Exploring Factors Influencing Insurance Pricing: A Comparative Analysis of Fixed and Random Effects

Anton Yang

2024-12-04

Introduction

Insurance is a financial product that provides protection against financial loss or risk. It is based on the principle of risk pooling, where individuals or entities pay regular premiums to an insurance company in exchange for coverage in case of unexpected events. These events could be accidents, illnesses, natural disasters, or property damage, depending on the type of insurance. Insurance distributes risk across a large number of policyholders, insurance companies are able to provide compensation to those who experience covered events while maintaining financial stability.

Specifically, health insurance focuses on providing financial assistance for medical expenses incurred by policyholders due to illness or injury. Health insurance plans cover a range of services, including hospital stays, doctor visits, prescription medications, preventive care, and surgeries. The premiums for health insurance are often determined by various factors such as age, gender, lifestyle habits (like smoking, jobs, hobbies, etc.), medical history, and the overall health of the individual. To assess the risk of insuring individuals use actuarial models and statistical techniques.

For this project, we'll use Medical Insurance Cost dataset from Kaggle [1]. This dataset contains about 2700 observations, and 7 variables (details shown in Table 1). The rarget variable, Charges, represents the total medical expenses for each policyholder, making it a valuable tool for modeling and prediction. We are interested in this data because we want to analyze the factors that influences the medical insurance charges, and this is crucial for actuarial work.

This project aims to investigate whether incorporating random effects can enhance the predictive accuracy of models built on this dataset. By exploring group-level variability, random effects can capture unobserved heterogeneity within groups, such as regions or individual characteristics. Additionally, the project goals is to perform diagnostic on the models and compare it with ordinary linear model and generalized linear model.

The findings of this project will further our understanding the factors driving the medical expenses and for improving insurance pricing strategies. If random effects demonstrate an

improvement in model performance, they may provide new insights into the variability of medical charges across different groups. In additional, identifying the appropriate fixed and random effects can enhance the company's ability to predict expenses more accurately.

Explanatory Variables

Variables	Descriptions
age	Age of the person
sex	Gender of the person
bmi	Body Mass Index
children	Number of children
smoker	Smoker or Non-smoker
region	Region like northeast, northwest, southeast, southwest
charges	Yearly insurance price

Table 1: Explanatory Variables Descriptions

Data

First, we'll analyze the characteristics of the data to identify any significant differences in insurance prices based on factors like gender, smoking status, and the number of children. Next, we will examine the correlation between insurance price and variables such as age and BMI. After exploring the dataset, we'll split it into training and test sets using the 80-20 rule, and begin developing the model.

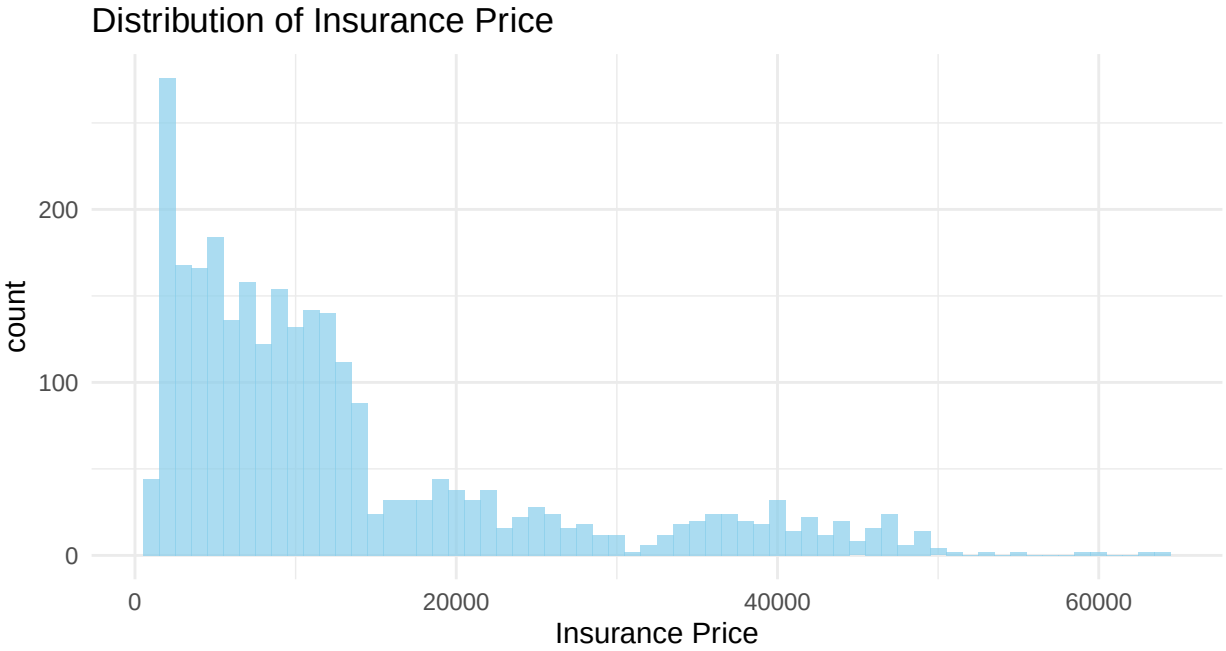

Figure 1: Distribution of Insurance Price

First, we examine the distribution of insurance prices. As shown in Figure 1, the distribution is right-skewed, which aligns with expectations. Most individuals are not charged high premiums unless they fall into the rated lives category. The majority of policyholders belong to either the preferred lives or standard lives groups, resulting in a concentration of lower insurance prices.

Based on Figure 2, we observe that the number of children does not have a significant impact on insurance price. There is no strong evidence to suggest that the number of children influences the annual insurance premium. Similarly, region and sex appear to have little effect on the price. However, smoking status is a notable factor—smokers tend to have higher insurance premiums than non-smokers. This is clearly reflected in the box plot, where smokers consistently show higher insurance prices. Thus, we can conclude that certain factors, such as smoking status, do influence insurance prices. Now we want to check the distribution of each categorical variables.

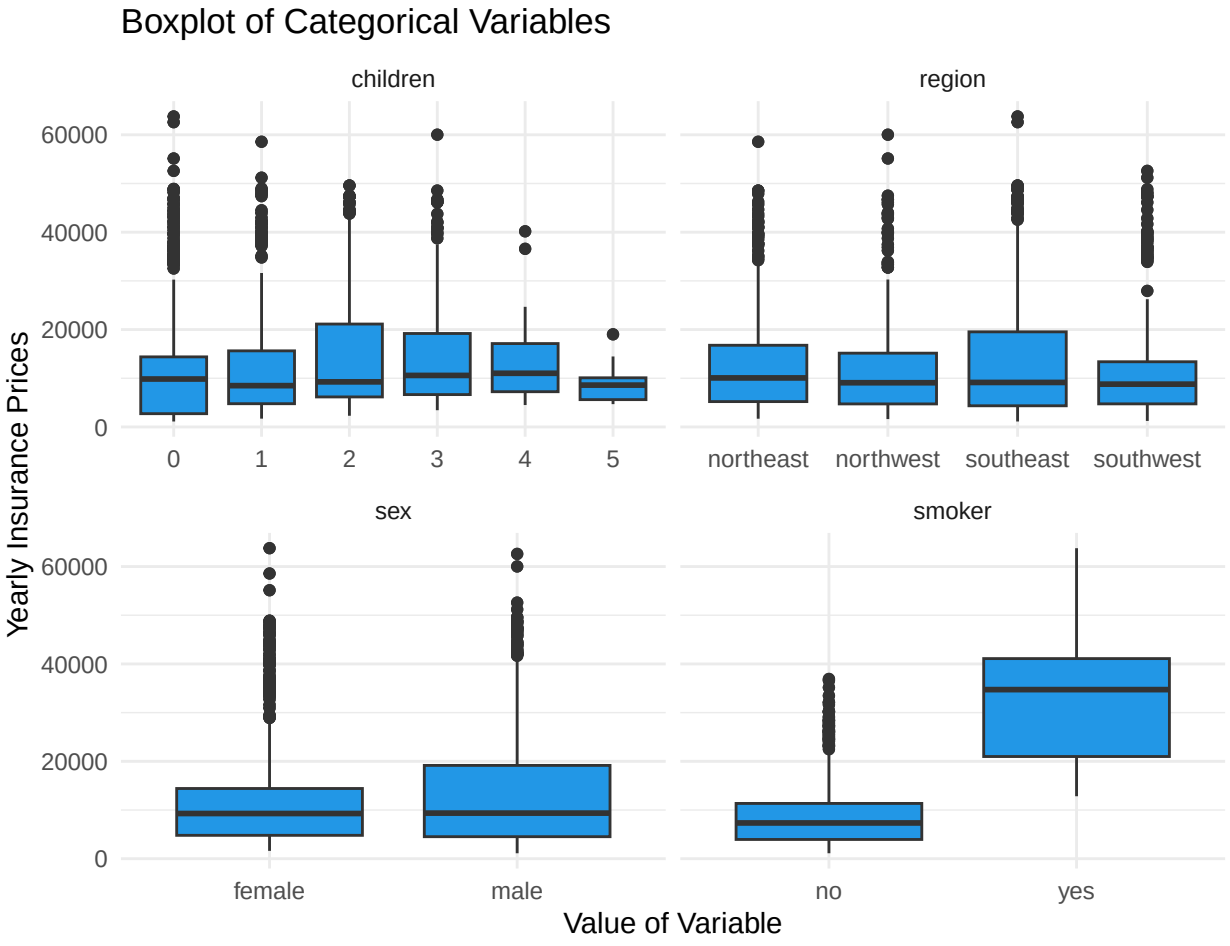

Figure 2: Box Plots of Categorical Variables

According to Figure 3, the distribution of insurance prices between smokers and non-smokers is distinctly different. Non-smokers exhibit a lower insurance price, with a right-skewed distribution. In contrast, smokers show a bimodal distribution, with peaks around 20,000

dollars and 40,000 dollars. This indicates that the two groups have fundamentally different insurance price structures. Given these differences, it may be beneficial to consider incorporating smoker status as a random effect, as the variation in distributions suggests that separate modeling approaches for each group could improve prediction accuracy.

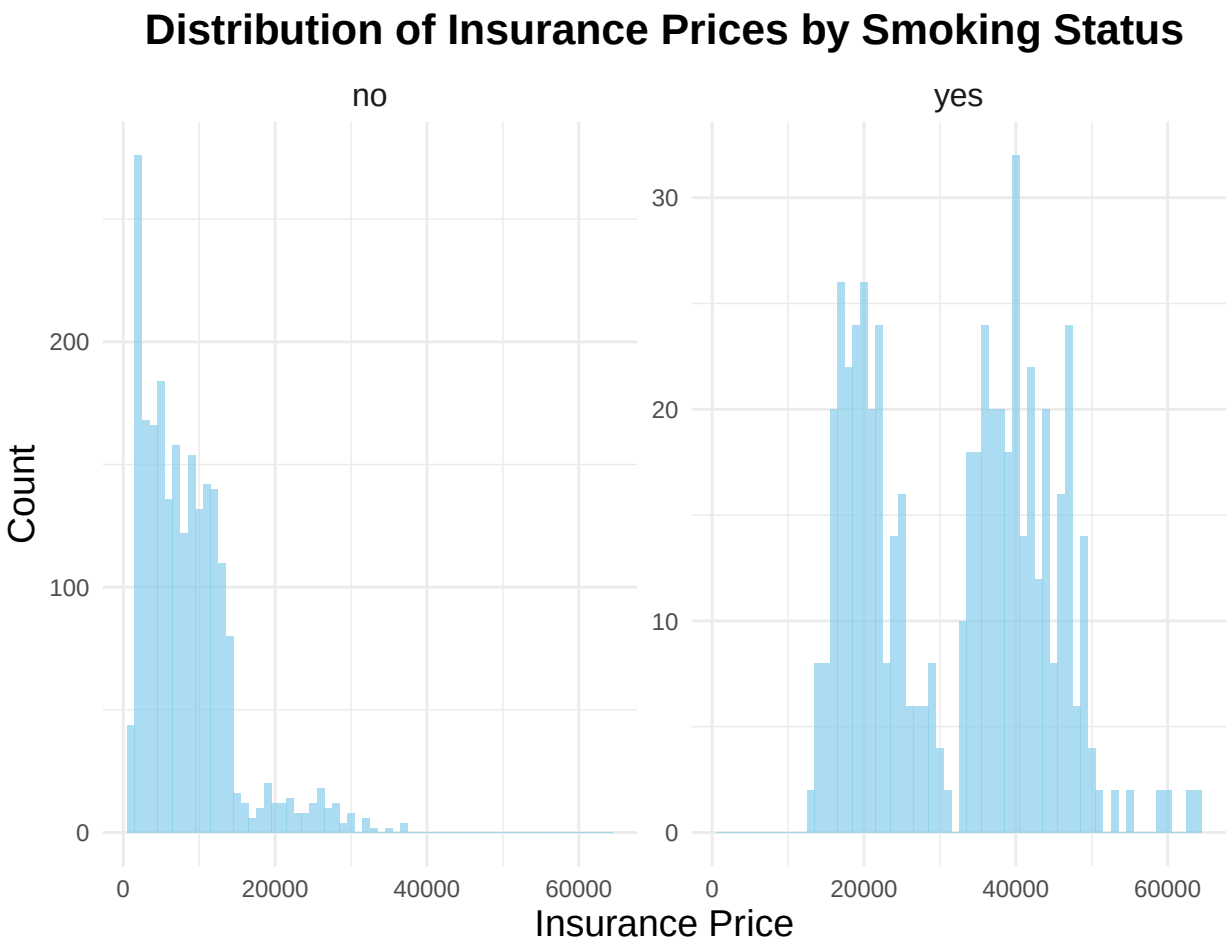

Figure 3: Distribution of Insurance Price for Smokers and Non-Smokers

We also aim to explore the distribution of insurance prices across different regions. We hypothesize that each region may exhibit varying insurance costs, with the Northeast, in particular, typically having higher premiums compared to other regions. This regional variation could provide valuable insights into how location influences insurance pricing.

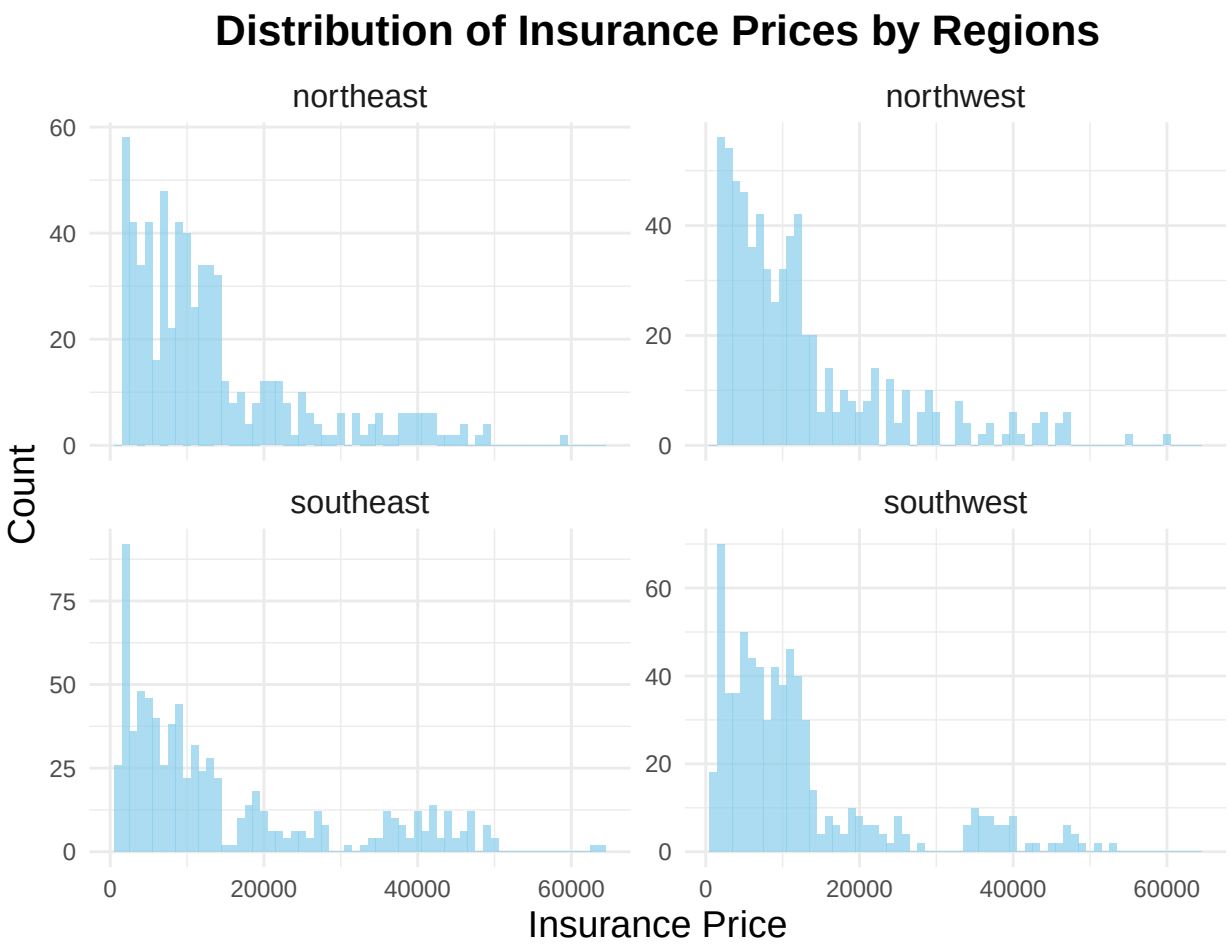

Figure 4: Distribution of Insurance Price for each Region

Based on Figure 4, we observe that the distribution of insurance prices across all regions is right-skewed. This suggests that the region variable is likely best treated as a fixed effect, as the patterns do not indicate the need for a random effect to capture significant regional variability.

We believe that the number of children might show a different distribution, as intuitively, more children could lead to higher insurance prices and greater variability in costs due to increased risk. However, as shown in Figure 5, the distribution for all categories of children is right-skewed, similar to the pattern observed with region. Therefore, we are likely to treat the number of children as a fixed effect, as the distribution does not suggest a need for a random effect.

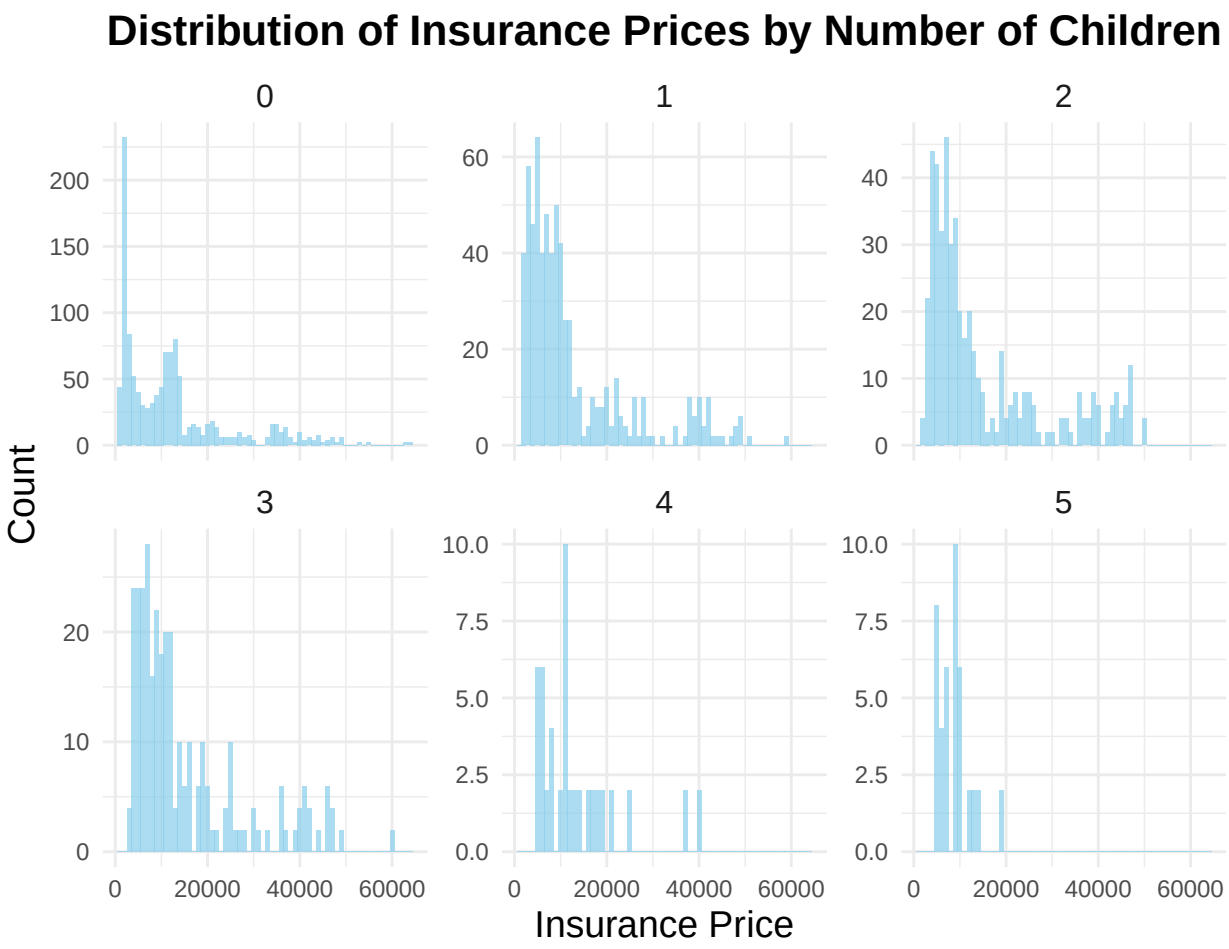

Figure 5: Distribution of Insurance Price by Number of Children

Finally, we examine the distribution of gender. As shown in Figure 6, both male and female distributions are right-skewed. This pattern further suggests that a random effect is unnecessary, and gender can be treated as a fixed effect in our model.

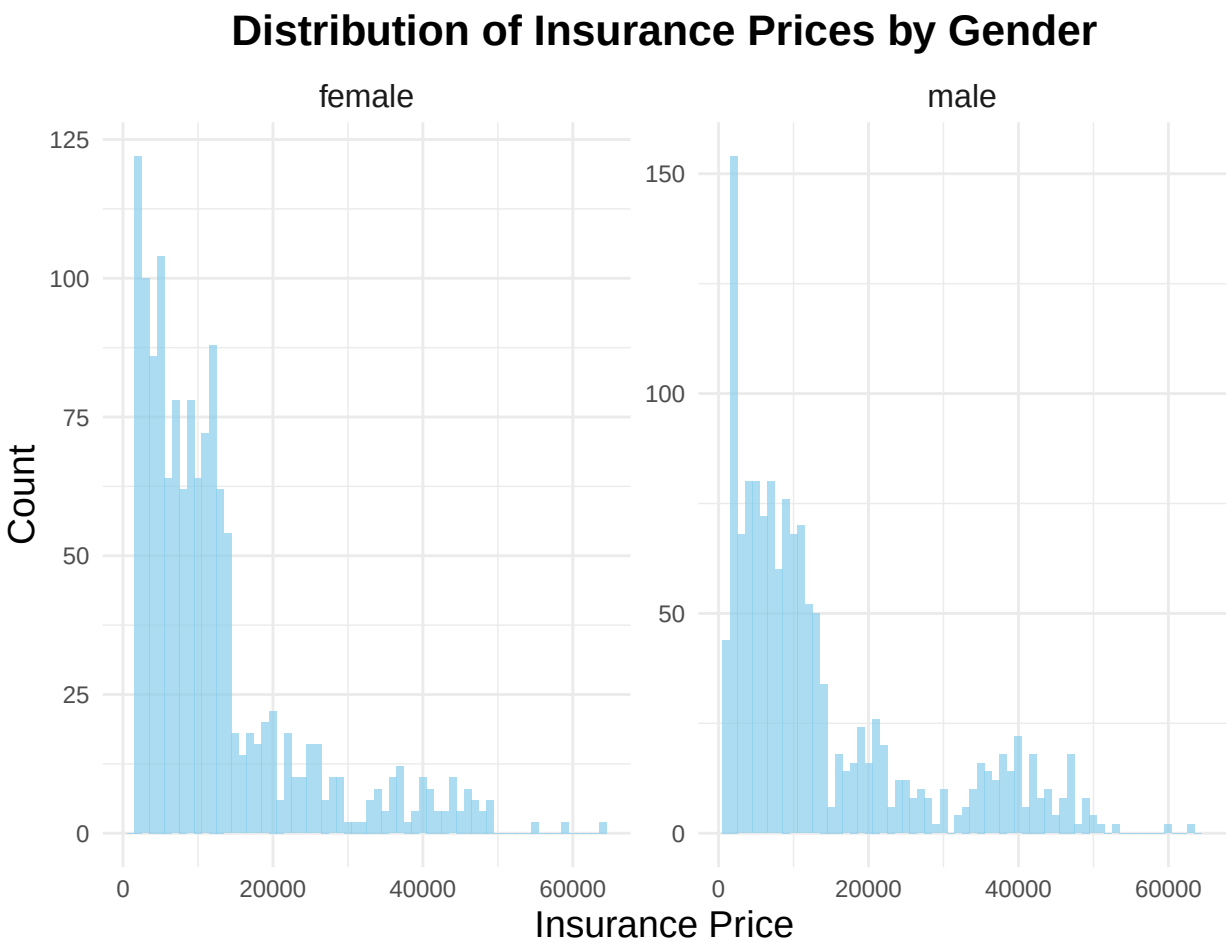

Figure 6: Distribution of Insurance Price by Gender

Next, we will explore the relationship between numeric variables and the insurance price. We anticipate a positive correlation between age and insurance price, as older individuals are considered higher risk due to an increased mortality rate. Additionally, we expect to observe a positive relationship between BMI and insurance price, since individuals with higher BMI may face greater health risks, leading to higher medical expenses.

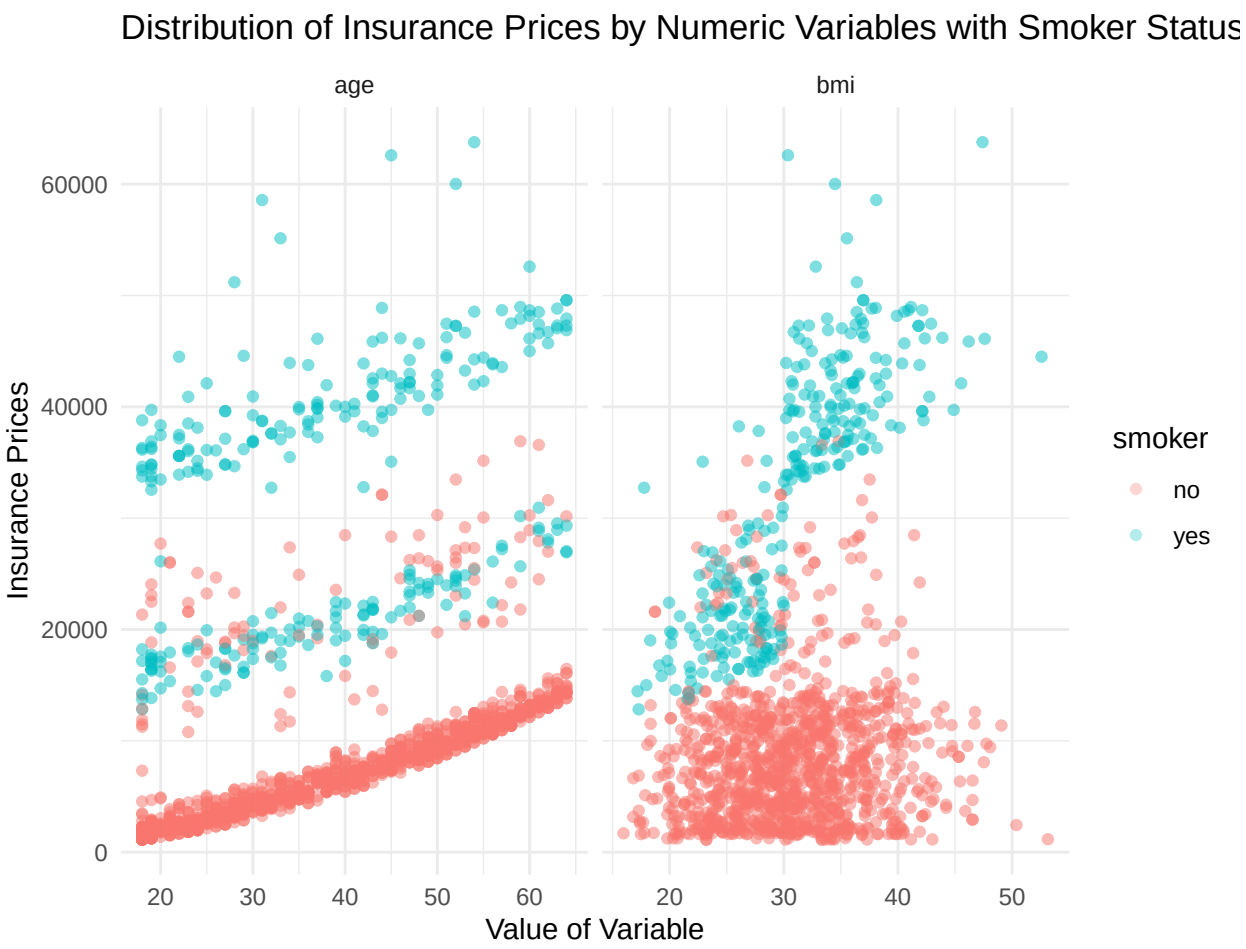

Figure 7: Scatter plot of Numeric Variables with Smoker Status

According to Figure 7, age shows a clear positive correlation with insurance price. The scatter plot reveals three distinct lines: the lowest line represents non-smokers, the middle line shows a mixture of smokers and non-smokers, and the highest line consists entirely of smokers. This provides stronger evidence for treating smoking status as a random effect. On the other hand, while BMI does not show a clear positive correlation with insurance prices, we do observe extreme outlier insurance prices as BMI increases.

Generalized Linear Model

We will analyze three models: the ordinary linear model, the generalized linear model (Gamma), and the random effects model. Diagnostic tests will be performed on each model to identify the best-performing one. Finally, we will evaluate the predictive accuracy of each model to determine which provides the most reliable predictions.

First, we want to build a full linear model. We transformed the target variable charges to reduce the effect of high insurance price. To improve the model, we performed stepwise

selection (both directions since there are only 7 variables), and the variable sex is eliminated. According to Table 2, we can see that all variables are highly significant (lower than p-value of 0.05). We have a linear model:

$$\hat{y} = \beta_0 + \beta_1 x_{age} + \beta_2 x_{bmi} + \beta_3 x_{children} + \beta_4 x_{smoker} + \beta_5 x_{northwest} + \beta_6 x_{southeast} + \beta_7 x_{southwest} + \epsilon$$

LM Coefficients (AIC = 44922.99)

Variable	Coefficient	PValue
(Intercept)	-12380.6798	1.397487e-53
age	258.1784	1.029992e-147
bmi	334.9547	2.763110e-49
children	525.4390	8.613790e-07
smokeryes	23813.0588	0.000000e+00
regionnorthwest	-357.6551	3.388672e-01
regionsoutheast	-1212.8820	1.143209e-03
regionsouthwest	-1120.6436	2.703946e-03

Table 2: Ordinary Linear Model Summary

Based on Figure 8, the residuals display three distinct clusters, indicating that the linear model may struggle to accurately capture the dynamics of extremely low and high insurance prices. Furthermore, the Q-Q plot reveals significant deviations from normality, with many points falling outside the theoretical quantile line. These observations suggest that the ordinary linear model may not be the most effective for predicting insurance prices. To address this, we will explore a generalized linear model with a Gamma distribution, which is better suited for handling right-skewed data.

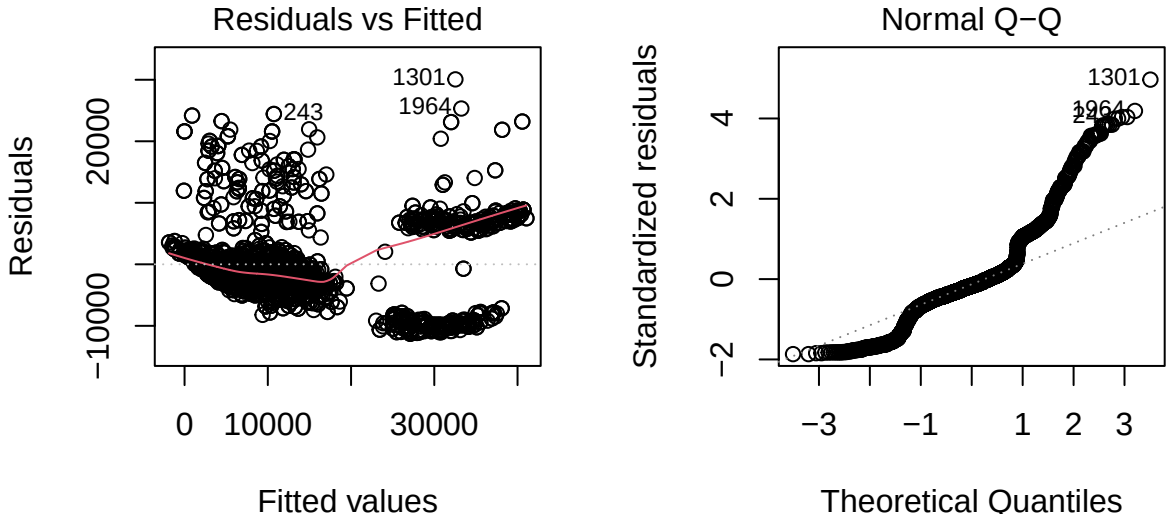

Figure 8: Ordinary Linear Model Diagnostic Plots

Next, we want to analyze generalized linear model with a Gamma distribution. The Gamma distribution has the form:

$$f(y) = \frac{1}{\Gamma(v)} \left(\frac{v}{\mu}\right)^{v} e^{\frac{-yv}{\mu}}, \text{ with y>0}$$

Note, v is the shape parameter, $E(Y) = \mu$ and $Var(Y) = \frac{\mu^2}{v}$. The canonical parameter for the Gamma distribution is $\theta = -\frac{1}{\mu}$. The canonical link is $\eta = -\frac{1}{\mu}$.

GLM Coefficients (AIC = 44501.21)

Variable	Coefficient	PValue
(Intercept)	1.924005e-04	3.280902e-251
age	-9.525509e-07	5.308682e-56
bmi	-1.076902e-06	6.940030e-17
children	-1.922309e-06	2.209312e-03
smokeryes	-8.216919e-05	1.052729e-268
regionnorthwest	2.340886e-06	2.786894e-01
regionsoutheast	5.844762e-06	2.487606e-03
regionsouthwest	3.955428e-06	6.908658e-02

Table 4: Generalized Linear Model with Gamma Distribution Summary

According to Table 4, all variables in the GLM model are statistically significant. Additionally, the GLM model exhibits a lower AIC compared to the ordinary linear model, indicating a potentially better predictive performance. Furthermore, as shown in Figure 9, the residual plot shows a similar behavior as the ordinary linear model. Additionally, we can see that residuals are still not normal according to the Q-Q plot.

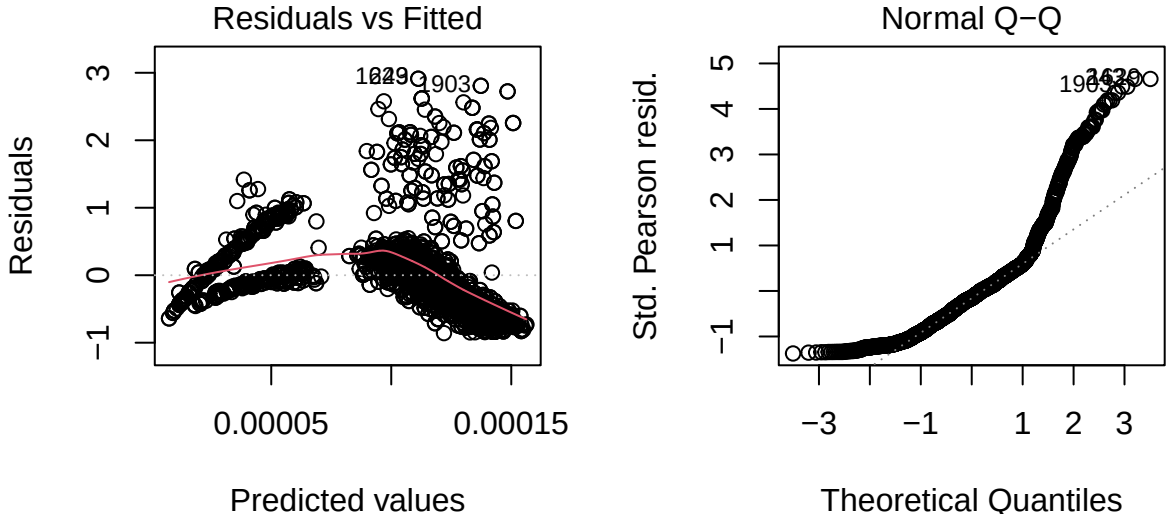

Figure 9: GLM Diagnostic Plots

Random Effect Model

A random effect is a factor with levels that are random selections from a broader population of possible levels, which means we assume a distribution for the factor. We can see that the variable smoker status clearly have a distinct distribution.

$$y_{ijk} = \mu + \tau_i + v_j + \epsilon_{ijk}$$

where μ is the intercept (fixed), τ_i is the effect of i-th level of factor A (fixed), and $v_j \sim N(0, \sigma_v^2)$ is the effect of j-th level of factor B. σ_v^2 and σ^2 are call the variance components.

We initially considered smoker as the random effect with a fixed slope. After constructing the model, we used Kenward-Roger method to test the fixed effect. We use Kenward-Roger method as the F-test with the adjusted degrees of freedom. We found out that all the variables are significant except for the variable sex. We can see that in Table 5, the p-value is greater than 0.05, which means that we fail to reject the null hypothesis that larger model is not significant.

Kenward-Roger Approximation

stat	ndf	ddf	F.scaling	p.value
0.02479147	1	2208	1	0.8749022

Table 5: Kenward-Roger Approximation for Fixed Effects

Based on Table 6, we found that all the variables are significant, so we'll include all the variables as a fixed effect. In addition, $\sigma^2 = 6062.30$ and $\sigma_\alpha^2 = 13528.04$, indicating a substantial amount of variation. According to the table, we can see as the age increase, it increases the age by about 258.178, which means there is a positive correlation between age and the insurance price.

Random Effects (AIC = 44923.01)

Group	Std_Dev
smoker	13528.04
Residuals	6062.30

Fixed Effects

Variable	Coefficient
(Intercept)	-12380.6798
age	258.1784

bmi	334.9547
children	525.4390
smokeryes	23813.0588
regionnorthwest	-357.6551
regionsoutheast	-1212.8820
regionsouthwest	-1120.6436

Table 6: Mixed Effect Model 1 Summary

According to Table 7, smoker status significantly impacts both the slope and intercept of the model, highlighting its critical role in predicting insurance prices. Specifically, for policyholders who smoke, the insurance cost increases by approximately 3525.38 for each additional year of age. In contrast, for non-smokers, the insurance price decreases by about 2483.26 with each additional year of age.

Mixed Effect Model 2 Coefficient by Smoker Status

Parameters	yes	no
(Intercept)	26.81576	-318.17585
age	3525.38302	-2483.25951
sexmale	271.32613	258.85791
bmi	1482.02040	17.72327
children	479.58508	560.14865
smokeryes	-27596.01397	-27596.01397
regionnorthwest	-500.03787	-571.07709
regionsoutheast	-1932.61007	-1148.14706
regionsouthwest	-934.72118	-1435.33495

Table 7: Smoker Status Coefficients

Based on our analysis of the explanatory variables, we identified that the smoker variable has a significant effect on insurance prices. This observation led us to hypothesize that incorporating both a random slope and intercept for the smoker variable could improve the model. Specifically, we allowed the smoker status to influence the slope of all variables and the intercept. As shown in Table 8, the standard deviations for the random effects across all variables are notably high with $\sigma^2 = 4820.513091$, indicating substantial variability associated with the smoker groups. Additionally, we can see that the coefficient for the smoker is typically higher than non-smoker, which means that if the policyholder is a smoker, the model will predict a higher insurance price. The model's AIC is lower than that of Mixed Effect Model 1, suggesting that this more complex random effects structure provides a better fit to the data. By including the random effects across all variables also have a lower residuals than Mixed Effect Model 1.

Random Effects (AIC = 44008.78)

Group	Name	Std_Dev
smoker	Intercept	5433.330170
	sexmale	275.589892
	children	64.840288
	regionnorthwest	197.762464
	regionsoutheast	583.233749
	regionsouthwest	399.284663
	age	8.949801
	bmi	1041.423631
Residuals		4820.513091

Fixed Effects

Variable	Coefficient
(Intercept)	-12380.6798
age	258.1784
bmi	334.9547
children	525.4390
smokeryes	23813.0588
regionnorthwest	-357.6551
regionsoutheast	-1212.8820
regionsouthwest	-1120.6436

Table 8: Mixed Effect Model 2 Summary

According to Figure 10, the residuals plot shows a significant improvement compared to both the ordinary linear model and the generalized linear model. The plot reveals a single cluster around the lower fitted values, while the remaining residuals are more evenly scattered. This indicates that the mixed-effect model is better at capturing high insurance prices, suggesting that it has a better predictive ability in comparison to the other models. The reduced clustering of residuals further supports the idea that this model provides a more accurate and reliable prediction for insurance costs.

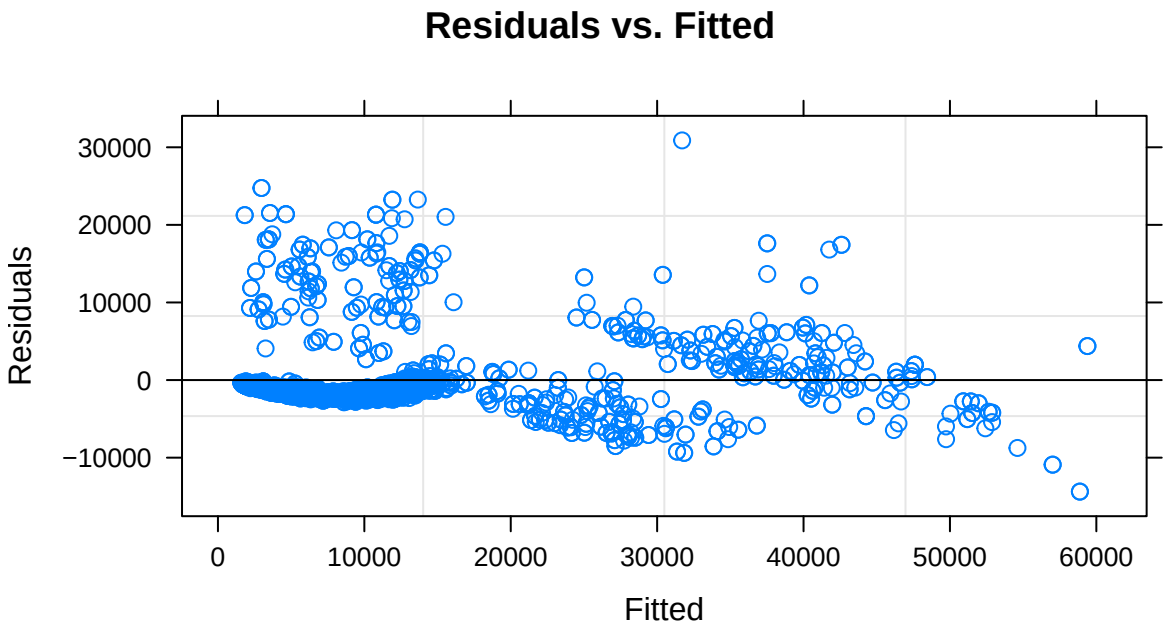

Figure 10: Mixed Effect Model 2 Diagnostic Plot

According to Table 9, the p-value is approximately 0, which is significantly less than the 0.05 threshold. This indicates that we reject the null hypothesis, which posits that the larger model is not significantly better. Therefore, we conclude that Mixed Effect Model 2 outperforms Mixed Effect Model 1. Additionally, the lower AIC value for the Random Effect Model, compared to the ordinary linear model, further supports the inclusion of random effects as a meaningful improvement to the model. This demonstrates that accounting for variability due to smoker status significantly enhances the model's predictive ability.

Anova Table

Model	npar	AIC	Deviance	Chisq	P_value
model_random1	10	44923.01	44903.01	NA	NA
model_random2	45	44008.78	43918.78	984.2345	5.990508e-184

Table 9: Anova Table for Comparing Mixed Effect Model 1 and 2

Results

To evaluate the predictive performance of the three models: linear regression, Gamma GLM, and mixed effect model. We will calculate the Mean Absolute Error (MAE) for each. MAE is a useful metric as it gives the average of the absolute differences between predicted and actual values, providing a clear indication of the model's accuracy. By comparing the MAE across these models, we can identify which one provides the most accurate predictions for

insurance prices, helping us determine the best approach for forecasting. After analyzing the model, we concludes that the Mixed Effect Model 2 performs the best with the lowest AIC. Now, we want to check the predictive ability of these models.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Absolute Error of each Model

Models	Mean.Absolute.Error
Ordinary Linear Model	4190.462
Generalized Linear Model with Gamma Distribution	5088.020
Mixed Effect Model with Smoker as Random Effect	2911.796

Table 10: Comparison of Mean Abosolute Error of each Model

According to Table 10, the Mixed Effect Model outperforms the other models with a significantly lower MAE of 2911.796, compared to 4190.462 for the linear model and 5088.020 for the GLM. This indicates that the Mixed Effect Model has the best predictive accuracy, highlighting the crucial role of smoker status in determining insurance prices. The linear model and GLM struggle to accurately capture the variability in insurance prices, particularly at the higher end, where smoker status has a substantial impact. This suggests that incorporating smoker status as a random effect improves the model's ability to predict insurance costs.

Conclusions

Determining the premium price is essential for insurance companies to balance their payouts to policyholders while maintaining financial stability. The models proposed in this project offer valuable insights that can help the company accurately predict the appropriate yearly premium to charge each policyholder. By incorporating factors such as smoker status and other key variables, these models enable more informed decisions, ensuring that premium pricing aligns with the risk profile of each policyholder and supports the company's long-term sustainability.

In this analysis, we aimed to develop models to predict insurance prices based on various factors like age, BMI, smoking status, and region. The models tested included a standard linear model, a generalized linear model (Gamma), and a mixed-effect model with smoker status as a random effect. By evaluating these models, we sought to identify the most accurate predictor of insurance costs, given that smoker status is likely a significant factor influencing prices.

After comparing the models, we found that the mixed-effect model performed the best, with a significantly lower mean absolute error (MAE) than both the linear and GLM models. The

key factor contributing to the mixed-effect model's success is the inclusion of smoker status as a random effect. This accounts for the variation in insurance prices between smokers and non-smokers, which the other models struggle to capture. The random effect of smoker status helps the model better fit the data, especially in cases of higher medical costs associated with smoking.

In conclusion, the mixed-effect model is the best option for predicting insurance prices, as it effectively captures the impact of smoker status on pricing. By treating smoker status as a random effect, the model improves its predictive accuracy and provides more reliable results. This analysis shows that incorporating random effects can significantly enhance the model's ability to handle complex data, leading to better predictions in scenarios like insurance pricing.

References

[1] Teertha. (n.d.). US Health Insurance Dataset. Kaggle. Retrieved December 5, 2024, from https://www.kaggle.com/datasets/teertha/ushealthinsurancedataset