Bayesian Analysis of Brain Cancer Incidence in the United States (1999-2021)

Vasudha Rohatgi
STATS 551 Bayesian Data Analysis Project
Submitted : 18th December 2024

Abstract

This report provides a Bayesian analysis of brain cancer incidence data in the United States from 1999 to 2021, with a focus on the environmental, occupational, and lifestyle factors that may affect the incidence of brain cancer. We utilize Bayesian regression techniques with Poisson distributions to model the total cancer incidence count and explore the relationships with various air quality metrics and demographic variables.

All figures generated are reported in the Quarto code and then at the end of the document as separate figures.

---

1. Introduction

Cancer incidence and mortality rates are influenced by a wide range of factors, including environmental, occupational, and lifestyle factors. The objective of this analysis is to investigate how these factors impact brain cancer incidence and mortality across different states in the United States over time (1999-2021). We specifically focus on using Bayesian modeling to quantify the uncertainty around the relationships between cancer incidence and various predictors such as air quality, population, and geographical features.

---

2. Data Description

The dataset includes information on cancer incidence, total population, and various air quality measures (e.g., AQI, days with unhealthy air quality) across all U.S. states for each year between 1999 and 2021. The primary factor which we examine here in it's relation to the rate of incidence of brain cancer is the air quality index. The data is structured with columns for the following variables:

- **`States`**: U.S. state names
- **`year`**: Year of the observation
- **`avg_max_aqi`**: Average maximum AQI per year
- **`avg_moderate_days`**: Average number of moderate AQI days
- **`avg_unhealthy_days`**: Average number of unhealthy AQI days
- **`Total_Count`**: Total number of brain cancer incidences
- **`Total_Population`**: Total population of the state in a given year

---

3. Modeling Approach

We begin by modeling the total cancer incidence count as a Poisson-distributed variable, with a log link function. The regression model includes environmental and demographic predictors, such as AQI, the number of unhealthy days, and the population. The model is defined as follows:

$$\boldsymbol{Incidence}_i \sim Poisson(\lambda_i)$$

where the log of the rate parameter λ_i is modeled as:

$$\log(\lambda\_i) = \alpha + \beta_1 \cdot avg\_\max\_aqi\_i + \beta_2 \cdot avg\_unhealthy\_days\_i + \beta_3 \cdot avg\_moderate\_days\_i + \ldots$$

Here, α is the intercept, and β_j are the regression coefficients for each predictor.

Prior Distributions

We use the following prior distributions for the model parameters:

- $\alpha \sim N(0, 10)$
- $\beta_j \sim N(0, 5)$
- $\sigma \sim N(0, 1)$
- $\lambda \sim Gamma(2, 0.1)$

These priors are chosen to reflect relatively weak assumptions about the model parameters, allowing the data to influence the posterior distribution.

---

4. Model Fitting and Inference

We fit the model using the Stan programming language and the `rstan` package in R. The model is compiled and sampled using 4 chains, with 2000 iterations per chain. Attached is the R code used to compile and fit the model:

5. Results

Posterior Summaries

We examined the posterior distributions of the regression coefficients. The coefficient estimates represent the average change in the log incidence rate for each one-unit increase in the predictor variable, adjusted for the other variables in the model.

(Insert figure here: Posterior distribution of the regression coefficients)

Model Diagnostics

The trace plots for the parameters showed good mixing, and the R-hat statistics were all close to 1, indicating that the chains had converged.

Posterior Predictive Checks

To assess the model fit, we performed posterior predictive checks by comparing the observed and predicted incidence counts. The model adequately captures the variability in the data, with the predictions closely matching the observed values.

---

6. Discussion

The Bayesian model reveals that factors such as air quality (measured by `avg_max_aqi` and `avg_unhealthy_days`) have a significant effect on brain cancer incidence. These findings suggest that states with worse air quality may experience higher cancer incidences, although further analysis is needed to confirm these relationships.

In future work, we could explore more complex models, such as hierarchical models, to account for state-level variations and improve predictive accuracy.

---

7. Conclusion

This study used Bayesian Poisson regression to model brain cancer incidence across U.S. states, incorporating environmental and demographic factors. The model provides valuable insights into the factors influencing cancer incidence and demonstrates the utility of Bayesian methods for understanding complex relationships in health data.

8. References

"AirData Website File Download Page." Data & Tools. Accessed December 18, 2024. https://aqs.epa.gov/aqsweb/documents/data_api.html#signup.

"AirData Website File Download Page." Data & Tools. Accessed December 18, 2024. https://aqs.epa.gov/aqsweb/airdata/download_files.html#Annual.

Baker, Nancy T. "Estimated Annual Agricultural Pesticide Use." U.S. Geological Survey, 2014. https://doi.org/10.5066/F7NP22KM.

———. "Estimated Annual Agricultural Pesticide Use." U.S. Geological Survey, 2014. https://doi.org/10.5066/F7NP22KM.

"Data Series." Data Series. Data Series, 2015.

"Data Series." Data Series. Data Series, 2015.

"NCHS - Age-Adjusted Death Rates for Selected Major Causes of Death | Data | Centers for Disease Control and Prevention." Accessed December 18, 2024. https://data.cdc.gov/NCHS/NCHS-Age-adjusted-Death-Rates-for-Selected-Major-C/6rkc-nb2q/about_data.

"NCHS - Leading Causes of Death: United States | Data | Centers for Disease Control and Prevention." Accessed December 18, 2024. https://data.cdc.gov/NCHS/NCHS-Leading-Causes-of-Death-United-States/bi63-dtpu/about_data.

"NVSS - Leading Causes of Death," May 9, 2024. https://www.cdc.gov/nchs/nvss/leading-causes-of-death.htm.

Toccalino, Patricia L. ; Norman. "Health-Based Screening Levels for Evaluating Water-Quality Data." U.S. Geological Survey, 2014. https://doi.org/10.5066/F71C1TWP.

———. "Health-Based Screening Levels for Evaluating Water-Quality Data." U.S. Geological Survey, 2014. https://doi.org/10.5066/F71C1TWP.

"Underlying Cause of Death, 1999-2020 Request." Accessed December 18, 2024. https://wonder.cdc.gov/ucd-icd10.html.

"Underlying Cause of Death 2018-2022 by Single Race." Accessed December 18, 2024. https://wonder.cdc.gov/wonder/help/ucd-expanded.html#ICD-10%20113%20Cause%20List.

"United States Cancer Statistics: Incidence Public Information Data." Accessed December 18, 2024. https://wonder.cdc.gov/wonder/help/cancer-v2021.html#How%20are%20age-adjusted%20rates%20calculated?

US EPA, ORD. "EPA's Report on the Environment (ROE)." Reports and Assessments, February 6, 2015. https://www.epa.gov/report-environment.

———. "Report on the Environment (ROE)." Collections and Lists. US EPA, February 6, 2015. https://www.epa.gov/report-environment.

"US EPA Search." Accessed December 18, 2024. https://search.epa.gov/epasearch/?querytext=data&areaname=&areacontacts=&areasearchurl=&typeofsearch=epa&result_template=#/.

Wieben, Christine M. "Estimated Annual Agricultural Pesticide Use by Major Crop or Crop Group for States of the Conterminous United States, 1992-2019." U.S. Geological Survey, 2021. https://doi.org/10.5066/P900FZ6Y.

———. "Estimated Annual Agricultural Pesticide Use by Major Crop or Crop Group for States of the Conterminous United States, 1992-2019." U.S. Geological Survey, 2021. https://doi.org/10.5066/P900FZ6Y.

Wieben, Christine M. "Estimated Annual Agricultural Pesticide Use for Counties of the Conterminous United States, 2013-17 (Ver. 2.0, May 2020)." U.S. Geological Survey, 2020.