Hierarchical Normal Model for Deaths among 25-34 Year Olds

Introduction

This analysis investigates the leading causes of death among individuals aged 25-34 in the United States using a hierarchical Bayesian model. The data consists of annual death counts for five leading causes over the period 2018–2022. The goal is to model the annual deaths for these causes, estimate the posterior distributions of the mean number of deaths for each cause, and provide a predictive distribution for a hypothetical sixth cause of death. A hierarchical normal model with Gibbs sampling is employed to infer the posterior distributions.

Methodology

Data Description

The dataset contains annual death counts from 2018 to 2022 for five leading causes of death among individuals aged 25-34. The causes are:

- Accidents
- Homicide
- Suicide
- Heart Disease
- Malignant Neoplasms

Model Specifications

The model used for this analysis is a hierarchical normal model. For each cause *i* and year *j*, the number of deaths y_{ij} is assumed to be normally distributed with a mean μ_i and a common variance σ^2:

$$y_{ij} \sim N(\mu_i, \sigma^2)$$

The means μ_i for each cause are modeled as being drawn from a global mean μ_0 with variance τ^2:

$$\mu_i \sim N(\mu_0, \tau^2)$$

Priors for the parameters μ_0, τ^2, and σ^2 are set as weakly informative.

Gibbs Sampler Setup

A Gibbs sampling procedure was employed to estimate the posterior distributions of the parameters. The Gibbs sampler was run for 10,000 iterations, and the first 2,000 iterations were discarded as burn-in. Weakly informative priors were set for the variance parameters. The initial values for μ_0, τ^2, and σ^2 were set based on the data's overall mean and variance.

Results

Posterior Distribution of Annual Deaths (5 Causes)

Accidents have the highest posterior mean (27,498.39), with the credible interval strongly positive.

Homicide and **Suicide** also have substantial posterior means but with much wider credible intervals.

Heart Disease and **Malignant Neoplasms** have lower posterior means, with the latter showing a credible interval that includes negative values, indicating considerable uncertainty around the estimate.

	Posterior Mean	**Lower 95% CI**	**Upper 95% CI**
Accidents	**27498.39**	**21515.22**	**32286.60**
Homicide	**11211.55**	**6640.61**	**15650.72**
Suicide	**8735.34**	**4349.60**	**13269.76**
Heart Disease	**4679.48**	**158.15**	**9370.86**
Malignant Neoplasms	**4444.52**	**-76.00**	**9135.32**

Predictive Distribution for the 6th Cause

Using the posterior estimates for μ_0, τ^2, and σ^2, the predictive distribution was simulated for a hypothetical sixth cause of death among 25-34-year-olds in the U.S.

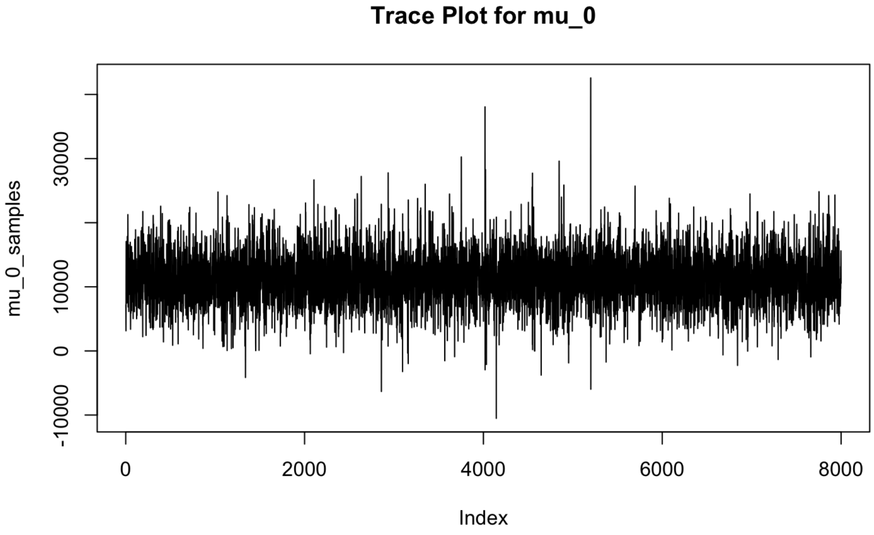
The predicted mean number of deaths for this 6th cause is **11,405.74**. The 95% credible interval for this prediction is **[-7,311.34, 28,549.61]**.

This wide credible interval reflects the substantial uncertainty around the estimate, with the possibility of observing fewer than 0 deaths (negative lower bound, which should be interpreted as a near-zero lower bound in practical terms) to as many as 28,549 deaths in a year for the hypothetical cause. This uncertainty arises due to the variances in the model: the within-cause variance (σ^2) and the between-cause variance (τ^2).

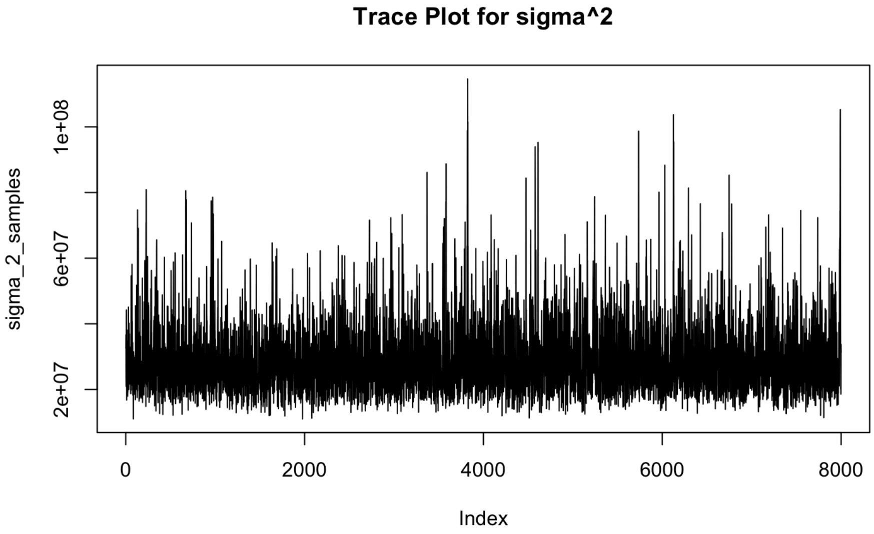
The predicted mean (11,405.74) is close to the global mean μ_0 (11,337.9), as expected because the 6th cause is modeled as part of the same population as the five observed causes. The large credible interval suggests significant uncertainty, indicating that this model predicts a wide range of possible outcomes for the number of deaths for this hypothetical cause.

Convergence Diagnostics

The μ_0 trace plot indicates the values of μ_0 bouncing around a central region, showing a good sign of stationarity and suggesting that the chain is exploring the parameter space well. No long-term trends are seen, suggesting the chain is not drifting and the sampler is likely converging correctly. Additionally, the chain appears to mix well since it moves through different values of μ_0 without repeating in a specific region. While there are occasional spikes, they do not seem to be the norm.



The values of σ^2 also fluctuate around a central region, indicating that the chain has settled into its stationary distribution. Again, this is a good sign that the sampler is exploring the posterior distribution adequately. The σ^2 trace plot has more noticeable spikes than the μ_0 trace plot. The spikes are relatively spread out and do not appear to be problematic. However, future analysis should ensure that the spikes do not heavily influence the posterior inference. The chain for σ^2 appears to mix fairly well and no long-term trends are noticeable, suggesting the parameter space has been properly explored and the chain has likely converged.



4. Discussion

The hierarchical Bayesian analysis highlights **Accidents** as the leading cause of death among 25-34-year-olds, with a posterior mean of 27,498 deaths. **Homicide** and **Suicide** also show significant numbers, but their wider credible intervals reflect more uncertainty. **Heart Disease** and **Malignant Neoplasms** have much lower posterior means, and in the case of Malignant Neoplasms, the credible interval includes negative values, indicating high uncertainty.

The predictive distribution for the 6th cause, with a mean of 11,405 and a 95% credible interval ranging from -7,311 to 28,550, shows substantial variability. This wide interval stems from the large variances in the model, indicating that predictions for an unseen cause are highly uncertain.

Convergence diagnostics, including trace plots, show that the chains have reached stationarity with no long-term trends. Occasional spikes in the variance parameter σ^2 are noted but do not appear to heavily affect overall convergence. However, more comprehensive diagnostics like effective sample size (ESS) and autocorrelation checks could further verify robustness.

The key limitation of this analysis is the small sample size (5 years of data), which leads to wide credible intervals, especially for less common causes. While the hierarchical model helps in borrowing strength across causes, more data would significantly reduce uncertainty.

5. Conclusion

This analysis confirms **Accidents** as the predominant cause of death for 25-34-year-olds, with **Homicide** and **Suicide** also contributing significantly. The model shows large uncertainties for less frequent causes, such as **Malignant Neoplasms**. The predictive distribution for a hypothetical 6th cause suggests high variability, with a wide credible interval.

While the hierarchical model is effective in handling limited data and pooling information, the small sample size leads to substantial uncertainty. More data would improve the reliability of the estimates and predictions. Future analyses should consider longer time periods to reduce this uncertainty and improve the model’s precision.