Scientific Research Publishing

Modern Economy, 2026, 17(4), 587-602
https://www.scirp.org/journal/me
ISSN Online: 2152-7261
ISSN Print: 2152-7245

Bayesian Hierarchical Analysis of US Homeowners' Insurance Markets: Integrating Socioeconomic Indicators with Actuarial Risk Metrics

Prabuddha Sanyal

Straightdeal Mortgage, Irvine, California, USA
Email: prabuddha@straightdealmortgage.com

How to cite this paper: Sanyal, P. (2026). Bayesian Hierarchical Analysis of US Homeowners' Insurance Markets: Integrating Socioeconomic Indicators with Actuarial Risk Metrics. *Modern Economy, 17,* 587-602.
https://doi.org/10.4236/me.2026.174031

Received: January 23, 2026
Accepted: April 25, 2026
Published: April 28, 2026

Abstract

This study presents a comprehensive Bayesian hierarchical analysis of the U.S. homeowners' insurance market from 2018 to 2022. By integrating federal insurance metrics with American Community Survey (ACS) income data, we model the drivers of premiums across more than 25,000 ZIP codes. Our findings reveal that policy cancellations and temporal trends are the dominant drivers of premium variation, while geographic and socioeconomic factors play secondary roles. Comparative regularization analysis shows that the Bayesian model with Lasso priors exhibits strong shrinkage, effectively reducing variables such as ZIP Code and Income to near-zero coefficients, whereas the Ridge model retains all predictors with moderate shrinkage to offer a more balanced view of the feature space. The Ridge and Pitman-Yor models achieved the highest predictive accuracy, explaining approximately 7% of variance, compared with 4.8% for the Dirichlet Process Mixture model. The Bayesian model with Lasso priors was used primarily for variable selection rather than predictive accuracy, and its R^2 is therefore not directly compared here. Notably, both the Dirichlet Process Mixture (DPM) and Pitman-Yor models identified only a single unified cluster, indicating that insurance risk exists on a continuous spectrum rather than in distinct, isolated risk pools. This finding was further confirmed by the Normal Random Intercepts Model (NREM), which showed minimal variance explained by decile groupings. Across all models, cancellation rates, particularly non-payment and other cancellations, and year consistently emerge as the most influential predictors of homeowners' insurance premiums, though the magnitude of these effects varies according to the specific regularization approach employed. These results suggest that market stability and policy retention are more critical to understanding premium behavior than traditional geographic or income-based risk segmentation.

Keywords

Bayesian Hierarchical Modeling, Claim Frequency and Severity, American Community Survey, Homeowners' Insurance

1. Introduction

The homeowners' insurance market is currently facing unprecedented volatility driven by shifting climate patterns, rising construction costs, and evolving socioeconomic landscapes. These factors have intensified the complexity of accurately pricing premiums, a task vital for maintaining market stability and ensuring fairness for consumers. Premiums can vary substantially even between neighboring areas, reflecting nuanced geographic and socioeconomic influences that traditional actuarial models often fail to capture fully.

This study analyzes homeowners' insurance premiums from 2018 to 2022 across more than 25,000 ZIP codes in the United States of America (USA), integrating federal insurance metrics with American Community Survey (ACS) data at a granular ZIP code level. Our primary focus is on modeling the relationship between average household income and insurance premiums using a Bayesian hierarchical premium regression framework. To enhance model flexibility and robustness, we employ Bayesian Lasso priors for coefficient shrinkage and variable selection, alongside Dirichlet and Pitman-Yor process priors to capture latent clustering and heterogeneity across ZIP codes. These advanced Bayesian methods allow the model to adaptively identify key predictors and uncover complex grouping structures without imposing restrictive parametric assumptions.

While the broader research agenda includes joint modeling of claim frequency and severity, nonlinear income effects, and spatially structured random effects to explicitly capture geographic dependence, the current analysis prioritizes a transparent and interpretable baseline premium regression model. ZIP-code level random intercepts are included to account for unobserved heterogeneity in the data. This baseline provides a foundation for future work that will extend the modeling framework to incorporate nonlinear income effects, spatially structured random effects (e.g., BYM2 or ICAR models), and joint-severity estimation.

By producing full posterior distributions for premiums and model parameters, this study advances uncertainty quantification in ratemaking, supporting regulatory review and consumer fairness. The results offer robust insights into the income-premium gradient while setting the stage for more detailed decomposition of risk drivers and spatial patterns in subsequent research.

In Section 2, we provide a brief overview of the existing literature on insurance determinants that have used Bayesian hierarchical models in analysis. Section 3 provides the data and method used in this study. Section 4 presents the main results of the study. Section 5 concludes with a few policy implications on what can be done to address the rising costs of homeowners' insurance.

2. An Overview of the Literature

Determining the factors that influence insurance costs using Bayesian hierarchical analysis is rather few in the insurance industry. We provide a few studies that uses Bayesian models to examine policy premiums and its determinants.

One of the first studies examining the issue affecting homeowners' insurance was by Kunreuther and Pauly (2006). For homeowners, the authors highlight a significant "demand-side" problem where individuals often ignore low-probability, high-consequence events until a disaster actually occurs. This behavior is driven by a sequential model of choice: if the perceived likelihood of a disaster falls below a certain "threshold of concern," homeowners simply assume it won't happen to them and fail to purchase insurance or invest in protective measures. Even when insurance is required by lenders, many homeowners cancel their policies after a few years if they haven't made a claim, viewing the premiums as a "wasted investment" rather than a necessary protective cost.

The findings of the paper reveal several key insights into the challenges and dynamics of homeowners' insurance in the face of catastrophic risks. One major finding is that homeowners often underestimate the likelihood of disasters and delay purchasing insurance until after experiencing a loss. This behavior is influenced by a threshold of concern. If the perceived risk is below this threshold, individuals tend to ignore it, which leads to low insurance uptake in hazard-prone areas. Additionally, many homeowners treat insurance as an investment rather than protection. This results in policy cancellations if no claims are made, which undermines the stability of insurance pools.

On the supply side, insurers face significant difficulties due to the highly correlated nature of catastrophic losses. Events like hurricanes or earthquakes cause simultaneous damage to many properties. This increases the risk of large and concentrated losses that require insurers to hold substantial capital reserves. This capital cost inflates premiums and makes insurance less affordable and less accessible, especially in high-risk areas. The paper also highlights that insurers often withdraw coverage or limit exposure after major disasters. These actions can exacerbate availability and affordability issues for homeowners.

To address these challenges, the study finds that a multi-layered approach involving both private insurers and public sector support is essential. Risk-based premiums that reflect true hazard exposure can encourage homeowners to invest in mitigation measures. At the same time, government programs can provide reinsurance and financial backstops for extreme losses. The findings also emphasize the importance of linking insurance with mitigation incentives. Examples include long-term loans for home improvements and the enforcement of building codes to reduce future losses and improve market sustainability.

Goldburd et al. (2025) used Generalized linear models (GLM) for modeling insurance risks. This study shows that GLMs, are highly effective and practical tools for setting insurance rates. They provide a flexible and reliable way to model insurance risks because they can handle the unique patterns found in insurance

data, such as skewed numbers or varying levels of risk exposure. This flexibility allows experts to build rating models that are easy to understand and match standard industry practices.

Choosing the right mathematical settings is a critical part of the process. The study highlights that using specific probability distributions for different goals, such as using Poisson for how often claims happen or Gamma for how expensive those claims are, is essential for accuracy. Using a log link function is also emphasized as the best choice because it allows different risk factors to be combined naturally to calculate a final premium.

Building a successful model requires balancing technical accuracy with real-world limits. The study finds that it is not just about the math; you also have to consider data quality, which variables to include, and how to avoid making the model too complex. It is important to balance statistical results with business needs, such as following government regulations and making sure the company's computer systems can actually run the model. By addressing these practical challenges and keeping the models transparent, companies can create pricing plans that are reliable and ready for the market.

Beyond GLMs, Bayesian Nonparametric (BNP) regression models such as the Dirichlet Process and Pitman-Yor Process Mixture Models provide a flexible alternative for insurance data that are multimodal, skewed, or heavy-tailed. Estimated using MCMC, these models improve predictive accuracy and naturally group policyholders by risk profile without requiring the number of groups to be specified in advance, and they have been shown to match or outperform classical parametric regressions in both simulations and motor insurance applications. Traditional actuarial models often rely on parametric regression, which assumes a fixed relationship between risk factors and claims. However, insurance data is typically complex because it is often multimodal, highly skewed, and heavy-tailed. These characteristics make rigid traditional models less accurate. To address this, the study proposes using Bayesian Nonparametric (BNP) regression models, specifically the Dirichlet Process Mixture Model and the Pitman-Yor Process Mixture Model. These models provide a flexible framework that allows each data point to have its own unique regression parameters.

The research explains the mathematical implementation of these models for both the number of claims and the monetary amount of those claims. For claim frequency, the data is modeled as a mixture of Poisson regressions. For claim severity, the data is modeled as a mixture of normal regressions on a logarithmic scale. A significant contribution of this work is the development of Markov Chain Monte Carlo (MCMC) sampling methods to handle complex settings where standard mathematical shortcuts do not apply. These models improve prediction accuracy and naturally group policyholders into clusters based on their risk profiles without requiring the researcher to specify the number of groups in advance.

To validate these methods, the study includes extensive simulations and an application to a real-world dataset of French motor insurance claims. The results show that the BNP models consistently match or outperform classical models. In

scenarios where the data comes from a mix of different distributions, the BNP models capture the underlying structure far more effectively than traditional regressions. The study concludes that these flexible Bayesian approaches offer a robust alternative for actuaries who want to set more accurate premiums and better understand individual risk factors.

Zhang et al. (2024) introduce Bayesian Classification and Regression Tree (BCART) models specifically designed to model the frequency of insurance claims. This approach addresses common challenges in insurance data such as exposure variations and datasets with a high number of zero claims. Unlike traditional tree models that use a simple step-by-step search to split data, BCART uses a Bayesian approach with advanced algorithms to explore a wider range of possible tree structures. This method improves both the stability of the model and its overall prediction accuracy. The authors also expand BCART to work with Poisson, Negative Binomial, and Zero-Inflated Poisson distributions, which are the standard mathematical tools used to describe insurance claim patterns.

Through detailed simulations and analysis of real data, the study shows that BCART models perform better than classical tree models. This is especially true when the data contains irrelevant information or when there is a high level of variation in the number of claims. The BCART models are particularly good at ignoring noisy variables and finding the true underlying structure of the data. The research also clarifies which specific model to use based on the data. For example, the Negative Binomial version works best when there is moderate variation, while the Zero-Inflated Poisson version is superior when there are many policyholders with no claims at all.

The analysis of a real insurance dataset confirms that these models are practical for the industry. The study highlights that BCART provides a perfect balance between high accuracy and easy interpretation. Because the models produce transparent and readable trees, they meet the strict regulatory and business requirements necessary for insurance pricing. Overall, the study establishes BCART as a flexible and powerful tool that combines the benefits of Bayesian statistics with tree-based modeling to help insurance companies better classify risk and set fairer premiums.

Fung et al. (2019) uses the Erlang Count Logit-weighted Reduced Mixture of Experts (EC-LRMoE) model, which is a flexible statistical framework designed for modeling multivariate insurance claim frequencies. This model is specifically built to accommodate dependencies among different claim types, such as third-party liability and own-damage claims, which are commonly observed in automobile insurance portfolios. The EC-LRMoE extends traditional mixture models by incorporating expert functions that allow for nonlinear relationships between covariates and response variables, thereby improving the representation of heterogeneous risk profiles.

The authors establish the theoretical foundation of the EC-LRMoE model by demonstrating its "denseness" property. This characteristic ensures that the model class is sufficiently rich to approximate virtually any multivariate frequency dis-

tribution, including complex dependence structures and nonlinear regression patterns. Additionally, the researchers provide a formal proof for the identifiability of the model parameters. This proof guarantees unique statistical inference and avoids common issues such as label switching or multiple interpretation problems often encountered in conventional mixture models.

To facilitate practical implementation, the authors developed an Expectation-Conditional-Maximization (ECM) algorithm for parameter estimation. This algorithm decomposes the complex high-dimensional optimization problem into a series of simpler conditional maximization steps, enabling efficient computation even when dealing with large datasets. This approach is particularly advantageous for handling the integer-valued shape parameters inherent in count data models. The study also employs bootstrapping techniques to obtain standard errors and account for parameter uncertainty in downstream applications.

The model was evaluated using a European automobile insurance dataset characterized by a high degree of heterogeneity, including excess zeros, over-dispersion, and nonlinear covariate effects. The researchers assessed the performance of the model in capturing key features such as under-dispersion and the correlations between different claim types. The analyzed covariates included policyholder demographics, vehicle characteristics, and historical claim indicators. All computations were conducted using customized routines designed to handle the specific requirements of the EC-LRMoE framework.

The authors utilized the fitted model to perform risk classification by identifying latent subpopulations of policyholders with distinct risk profiles, such as safest versus most dangerous drivers. The ability of the model to visualize interactions among risk factors enhances its interpretability for actuarial decision-making. For ratemaking applications, the study incorporated parameter uncertainty through bootstrapped estimates to evaluate the impact of claim correlations. By comparing the results to traditional models, the authors demonstrated that assuming independence between claim types can lead to a significant underestimation of the necessary premiums for certain risk profiles.

The next section introduces the data and method used in this study.

3. Data and Methods

We used a variety of Bayesian Hierarchical models and Frequentist ridge regression models in order to assess the determinants of Homeowners' insurance costs. The model was used for 25,000 + zip codes in the United States. All continuous variables were log-transformed and standardized (Z-score) to allow for direct comparison of effect sizes.

The unit of observation for this study is the ZIP code-year combination, representing aggregated homeowners' insurance and income data for each ZIP code in each year from 2018 to 2022. The final sample size after data cleaning consists of 125,888 ZIP code-year observations. To ensure the integrity of the statistical analysis, the dataset was pre-processed to handle missing, zero, and extreme values. Specifically, observations with zero or negative values for claim severity (2022 in-

stances) and average income (60 instances) were excluded prior to log transformation to avoid undefined mathematical results. Following these exclusions, the variables for premiums, claim severity, and income were log-transformed to normalize their distributions and then standardized to a mean of zero and a standard deviation of one. This rigorous cleaning process addressed potential skewness from extreme values and ensured that the subsequent Bayesian modeling and regression analyses were based on a consistent and reliable sample of the homeowners' insurance market in the United States of America (USA).

3.1. Data Sources and Key Variable Definitions

We integrated two primary datasets: the U.S. Treasury's property and Casualty Market Intelligence (PCMI) data and the American Community Survey (ACS) 5-year estimates[1]. The resulting dataset contains 127.965 observations, ZIP codes were standardized to 5-digit strings to ensure a 100% match. Income data were averaged over the 2018-2022 period to provide a stable measure based on the median household income.

The following key variables are used in the modeling and interpretation:

Premiums Per Policy represents the average insurance premium charged per active policy within each ZIP code and year. It is calculated as the total premiums collected divided by the number of active policies, reflecting the average cost borne by policyholders in that area.

Loss Ratio is defined as the ratio of total claims paid to total premiums earned within a ZIP code-year. This metric indicates the proportion of premium revenue paid out in claims, serving as a measure of underwriting profitability or risk exposure.

Nonrenewal Rate measures the proportion of policies that were not renewed at the end of their term for reasons other than nonpayment. It captures voluntary or insurer-initiated discontinuations excluding cancellations due to missed payments.

Other than Nonpayment Cancellation Rate quantifies the proportion of policies canceled during the policy term for reasons other than nonpayment, such as underwriting decisions, fraud, or changes in risk profile.

Policy Decile Grouping categorizes ZIP code-year observations into ten groups (deciles) based on the distribution of average premiums per policy. The observations are ranked from lowest to highest premiums and divided into ten equal-sized groups, each representing approximately 10% of the sample. This grouping facilitates comparative analysis across different risk or cost strata.

These variables form the basis for understanding geographic and socioeconomic variation in homeowners insurance premiums and are incorporated as predictors or outcomes in the Bayesian hierarchical modeling framework described below.

[1]The exact data source is currently not available. Interested readers may consult the following: U.S. Department of the Treasury, Federal Insurance Office. (2023). Homeowners Insurance: Availability and Affordability Data Call. Washington, DC: U.S. Department of the Treasury.

3.2. Methods

Our primary outcome variable is the log-transformed premium per policy for zip code i in year t. We model this outcome using a Bayesian hierarchical regression framework as follows:

$$\log\left(y_{it}\right)=\alpha+\beta\log\left(X_{it}\right)+\sum_{k}\gamma_{k}Z_{kit}+v_{t}+\varepsilon_{it} \tag{1}$$

where, y_{it} is the premium per policy, X_{it} is the average household income, Z_{kit} are additional covariates, γ_{k} are their coefficients, and ε_{it} is the residual error.

3.2.1. Bayesian Priors and Regularization

To improve model parsimony and prevent overfitting, we apply Bayesian Lasso priors on regression coefficients (γ_{k}), which performs shrinkage and variable selection within the hierarchical framework. Additionally, Dirichlet and Pitman-Yor process priors are used to model latent clustering that allow for flexible, data-driven grouping structures among zip codes. These nonparametric priors enable the model to capture complex heterogeneity without imposing restrictive assumptions on the distribution of random effects.

We consider four model specifications in this study. They are: (a) Bayesian model with Lasso priors; (b) Dirichlet process mixture of random intercepts model; (c) Normal Random intercepts model (2-level); and (d) Pitman-Yor process mixture of random intercepts model.

A. Bayesian Model with Lasso Priors

Likelihood:

$$y_{i}\sim N\left(\mu_{i},\ \sigma^{2}\right)$$

$$\mu_{i}=x_{i}^{T}\beta$$

Priors:

$$\beta\sim\prod_{j=1}^{p}Laplace\left(0,\ \lambda\right)$$

$$\sigma^{2}\sim Inverse-Gamma\left(a,\ b\right)$$

Scale Mixture Representation:

$$\beta_{j}\left|\tau_{j}^{2}\right.,\ \lambda\sim N\left(0,\ \tau_{j}^{2}\right) \tag{2}$$

$$\tau_{j}^{2}\left|\lambda\right.\sim Exp\left(\left(\frac{\lambda^{2}}{2}\right)\right) \tag{3}$$

B. Dirichlet Process Mixture of Random Intercepts Model

Likelihood:

$$y_{ij}\sim N\left(\mu_{ij},\ \sigma^{2}\right)$$

$$\mu_{ij}=x_{ij}^{T}\beta+b_{i}$$

Priors:

$$\beta\sim N\left(0,\ \sum\nolimits_{\beta}\right)$$

$$\sigma^2 \sim Inverse-Gamma\left(a_\sigma,\ b_\sigma\right)$$

$$\sigma_b^2 \sim Inverse-Gamma\left(a_b,\ b_b\right)$$

$$\alpha \sim Gamma\left(a_\alpha,\ b_\alpha\right)$$

Random Effects Distribution:

$$b_i \sim G \tag{4}$$

$$G \sim DP\left(\alpha,\ G_0\right) \tag{5}$$

C. Normal Random Effects Model (*NREM*)

Likelihood:

The observed data $Y_i(h)$ for group h follows a normal distribution:

$$Y_i(h)\left|X_i(h) \sim N\left(X_i(h)\beta_{rh},\ \sigma^2\right)\right.$$

The regression coefficients for group h are modeled as:

$$\beta_{rh} = \beta + u_h$$

Priors:

$$\beta_0\left|\sigma^2 \sim N\left(0,\ \sigma^2/v_0\right),\ v_0 \to \infty\right. \text{ (non-informative prior)}$$

$$\beta_k\left|\sigma^2 \sim N\left(0,\ \sigma^2 v_k\right),\ k = 1,\cdots,p\right.$$

$$\mu_h\left|T \sim N\left(0,\ T\right),\ h = 1,\cdots,N_h\right.$$

$$\sigma^2 \sim Inverse-Gamma\left(\frac{a_0}{2},\ \frac{a_0}{2}\right)$$

$$T \sim Inverse-Wishart\left(p+3,\ S_0\left|p+1\right.\right)$$

Posterior:

The posterior distribution is proportional to the product of the likelihood and priors:

$$\begin{aligned}&\Psi\left(\beta,\ u,\ \sigma^2,\ T\left|D_n\right.\right) \propto L\left(D_n;\ \beta,\ u,\ \sigma^2,\ T\right) \times \Psi\left(\beta_0\left|\sigma^2\right.\right)\\ &\times\prod_{k=1}^{p}\Psi\left(\beta_k\left|\sigma^2\right.\right)\times\prod_{h=1}^{N_h}\Psi\left(u_h\left|T\right.\right)\times\Psi\left(\sigma^2\right)\times\Psi\left(T\right)\end{aligned} \tag{6}$$

D. Pitman-Yor Process Mixture Model (PYHLM)

Likelihood:

The data likelihood is modeled as an infinite mixture of normal regressions:

$$f\left(y_h\left|X_h\right.\right) = \sum_{j=1}^{\infty}\prod_{i(h)=1}^{n_h} N\left(y_{i(h)}\left|x_{i(h)}\beta_j,\ \sigma^2\right.\right) * w_j$$

Priors:

Stick-breaking weights:

$$w_j = v_j\prod_{l=1}^{j-1}\left(1-v_l\right)$$

With

$$v_j \sim Beta\left(1-d,\ \theta+jd\right)$$

Regression parameters and variance:

$$\beta_j \sim N(\mu,\ T)$$

$$\sigma^2 \sim Inverse-Gamma\left(\frac{a_0}{2},\ \frac{a_0}{2}\right)$$

$$(\mu,\ T) \sim Normal-inverse-wishart(0,\ r_0,\ S_0,\ p+1)$$

where, $d \in [0,1)$ is the discount parameter and $\theta > -d$ is the concentration parameter.

Posterior:

The posterior distribution again is proportional to the product of the likelihood and the priors[2].

Ridge Regression Specifications

We included L2-regularized linear regression (Ridge) as one of the candidate predictive models. Ridge regression was applied to the log-transformed premiums (log_premium) using the same set of predictors employed throughout this study, including standardized log-income, standardized log-severity, standardized frequency, policy decile, and other relevant covariates. Prior to model fitting, all predictors were standardized to have a mean of zero and a standard deviation of one. This standardization ensures that coefficient magnitudes represent the effect of a one standard deviation change in the predictor and allows the L2 penalty to be applied uniformly across all predictors. The Ridge penalty parameter (lambda) was chosen via 10-fold cross-validation by minimizing the out-of-sample root mean squared error (RMSE). Final Ridge coefficients are reported on the standardized predictor scale; where appropriate, we also present back-transformed effects (exp(β)) to approximate percent changes on the original premium scale.

When comparing our models, we looked at more than just the raw predictive scores like RMSE and LOO/WAIC. While the Pitman-Yor (PY) model performed similarly to Ridge regression, we ultimately chose to stick with Ridge for the final analysis. The main reason was consistency: Ridge gave us much more stable coefficient estimates during cross-validation, whereas the PY model was more sensitive to how it was tuned, leading to 'jumpier' results across different data folds. Because Ridge is simpler to reproduce and its results are easier to interpret and communicate to others, it offered a better balance of reliability and performance[3].

3.2.2. Model Estimation and Inference

The model is estimated using Markov Chain Monte Carlo (MCMC) methods, producing full posterior distributions for all parameters. This approach facilitates

[2]We do not specify this equation as it is very clumsy. We can provide a derivation of this equation from interested author upon reasonable request.

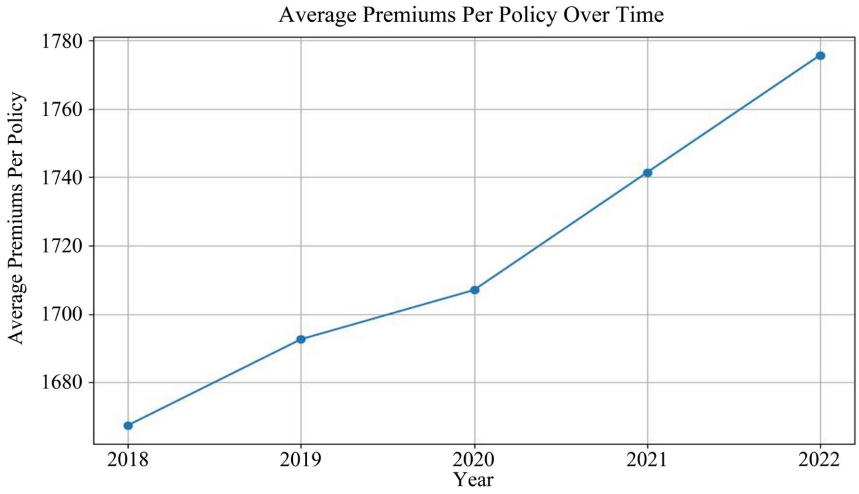
[3]To ensure direct comparability, all models use a consistent log-transformed outcome and standardized predictors (mean 0, SD 1). Consequently, each coefficient represents the effect of a one-standard-deviation increase in the predictor on log premium. For practical interpretation, we also report exponentiated coefficients (eβeβ) to show the approximate percentage change in premium per standard deviation. Any models originally estimated without standardization were rescaled prior to comparison to maintain a uniform scale across all reported result.

uncertainty quantification in premium predictions and parameter estimates, supporting robust inference and regulatory applications (Karabatsos, 2017).

4. Main Results

This section is organized into 3 parts: (a) Descriptive Statistics; (b) Model Evaluation; and (c) Main Results of Bayesian Regressions.

4.1. Descriptive Statistics



Source: Computed by the author using US Treasury Data.

Figure 1. Average premiums per policy over time.

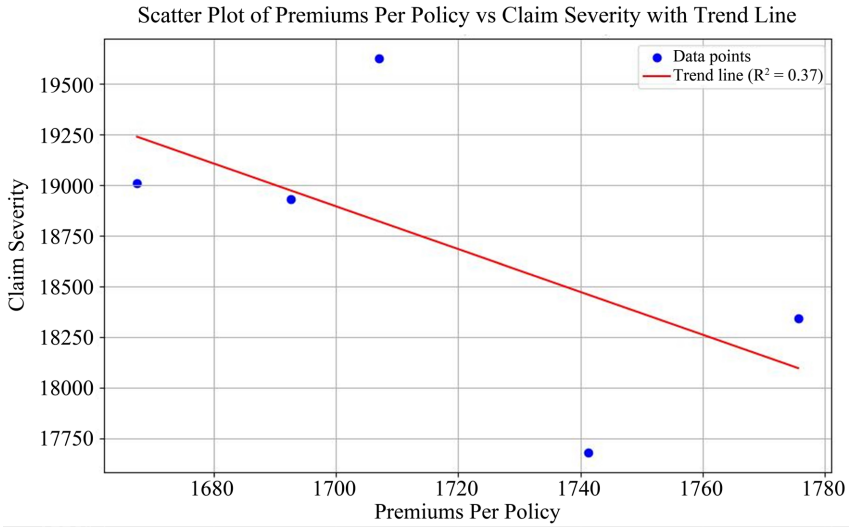
Figure 1 shows the trend of the average premiums per policy over time. The X-axis represents the years, while the vertical axis shows the average premium amount in dollars, ranging from roughly $1660 to $1780.

There is a clear upward trend throughout the five-year period. Every year, the average premium was higher than the year before. While the costs rose every year, the increase became more aggressive in the later years. Between 2020 and 2022, the line becomes steeper, indicating that premiums were rising at a faster rate during that time compared to the period between 2019 and 2020.

Figure 2 illustrates a negative correlation between average annual premiums and claim severity from 2018 to 2022, as evidenced by the downward-sloping red regression line. By aggregating the normalized_insurance_data.csv dataset into yearly means, the analysis reveals that as the average premium per policy increased from approximately $1675 to $1765 over the five-year period, the average claim severity concurrently declined from nearly $19,500 to just above $18,500. The linear model fits the data with an R-squared value of 0.37 and a negative slope of −11.52, suggesting that for every one dollar increase in the average premium, claim severity decreased by roughly $11.52. This inverse relationship indicates that while policy costs for homeowners trended upward during this window, the average cost of individual claims processed by insurers saw a consistent reduction.

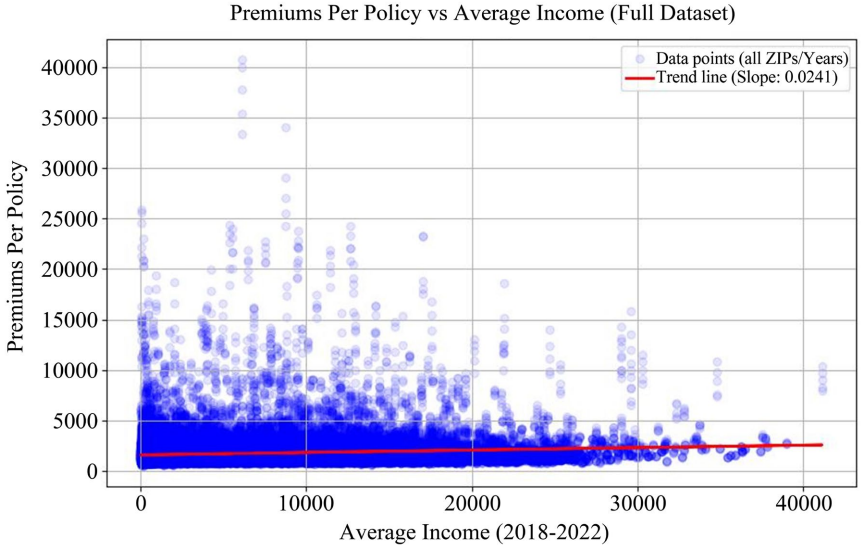
Figure 3 looks at the relationship between how much people earn (Average In-

come)[4] and how much they pay for insurance (Premiums Per Policy) across various ZIP codes and years. The horizontal axis represents the "Average Income" from 2018 to 2022, ranging from $0 to over $40,000. The vertical axis represents "Premiums Per Policy" on a scale that goes from $0 to $40,000, but the data points are heavily concentrated at the bottom. While the axis itself reaches $40,000 to accommodate those extreme "outlier" points at the top, the vast majority of the blue dots (the thickest part of the "cloud") are actually clustered much lower, specifically between $0 and $5000.



Source: Computed by the author using US Treasury Data.

Figure 2. Scatter plot of premiums per policy vs claim severity.



Source: Computed by the author using US Treasury Data.

Figure 3. Premiums per policy vs average income (2018-2022).

[4]Average income at the ZIP code level (median household income from ACS data, averaged across 2018-2022).

This graph shows that as average income increases, insurance premiums tend to rise only very slightly, with the red trend line indicating that every extra $1000 in income only adds about $24 to the premium cost. The most notable feature is the massive cluster of data points at the lower end of the income scale, where most people pay relatively low premiums under $5000, yet there are many extreme "outliers" paying as much as $25,000 to $40,000. As you move toward higher income levels on the right, the data becomes much thinner and these dramatic price spikes disappear, suggesting that while higher earners pay slightly more on average, the most extreme insurance costs are actually being hit by people in lower-income areas.

4.2. Model Evaluation

Our final model choice was guided by the following considerations: (a) Best predictive accuracy with all variables; (b) Modeling group-level effects explicitly; (c) Modeling latent clusters or heterogeneity; and (d) Balance between interpretability and complexity. The results are summarized in Table 1.

Table 1. Comparison of models.

Model Name	Type	Key Features	Computation Time	R2 (Explained Variance)	Notes on Clustering/Grouping
LASSO Prior Model	Linear regression with L1 penalty	Variable selection, coefficient shrinkage	Not reported	Not explicitly reported	No grouping
Dirichlet Process Mixture (DPM)	Bayesian nonparametric mixture	Random intercept clustering	High (MCMC)	~4.8%	1 cluster found (no meaningful clustering)
Pitman-Yor (PY) Process Mixture	Bayesian nonparametric mixture	Flexible clustering prior	High (MCMC)	~7%	1 cluster found (no meaningful clustering)
Ridge Regression (MML)	Penalized linear regression (L2)	Shrinks coefficients, no variable selection	Very fast (−0.23 s)	~7%	No grouping
Normal Random Intercepts Model (NREM)	2-level hierarchical linear model	Explicit random intercepts for grouping	Moderate (MCMC)	Not explicitly reported	Grouping modeled, small variance explained

Source: Compiled by the author.

4.2.1. Predictive Performance and Model Fit

The analysis of insurance premium determinants reveals that the Ridge Regression and Pitman-Yor (PY) Process models achieved the highest predictive accuracy, each explaining approximately 7% of the total variance in premium costs. While this figure may appear modest, it represents a statistically significant improvement over the Dirichlet Process Mixture (DPM) model, which accounted for only 4.8% of the variance. The Bayesian model with Lasso priors, while effective for identifying primary drivers such as geography and income, prioritized model simplicity over exhaustive predictive power. These results suggest that

while the identified risk factors are influential, a substantial portion of premium volatility is likely driven by external market conditions or unobserved policyholder characteristics not captured in the current dataset.

4.2.2. Dynamics of Grouping and Latent Clustering

A primary goal of this study was to see if policyholders naturally fall into separate risk groups based on their decile rankings. Interestingly, both the DPM and PY models identified only one single, unified cluster, even though these models are specifically designed to find hidden patterns. This shows that the "Policy Decile Grouping" is not a set of separate and isolated risk pools. Instead, it represents a continuous range of risk. Additionally, the Normal Random Intercepts Model (NREM) confirmed that these groupings explained very little of the overall variation. In simple terms, the data suggests that insurance risk is more consistent across these groups than we originally thought. This means that using complex models to find multiple clusters does not actually help us understand premium behavior any better in this case.

Thus, we select the Bayesian regression model with Lasso priors and the Ridge regression model in our final analysis.

4.3. Main Results of the Bayesian Models

Table 2. Comparison of covariate effects: Bayesian model with lasso priors vs ridge regression.

Covariate	LASSO beta	Ridge beta	Notes
ZIP Code	−0.001	−23.66	LASSO shrinks this to nearly zero; Ridge shows a larger standardized effect
Year	28.99	41.59	Both positive; Ridge slightly higher
Claim Severity	0.005	189.80	LASSO very small; Ridge large (possibly due to standardization differences)
Loss Ratio	−18.11	−125.24	Both negative; Ridge more extreme
Nonrenewal Rate	−621.26	−4.66	LASSO much more negative; Ridge near zero (possibly due to standardization)
Nonpayment Cancellation Rate	−3413.15	−51.21	LASSO much more negative
Other than Nonpayment Cancel	−6960.24	−80.46	LASSO much more negative
Income_Avg_2018_2022	0.022	120.65	LASSO small; Ridge large

Source: Computed by the author. Note: Although predictors were standardized for Ridge cross-validation, the coefficients reported in this table are back-transformed to the original premium scale in dollars for ease of interpretation. Values across the two models are therefore on comparable scales but are not standardized regression coefficients.

The comparison between the Bayesian model with Lasso priors and Ridge regression models reveals how different statistical approaches interpret the factors driving insurance premiums (Table 2). The Bayesian model with Lasso priors acts as a strict filter, aggressively shrinking less important variables like ZIP code and

average income to nearly zero[5]. This suggests that, from a simplified perspective, these factors may not be the primary drivers of cost. In contrast, the Ridge model retains all variables but scales them differently, showing much larger standardized effects for factors like claim severity and income. This difference highlights that while some factors may seem insignificant in a simple model, they can still play a substantial role when the model is allowed to consider the full complexity of the data.

The most striking finding across both models is the massive impact of policy cancellations and non-renewals. Specifically, the "Other than Nonpayment" and "Nonpayment" cancellation rates showed the most extreme negative values, particularly in the Bayesian model with Lasso priors. This indicates a strong negative association between cancellation rates and average premiums. The direction of this relationship should be interpreted with care. It is consistent with a selection story in which lower-priced policies are more likely to lapse or be cancelled, rather than with cancellations directly causing premiums to fall. Disentangling these explanations would require policy-level data and is beyond the scope of the present analysis. Additionally, both models agreed that "Year" has a consistent positive effect, confirming that premiums have been steadily rising over time regardless of which mathematical approach is used to analyze the data.

Zip codes i.e. explaining geographic variation and average income are not the main drivers of variation of premiums per policy. While where you live and how much you earn do influence premiums, the data shows that policy cancellations, non-renewals, and the general passage of time are much more powerful at explaining why one person's premium is different from another's. This includes factors like inflation and market trends. ZIP code and income are secondary factors that provide extra detail, but they are not the main reasons for the price changes we are seeing.

5. Conclusion and Policy Implications

The findings of this study suggest that the primary drivers of homeowners' insurance premiums are not necessarily the factors most commonly discussed in public discourse, such as specific ZIP codes or household income levels. While these geographic and socioeconomic indicators do play a role, they are secondary to more direct operational and market-driven metrics. The data clearly shows that policy cancellations, non-renewals, and the general passage of time are far more powerful at explaining premium variations. This indicates that the market is being shaped more by broad economic trends like inflation and internal policy stability than by local neighborhood characteristics alone.

From a policy perspective, these results suggest that regulators and industry leaders should shift their focus toward the stability of policy renewals and the un-

[5]Preliminary work using a BYM2 spatial specification suggests that income effects may become statistically significant once spatial correlation is modeled explicitly, and that spatial structure itself accounts for a modest share of premium variance. A full treatment of this specification is left to future work.

derlying causes of high cancellation rates. Since high rates of non-payment and other cancellations are so strongly linked to premium fluctuations, programs that help homeowners maintain continuous coverage could potentially lead to more stable pricing across the market. Furthermore, because the study found that insurance risk is relatively uniform across different decile groups, there is a strong argument for simplifying the complex clustering models often used in ratemaking. A more streamlined approach could improve transparency for consumers and make it easier for regulators to ensure that pricing remains fair and predictable.

Finally, the consistent upward trend in premiums over the five-year period highlights the need for long-term strategies to address rising costs. Since the "Year" variable was a significant positive driver in every model, it is clear that external pressures like rising construction costs and shifting climate risks are creating a baseline of inflation that individual policyholder characteristics cannot overcome. Policymakers might consider incentivizing home resilience improvements or exploring state-backed reinsurance options to help buffer the market against these broad, time-driven increases. By focusing on these systemic drivers rather than just individual risk factors, the insurance industry can move toward a more sustainable and equitable future for all homeowners.

Acknowledgements

We thank an anonymous referee for excellent comments on the draft of the manuscript.

Disclaimer

Any remaining errors or omissions remain the sole responsibility of the author.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

Fung, T. C., Badescu, A. L., & Lin, X. S. (2019). A Class of Mixture of Experts Models for General Insurance: Theoretical Developments. *Insurance: Mathematics and Economics, 89*, 111-127. https://doi.org/10.1016/j.insmatheco.2019.09.007

Goldburd, M., Khare, A., Tevet, D., & Guller, D. (2025). *Generalized Linear Models for Insurance Rating* (2nd ed.). Casualty Actuarial Society.

Karabatsos, G. (2017). A Menu-Driven Software Package of Bayesian Nonparametric (and Parametric) Mixed Models for Regression Analysis and Density Estimation. *Behavior Research Methods, 49*, 335-362. https://doi.org/10.3758/s13428-016-0711-7

Kunreuther, H., & Pauly, M. (2006). *Insuring Against Catastrophes*. Paper for Financial Risk Management in Practice: The Known, The Unknown and The Unknowable.

Zhang, Y., Ji, L., Aivaliotis, G., & Taylor, C. (2024). Bayesian CART Models for Insurance Claims Frequency. *Insurance: Mathematics and Economics, 114*, 108-131. https://doi.org/10.1016/j.insmatheco.2023.11.005