Predicting Healthcare Costs: A Regression-Based Analysis of Key Cost Drivers and Model Optimization

Abstract

This report focuses on predicting individual medical charges billed by health insurance companies using linear regression models to identify key factors that influence healthcare costs and develop a reliable and interpretable predictive framework. However, constructing a robust and accurate model required overcoming several challenges, including heteroscedasticity, multicollinearity, and influential data points.

Heteroscedasticity, characterised by unequal residual variances across fitted values, posed a major threat to the reliability of the model by distorting p-values and confidence intervals. Multicollinearity, which occurs when predictor variables are highly correlated, was carefully assessed to ensure the independence of variables and the validity of their contributions to the model. Additionally, influential data points with high leverage or extreme values were identified and managed appropriately, as they had the potential to disproportionately skew model estimates.

Log transformations were used to stabilise the variance of the response variable, effectively mitigating heteroscedasticity. Furthermore, the Weighted Least Squares (WLS) method was employed to assign appropriate weights to observations with higher variability, thereby improving the model's robustness and accuracy. These measures collectively enhanced the reliability of the regression model and ensured its assumptions were met.

The refined model provided improved interpretability and predictive accuracy, making it a valuable tool for various practical applications. The findings offer actionable insights for healthcare policymakers, enabling them to identify high-cost drivers, such as smoking, and design targeted interventions to reduce costs. Healthcare providers and insurers can leverage the model to optimise resource allocation and predict future expenditures. Moreover, these insights support the development of cost management strategies that enhance the efficiency and effectiveness of healthcare delivery systems.

By addressing diagnostic issues and refining predictors, this study presents a comprehensive and methodologically sound approach to understanding and predicting healthcare costs.

Introduction

The report focuses on the challenge of predicting medical expenses and identifying the main factors influencing healthcare costs billed by insurance companies. As healthcare costs continue to rise globally, insurers, policymakers, and providers must understand what drives these expenses. This study aims to uncover the key factors behind healthcare costs and create a predictive model that not only explains these variations but also helps design strategies to manage and reduce expenses. Insights from this study are particularly important for policymakers looking to allocate resources wisely, improve cost efficiency, and ensure healthcare systems remain financially sustainable.

Accurately predicting medical expenses benefits several stakeholders. Insurers can use this information to set fair premiums and assess risks more effectively. Healthcare providers can better allocate their resources, and policymakers can develop targeted programs, like anti-smoking campaigns or preventive healthcare initiatives, to reduce overall costs. Building a reliable model, therefore, has practical and far-reaching benefits.

However, developing such a model presents challenges. For example, heteroscedasticity—where data inconsistencies can affect statistical accuracy—needs to be addressed to ensure reliable results. Multicollinearity, or strong correlations between predictors, must also be carefully managed to ensure each factor's influence is clear and stable. Outliers and unusual data points add another layer of complexity, as they can distort results if not handled properly. Overcoming these challenges is critical to creating a trustworthy and understandable model.

The study identifies three key factors that influence medical charges: age, Body Mass Index (BMI), and smoking status. Among these, smoking stands out as the most significant driver, showing how it not only raises individual costs but also places a heavy burden on healthcare systems. Age and BMI are also important, highlighting the role of both demographic and health-related factors in determining costs.

The report follows a structured approach. It starts with an analysis of the dataset, including its characteristics and variables, and outlines the methods used to build the Multiple Linear Regression model. Various issues like heteroscedasticity, multicollinearity, and outliers are identified and resolved using techniques like log transformations and Weighted Least Squares (WLS) to improve the model's accuracy. The findings are then discussed in the context of their implications for healthcare policy, cost management, and future research.

By addressing these challenges and offering clear, actionable insights, the study provides valuable guidance. Policymakers can focus on tackling high-cost drivers like smoking, while insurers and healthcare providers can optimize their strategies for managing expenses. Ultimately, this report serves as a comprehensive guide to understanding and predicting healthcare costs, supporting efforts to make healthcare systems more efficient, effective, and fair.

Problem Statement and Data Source

The objective of this study is to predict individual medical charges billed by health insurance and identify the factors strongly associated with these costs. Understanding these drivers is crucial for enabling insurers, healthcare providers, and policymakers to make informed decisions regarding cost management and resource allocation. Factors such as age, BMI, and smoking status are evaluated for their influence on medical expenses, with smoking identified as a particularly significant driver. The study also addresses challenges like heteroscedasticity, multicollinearity, and outliers to ensure the model's reliability and interpretability. By providing actionable insights, this study aims to support better financial planning, targeted interventions, and efficient healthcare resource distribution. The dataset used for this project was sourced from Kaggle (4).

Proposed Methodology

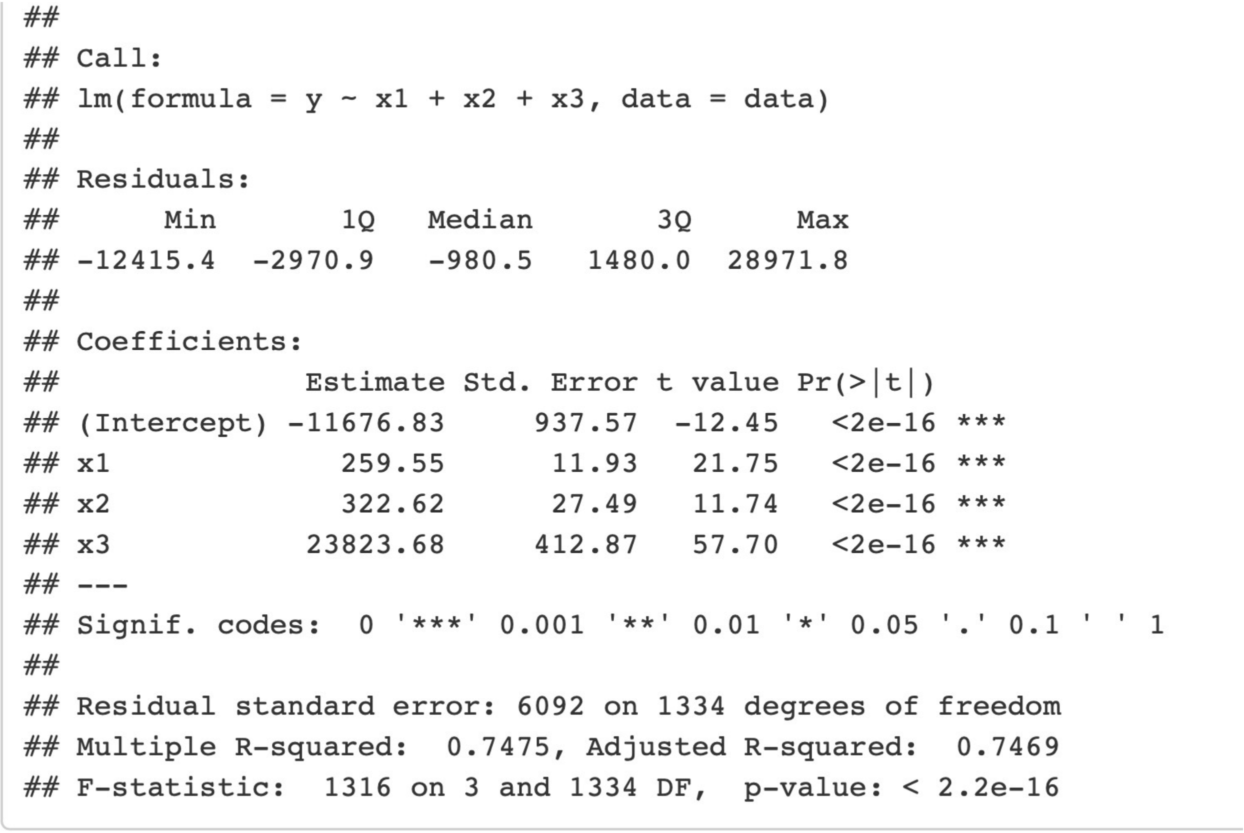
The proposed methodology for this study involves a systematic approach to model fitting, diagnostic evaluation, and remediation to address the identified challenges. In the first phase, a multiple linear regression model was developed to predict medical charges, with the response variable being continuous medical charges and the predictor variables including age (continuous), BMI (continuous), and smoking status (categorical). The model's significance was evaluated using the F-test, and all predictors were found to have p-values less than 0.05, indicating statistically significant relationships with medical charges. Key findings from this model included a baseline charge (intercept) of -11,676.83 units, an increase of 259.55 units in charges for each additional year of age, a rise of 322.62 units for every unit increase in BMI, and a striking 23,823.68-unit increase for smokers compared to non-smokers, highlighting smoking as a critical cost driver. The model's R^2 value of 0.7475 and adjusted R^2 confirmed that 75% of the variance in charges was explained without including unnecessary predictors.

In the second phase, regression diagnostics were performed to assess model assumptions. Residual plots revealed heteroscedasticity, where the variance of residuals was not constant, impacting the reliability of p-values and confidence intervals. Multicollinearity was evaluated using Variance Inflation Factor (VIF) values, which were low, indicating no significant issues among predictors. Influential data points were identified using Cook’s Distance, which highlighted observations with the potential to skew results.

To address these issues, the third phase involved remediation techniques. Heteroscedasticity was mitigated by applying a log transformation to medical charges, stabilising variance and improving interpretability. Additionally, a Weighted Least Squares (WLS) approach was used to assign weights to observations with higher variance, further enhancing robustness. Influential points were carefully examined and either retained or adjusted based on their validity to ensure they did not unduly affect the model. Comparisons between the original and refined models demonstrated significant improvements. The log-transformed model reduced heteroscedasticity and allowed percentage-based interpretations, while the WLS model offered enhanced robustness at the cost of added complexity. These steps collectively ensured the development of a reliable, interpretable, and robust predictive model for medical charges.

Analysis and Results

The figure below shows the final results for the dataset after all the algorithm implementations.



```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12415.4  -2970.9   -980.5   1480.0  28971.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11676.83     937.57  -12.45   <2e-16 ***
## x1             259.55      11.93   21.75   <2e-16 ***
## x2             322.62      27.49   11.74   <2e-16 ***
## x3           23823.68     412.87   57.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6092 on 1334 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7469
## F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

The regression analysis results show a very low p-value, which is below the significance threshold of α = 0.05. This indicates that the null hypothesis can be rejected, meaning that at least one predictor has a significant impact on the variation in the response variable, which is the charges. Notably, variables like age, smoker status, and BMI are found to have a statistically significant association with charges.

Examining the individual p-values reveals that each predictor's p-value is well below the significance level of α, confirming their statistical significance. The intercept of -11,676.83 represents the baseline charge when age, BMI, and smoker status are all zero. The age coefficient, 259.55, indicates that for every additional year of age, charges are predicted to increase by 259.55 units, assuming BMI and smoker status remain unchanged. Similarly, the BMI coefficient of 322.62 suggests that a one-unit rise in BMI corresponds to a 322.62-unit increase in charges, with other factors held constant. The smoker variable has the most substantial influence, with a coefficient of 23,823.68, showing that smokers are expected to incur $23,823.68 more in charges than non-smokers, emphasising the strong link between smoking and elevated healthcare costs.

The R-squared value of 0.7475 indicates that about 75% of the variability in charges is accounted for by the predictor’s age, BMI, and smoker status, highlighting the model's strong fit to the data. Furthermore, the adjusted R-squared value is nearly identical to the R-squared, suggesting that all included variables contribute meaningfully to the model and that none are unnecessary. Overall, the model successfully captures the relationships between the predictors and the response variable, offering valuable insights into the factors that impact healthcare charges.

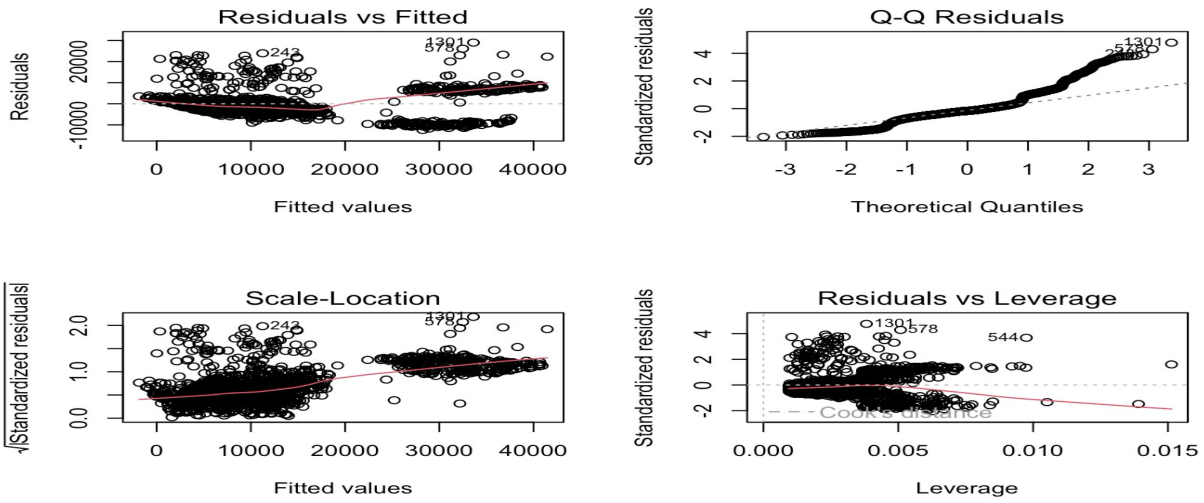


Figure 6.1

Figure 6.1 presents diagnostic plots that offer crucial insights into the linear regression model's performance and pinpoint areas needing improvement. The Residuals vs. Fitted plot shows evidence of heteroscedasticity, with residual variance increasing at higher fitted values. This violates the assumption of constant variance, indicating the model may not fully capture the relationship between the predictors and the response variable. The Q-Q plot reveals deviations from the diagonal line, particularly at the tails, suggesting that the residuals do not strictly follow a normal distribution and highlighting the need for transformations or adjustments to better meet model assumptions. The Scale-Location plot also displays a pattern in the residual spread as fitted values increase, further confirming heteroscedasticity and suggesting the potential use of methods like log transformation or Weighted Least Squares to stabilise the variance. Additionally, the Residuals vs. Leverage plot identifies data points such as 1301, 578, and 544 as having high leverage with significant Cook's Distance values, indicating they are influential and could heavily affect the model's estimates. These observations emphasise the need to address heteroscedasticity, assess the impact of influential points, and ensure that linear regression assumptions are met. Applying necessary transformations, refining predictor variables, or considering alternative modelling approaches can greatly improve the model's reliability and robustness.

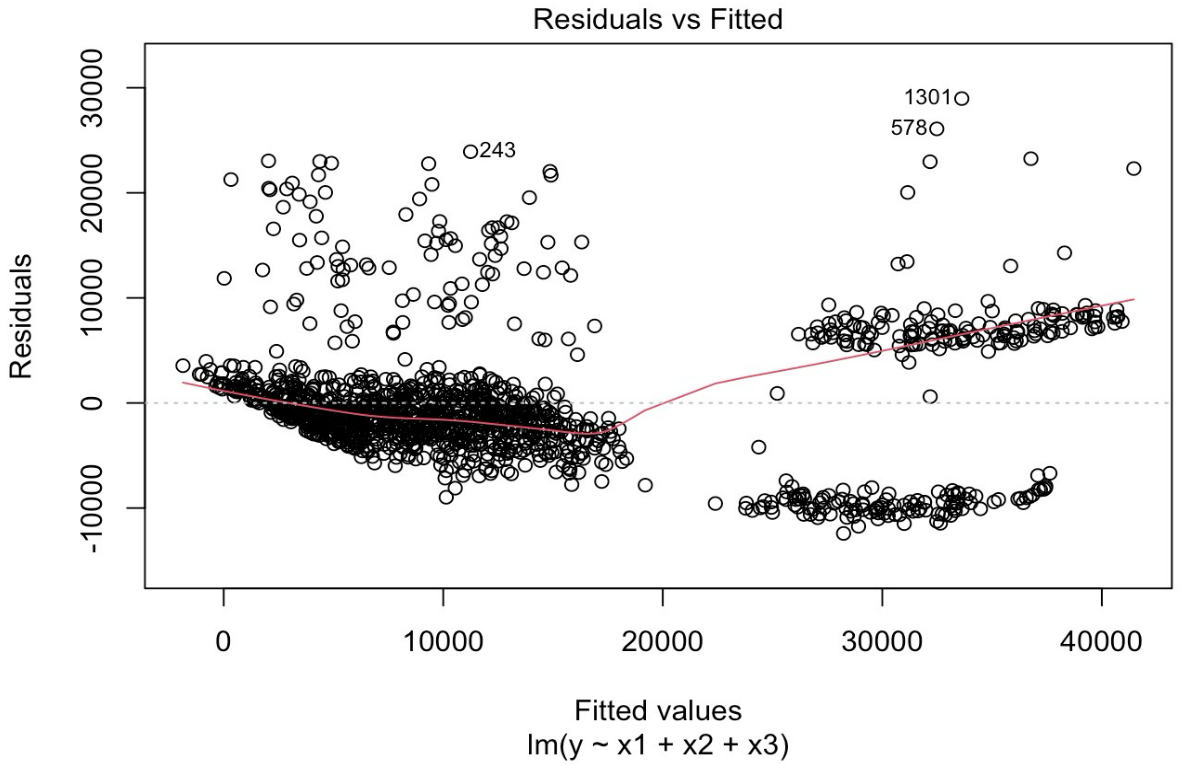


Figure 6.2

"Residuals vs. Fitted" plot, as shown in Fig 6.2, provides valuable insights into the model's performance and areas requiring refinement. The red smooth line indicating a non-linear pattern in the residuals suggests that the model may not adequately capture the underlying relationship between the predictors and the response variable, potentially necessitating the inclusion of non-linear terms or transformations. Moreover, the residuals exhibit increasing variance as fitted values rise, signalling heteroscedasticity and a violation of the constant variance assumption in linear regression. Addressing this issue may involve applying data transformations or using alternative techniques such as weighted least squares regression. Outliers, including points labelled 243, 578, and 1301, appear to exert a strong influence on the model and should be further investigated. Additionally, residual clustering at specific fitted values indicates that important variables or interactions may be missing from the model. These observations underline the need for model adjustments to enhance accuracy and reliability.

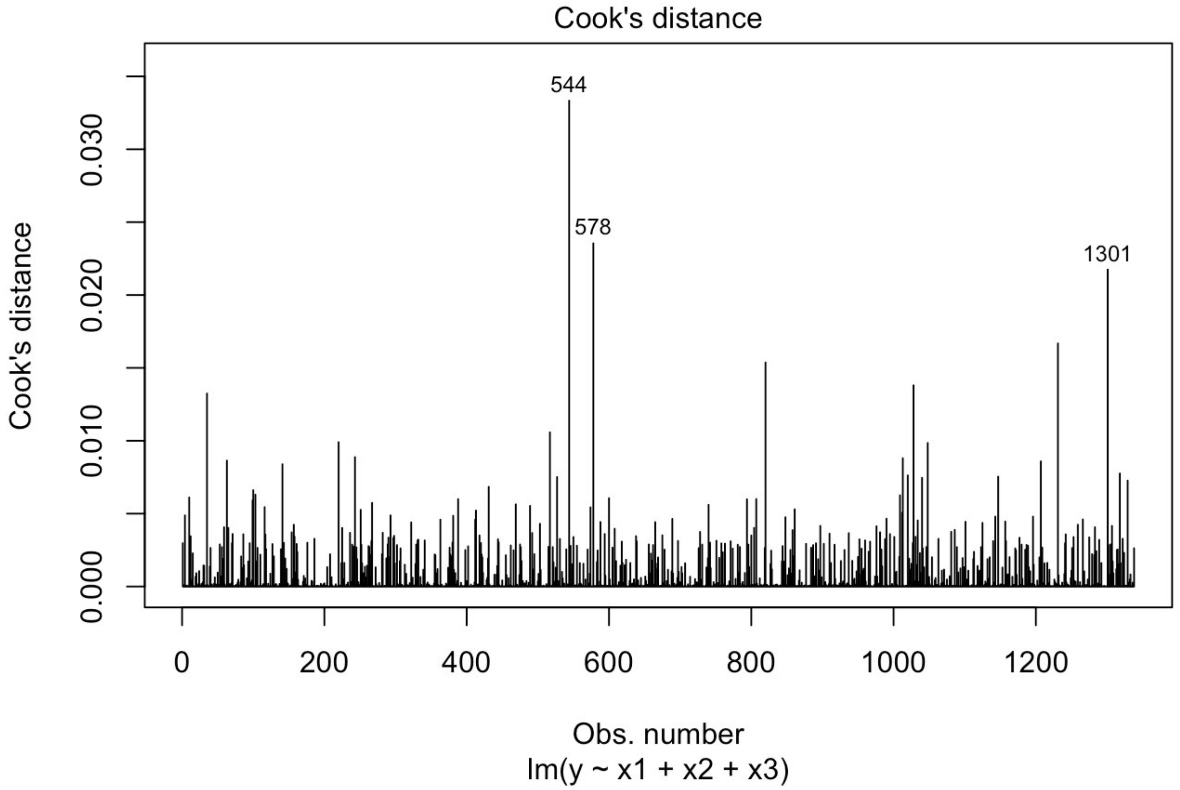


Figure 6.3

Cook's Distance plot, as shown in Fig 6.3, identifies key influential observations in the dataset, particularly points 544, 578, and 1301, which display significantly higher Cook's Distance values compared to the rest. These points have a considerable impact on the regression model, potentially compromising the stability and accuracy of the estimated coefficients. Their high Cook's Distance values indicate they may be outliers or extreme cases that require further investigation. It is important to determine whether these observations are due to data entry errors, valid extreme values, or represent a distinct subgroup that may need separate analysis. Ignoring them could lead to biased predictions and misinterpretation of the model's results. Conducting sensitivity analyses, such as refitting the model after excluding these points, can help evaluate their influence on the outcomes. Addressing these influential data points is critical for enhancing the model's reliability and validity.

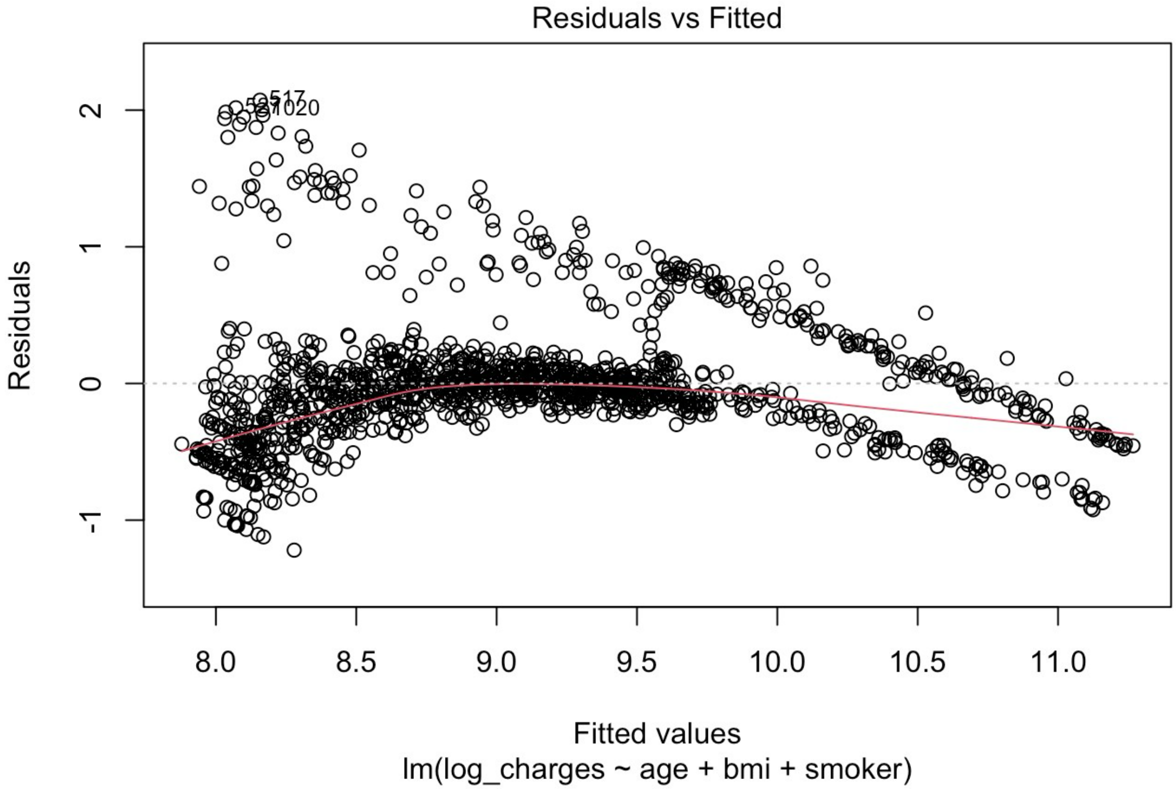


Fig 6.4

"Residuals vs. Fitted" plot, as shown in Fig 6.4, for the model using log-transformed charges as the response variable provides several observations. The red smooth line indicates a slight non-linear trend in the residuals, implying that the model may not fully capture the relationship between the predictors (age, BMI, and smoker status) and the log-transformed charges. While most residuals are centred around zero, signs of heteroscedasticity are present, with residual variance appearing higher at both lower and upper fitted values, suggesting a violation of the constant variance assumption in linear regression. Additionally, extreme residuals, such as those near points 515 and 1020, may represent outliers or influential observations that warrant further examination. These insights indicate that the model could benefit from adjustments to address non-linearity and heteroscedasticity, such as incorporating interaction terms polynomial features or exploring alternative modelling approaches that better fit the data's structure.

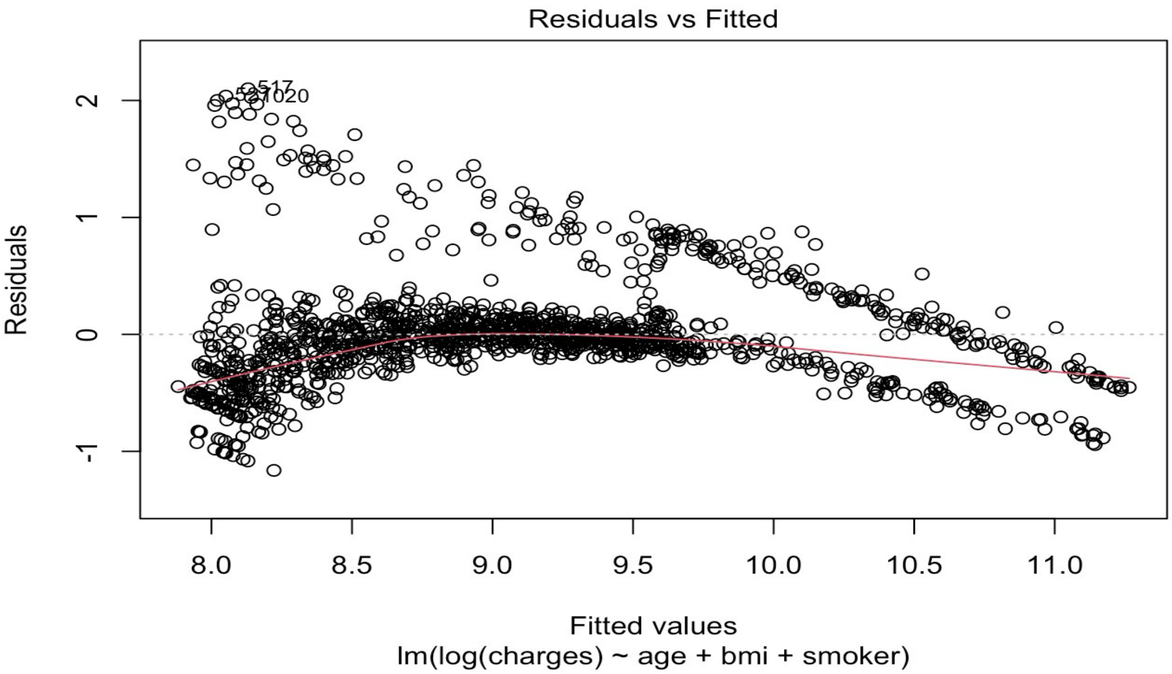


Figure 6.5

"Residuals vs. Fitted" plot, as shown in Fig 6.5, offers significant insights into the model's performance when using log-transformed charges as the response variable. The red smooth line shows a slight curvature, which hints at some non-linearity in the data. This suggests that the current model may not fully capture the complex relationship between the predictor's age, BMI, and smoker status—and the log-transformed charges. Addressing this non-linearity may require the addition of interaction terms or polynomial features to better align the model with the underlying structure of the data. Moreover, while the residuals are generally distributed around zero, there are clear signs of heteroscedasticity. Specifically, the residuals exhibit greater variability at the lower and upper ends of the fitted values, violating the constant variance assumption fundamental to linear regression. This inconsistency in residual spread could undermine the accuracy and reliability of the model's predictions. Techniques such as applying further transformations, using Weighted Least Squares, or considering generalised linear models may help address this issue. Additionally, specific observations, such as points 515 and 1020, deviate significantly from the main cluster of residuals. These outliers or high-leverage points may disproportionately influence the model's coefficients and predictions. A deeper investigation into these points is essential to determine if they result from data entry errors, are valid extreme cases, or represent unique subgroups in the data. In summary, the plot highlights the need for model refinement to address non-linearity, heteroscedasticity, and the impact of influential points. By implementing adjustments such as interaction terms, alternative modelling techniques, or robust regression methods, the model's accuracy and robustness can be significantly enhanced, providing more reliable insights into the factors influencing charges.

Conclusion

In summary, this study highlights that age, BMI, and smoking status are significant determinants of individual medical charges, with smoking status standing out as the most influential factor driving higher costs. The analysis underscores the importance of addressing diagnostic issues such as heteroscedasticity and the presence of influential data points to enhance the model's reliability and accuracy. By applying a log transformation to the response variable, variance was effectively stabilized, addressing a key violation of linear regression assumptions. Additionally, careful management of influential observations ensured that their impact did not distort the model’s estimates, thereby improving the robustness of the analysis. The final log-transformed regression model provides a strong foundation for predicting medical charges and delivers valuable insights into the primary factors influencing healthcare costs. These findings have practical implications for healthcare policy, enabling targeted interventions aimed at reducing costs, particularly for high-risk groups such as smokers. Furthermore, this model can guide financial planning and resource allocation by helping policymakers and healthcare providers better understand the drivers of medical expenses. Future research could expand on this work by incorporating additional predictors, such as lifestyle factors, comorbidities, or geographic variables, to capture a broader range of influences on medical costs. Exploring non-linear relationships or using advanced modeling techniques like machine learning could further improve predictive accuracy and align the model more closely with the complexities of real-world healthcare data. These advancements would enhance the utility of the model in informing policy decisions and optimizing healthcare resource distribution.

Significance and Interpretation of Predictors

The predictors age, BMI, and smoker status were found to be highly significant in predicting medical charges, with p-values below 0.001 for each variable. Specifically:

- **Age**: Each additional year of age was associated with an increase in medical charges, demonstrating a positive relationship.
- **BMI**: Higher BMI was linked to increased charges, consistent with the health risks associated with obesity.
- **Smoker Status**: Being a smoker had the most substantial impact, with smokers incurring significantly higher costs than non-smokers, underscoring the financial burden of smoking on healthcare.

Diagnostic Issues and Implications

Heteroscedasticity: The residual variance was observed to increase with fitted values, violating the assumption of constant variance. This issue could reduce the reliability of p-values and confidence intervals if not addressed.

Influential Points: Cook’s Distance identified a few influential observations with disproportionately high impacts on the model’s results, highlighting the need for careful examination and potential mitigation of these data points.

Effectiveness of Remediation Techniques

Log Transformation: Applying a log transformation to the response variable successfully stabilized the residual variance, addressing heteroscedasticity and simplifying interpretation. With this transformation, model coefficients now represent percentage changes in charges, enhancing the clarity of predictor effects.

Weighted Least Squares (WLS): WLS further reduced the impact of high-variance observations and provided additional robustness. However, it introduced interpretive complexity and did not significantly improve the model beyond the log transformation alone.

Comparison of Remediation Techniques

The log-transformed model proved more effective and straightforward than WLS in resolving heteroscedasticity. While WLS offered added robustness, the log transformation was sufficient to stabilize variance and ensure interpretability, making it the preferred method for this analysis.

Overall Suitability and Practical Recommendations

Suitability: The final log-transformed model is well-suited for predicting medical charges, striking a balance between interpretability and statistical reliability. It offers meaningful insights into the effects of age, BMI, and smoking status on healthcare costs, making it effective for both prediction and inference.

Future Recommendations: Future analyses could benefit from employing robust regression techniques if influential points remain problematic. Including additional predictors may capture more complex relationships, and periodic validation will ensure the model’s applicability to diverse datasets. The logtransformed model is an effective and practical tool for predicting medical charges. It addresses critical assumptions of linear regression while providing clear and interpretable insights into the key drivers of medical costs. By maintaining statistical rigor and balancing simplicity, the model serves as a valuable framework for understanding and managing healthcare expenses.

Bibliography and Credits

1) Aiken, L. S., West, S. G., & Pitts, S. C. (2003). Multiple linear regression

2) Kutner, M., Nachtsheim, C., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*.

3) G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 2nd ed. Springer, 2021.

4) Data Source: https://www.kaggle.com/datasets/mirichoi0218/insurance/data