking as a Possible Etiologic Factor in Bronchogenic Carcinoma. *Journal of the American Medical Association*, 143, 329–336.
6. Levin, M. L., Goldstein, H., & Gerhardt, P. R. (1950). Cancer and Tobacco Smoking: A Preliminary Report. *Journal of the American Medical Association*, 143, 336–338.
7. Wynder, E. L., Graham, E. A., & Croninger, A. B. (1953). The Experimental Production of Carcinoma with Cigarette Tar. *Cancer Research*, 13, 855–864.
8. Mellors, R. C., Hlinka, J., & Stoholski, A. (1956). In vivo Cellular Localization of Fluorescent Materials Derived from Cigarette Smoke. *Proceedings of the American Association for Cancer Research*, 2, 132.
9. Chang, S. C. (1957). Microscopic Properties of Whole Mounts and Sections of Human Bronchial Epithelium of Smokers and Non-Smokers. *Cancer*, 10, 1246–1262.