STATISTICAL ANALYSIS OF INSURANCE CHARGES: A PREDICTIVE MODELING STUDY

BY

PABLO LEYVA

pl33@njit.edu

Project One Report

Submitted in completion of the requirements
for the course of Statistical Learning Capstone

New Jersey Institute of Technology

Newark, New Jersey

Abstract

This study presents a comprehensive statistical analysis of insurance charges using demographic and health-related predictors. We employed linear regression and Ridge regression techniques to identify key factors influencing insurance costs and develop predictive models. Our analysis reveals that smoking status is the most significant predictor of insurance charges, followed by the number of children, BMI, age, region, and sex. The final linear regression model achieved an R-squared value of 0.769, explaining approximately 77% of the variance in insurance charges.

1 Introduction to the Study

Healthcare costs and insurance premiums have become increasingly important topics in modern society. Understanding the factors that influence insurance charges is crucial for both insurance companies in risk assessment and for individuals in making informed healthcare decisions. This study analyzes a comprehensive dataset of insurance charges to identify the key demographic and health-related factors that drive insurance costs.

2 Objective

The primary objectives of this study are:

1. **Exploratory Analysis**: To conduct comprehensive exploratory data analysis to understand the distribution of insurance charges and identify patterns in the relationship between predictor variables and insurance costs.
2. **Feature Assessment**: To evaluate the relative importance of demographic and health-related factors (age, sex, BMI, number of children, smoking status, and geographical region) in predicting insurance charges.
3. **Predictive Modeling**: To develop and evaluate linear regression models for predicting insurance charges, including regularized models to handle potential overfitting.
4. **Model Validation**: To assess model performance using appropriate statistical metrics and diagnostic tests to ensure model reliability and validity.
5. **Policy Insights**: To provide actionable insights that can inform insurance pricing strategies and help individuals understand factors affecting their insurance premiums.

These objectives aim to provide a comprehensive understanding of insurance charge determination and develop robust predictive models that can serve both academic and practical purposes in the insurance industry.

3 Statistical Analysis

3.1 Data Description and Preprocessing

The dataset comprises 1,338 observations with seven variables:

- **Age**: Continuous variable (integer values)
- **Sex**: Binary categorical variable (male/female)
- **BMI**: Continuous variable representing body mass index
- **Children**: Discrete variable indicating number of dependents
- **Smoker**: Binary categorical variable (yes/no)
- **Region**: Categorical variable with four levels (southeast, southwest, northeast, northwest)
- **Charges**: Continuous dependent variable (insurance charges in USD)

Data preprocessing involved encoding categorical variables using appropriate techniques:

- Sex and smoker status were encoded using dummy variables with drop_first=True
- Region was mapped to numerical values (southeast=0, southwest=1, northeast=2, northwest=3)
- Data types were optimized for memory efficiency: age, children, and region variables were converted to `uint8`, while BMI and charges were converted to `float32`, resulting in a 25% memory reduction (from 73.3+ KB to 55.0 KB)

3.2 Exploratory Data Analysis

The exploratory analysis revealed several key insights:

3.2.1 Distribution Analysis

Insurance charges exhibited a right-skewed distribution with mean charges significantly higher than median charges. Log transformation revealed an underlying normal distribution, suggesting the presence of multiplicative effects in the data.

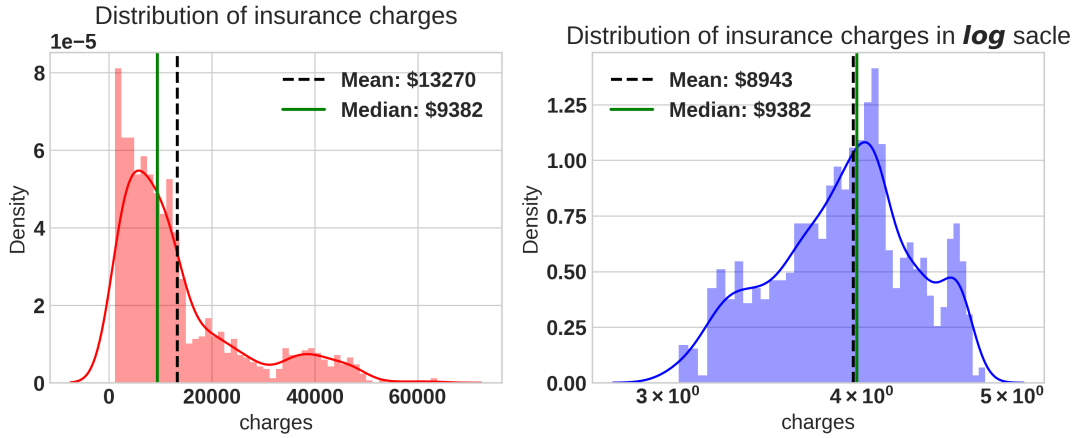

Figure 1: Distribution of insurance charges showing right-skewed pattern (left) and log-transformed distribution (right)

3.2.2 Correlation Analysis

Initial correlation analysis showed no strong linear correlations between predictor variables, indicating minimal multicollinearity concerns and that features cannot be expressed as linear combinations of each other.

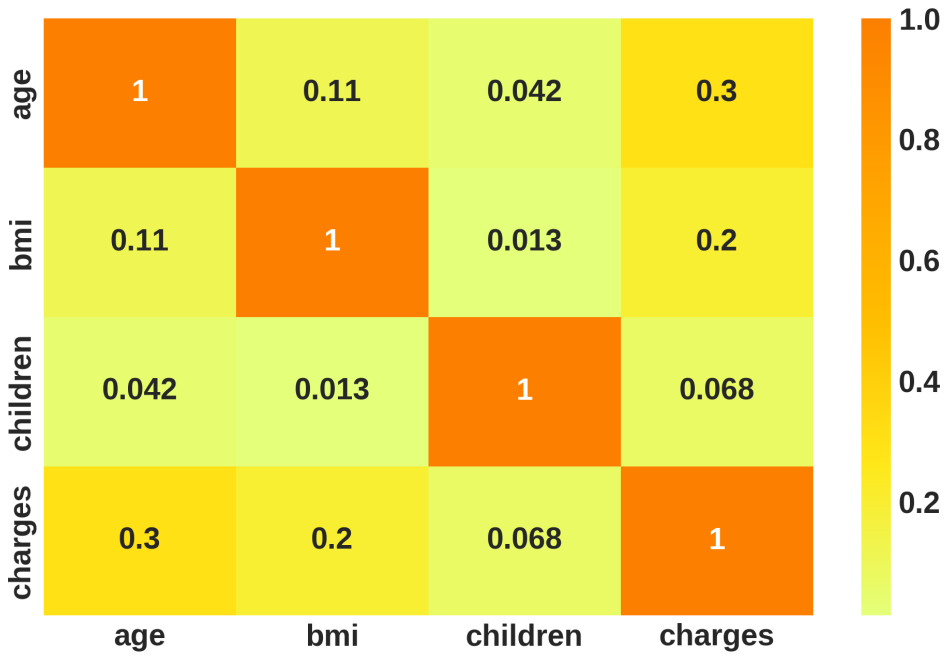

Figure 2: Correlation matrix heatmap showing relationships between numerical variables

Feature-Target Relationships:

Figure 3: Violin plots showing charge distributions by sex and smoking status

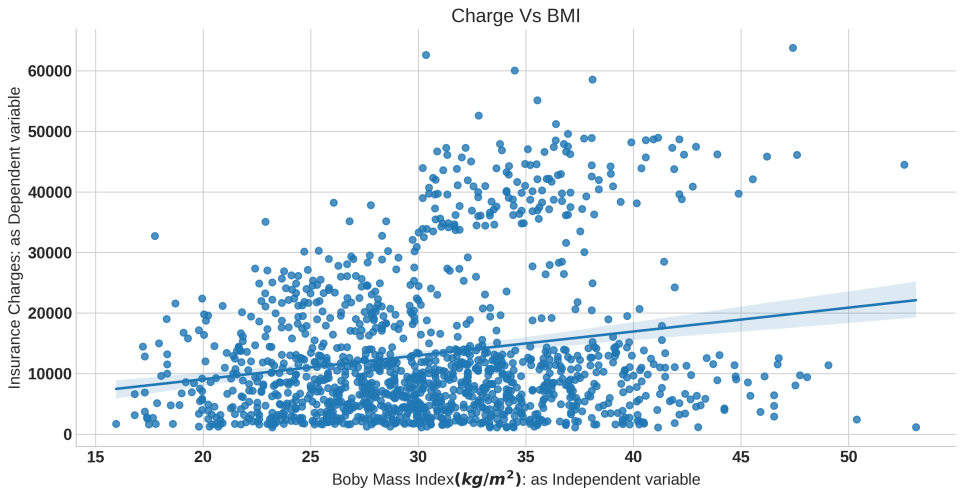

Figure 4: Scatter plot showing the relationship between BMI and insurance charges

- **Smoking Status**: Violin plots revealed a dramatic difference between smokers and non-smokers, with smokers showing both higher central tendency and greater variance in charges.
- **BMI**: Scatter plot analysis revealed a complex relationship where the impact of BMI on charges is amplified for smokers, particularly at higher BMI values (>30).
- **Age**: Showed a positive relationship with charges, with the effect varying by smoking status, revealing three distinct charge levels.
- **Sex**: Minimal difference in charge distributions between males and females.
- **Children**: Similar charge distributions across different numbers of children.
- **Region**: No significant differences in charge distributions across geographical regions.

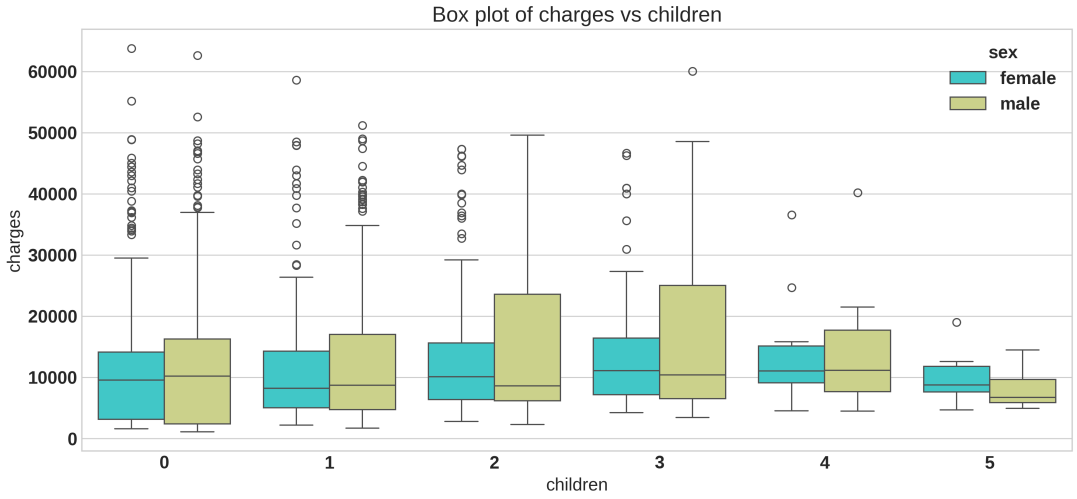

Figure 5: Box plots showing charge distributions by number of children and sex

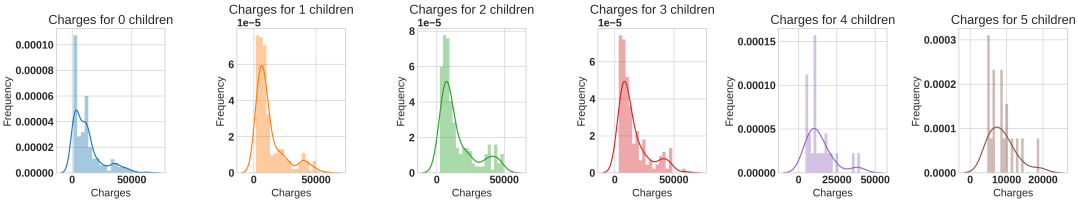

Figure 6: Distribution of charges segmented by number of children

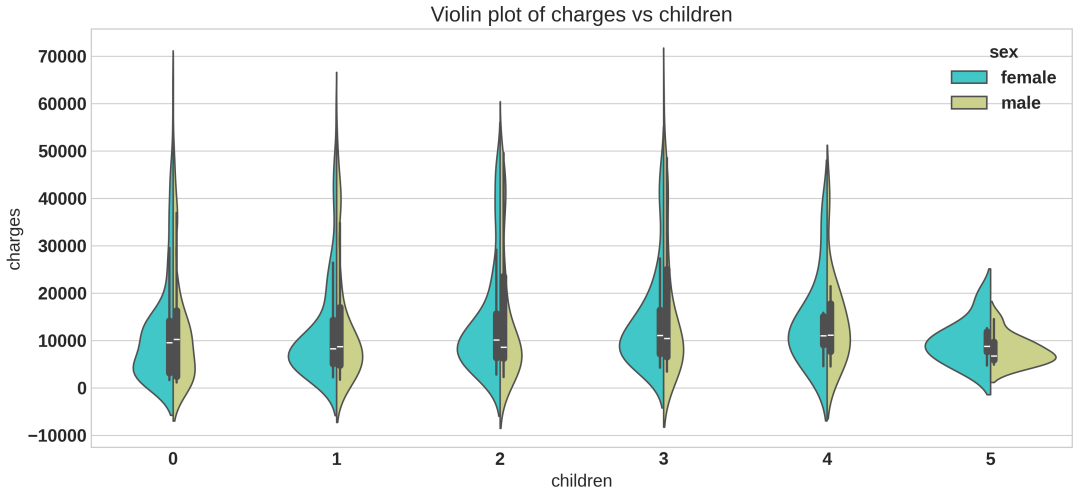

Figure 7: Violin plots of charges vs children by sex

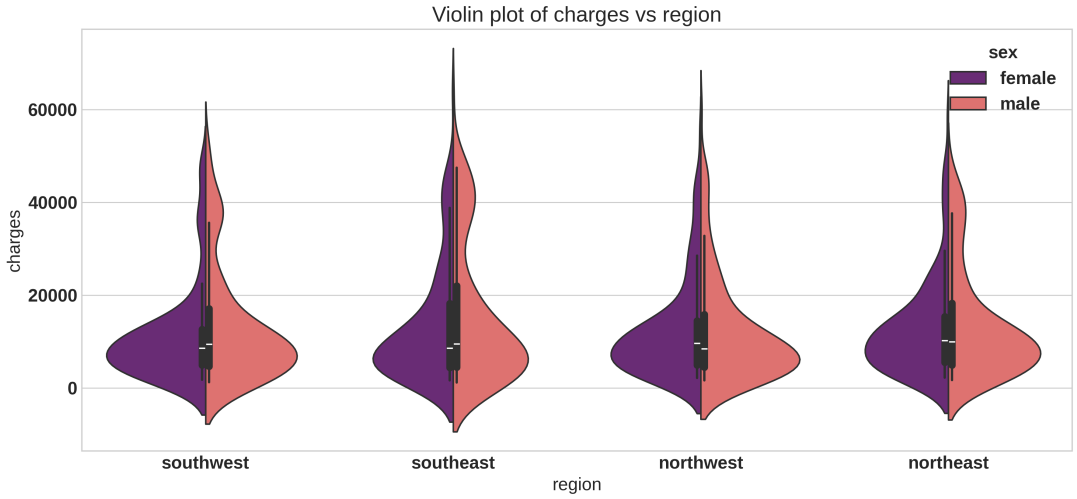

Figure 8: Violin plots of charges vs region by sex

3.3 Model Development

3.3.1 Linear Regression

A multiple linear regression model was initially fitted as the primary approach using all available predictors:

$$\hat{y} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{sex} + \beta_3 \cdot \text{BMI} + \beta_4 \cdot \text{children} + \beta_5 \cdot \text{smoker} + \beta_6 \cdot \text{region}$$

This baseline model served as the foundation for understanding the relationships between predictors and insurance charges.

3.3.2 Ridge Regression

Following the initial linear regression analysis, Ridge regression with L2 regularization was implemented:

$$\min_{\boldsymbol{\beta}} \left\{ \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda \|\boldsymbol{\beta}\|_2^2 \right\}$$

Regularization paths were analyzed across different alpha values (1 to 99) to understand coefficient behavior under varying regularization strengths and to validate the stability of the linear regression findings.

3.3.3 Why Ridge?

Given the large value of the Smoker coefficient observed in the initial linear regression model: Ridge regression was employed to check the weights and stabilize the coefficients to prevent potential overfitting. After analyzing the regularization paths, it was found that the Smoker coefficient was reduced to 14,374.39, which is still significantly larger than the other coefficients, confirming the robustness of this finding.

3.4 Model Diagnostics

Several diagnostic tests were performed to validate model assumptions:

- **Linearity**: Assessed through actual vs. predicted value scatter plots
- **Residual Normality**: Evaluated using distribution plots and Q-Q plots
- **Variance Inflation Factor**: Calculated to assess multicollinearity

4 Analysis Results

4.1 Model Performance

The linear regression model achieved the following performance metrics:

- **R-squared**: 0.7694 (76.94% of variance explained)
- **Mean Squared Error**: 33,806,944
- **Variance Inflation Factor**: 4.33

These results indicate that the model explains approximately 77% of the variance in insurance charges, representing good predictive performance for this type of analysis.

4.2 Feature Importance Analysis

The linear regression coefficients revealed the relative importance of different factors:

Feature	Coefficient	Absolute Value
Smoker	23,644.44	23,644.44
Children	425.61	425.61
BMI	344.77	344.77
Age	261.17	261.17
Region	234.58	234.58
Sex	106.59	106.59

Table 1: Feature importance based on linear regression coefficients

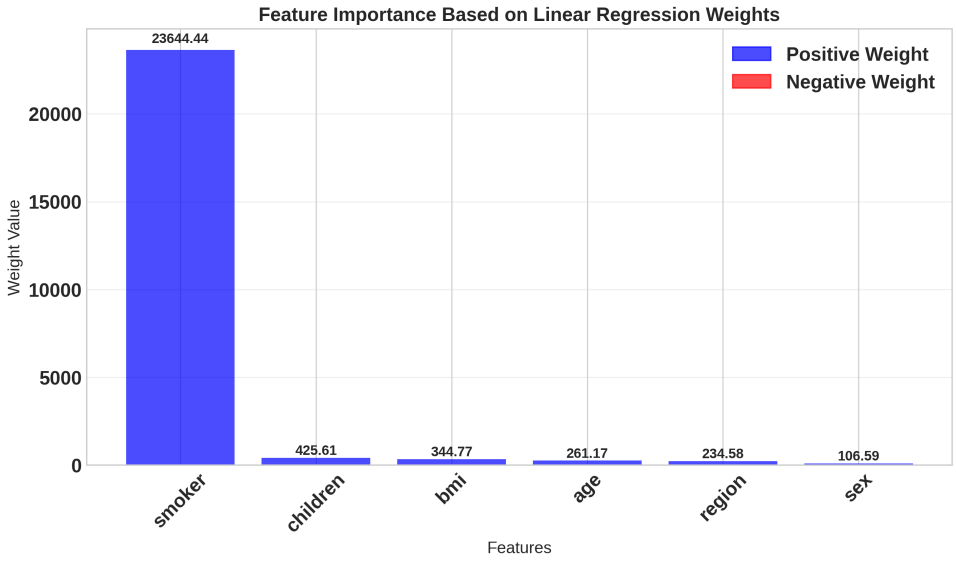

Figure 9: Visual representation of feature importance based on linear regression weights

Key Findings:

- **Smoking Status**: By far the most influential factor, with smokers paying approximately $23,644 more than non-smokers, all else being equal.
- **Number of Children**: Each additional child increases insurance charges by approximately $426.
- **BMI**: Each unit increase in BMI results in approximately $345 increase in charges.
- **Age**: Each additional year of age increases charges by approximately $261.
- **Region and Sex**: Relatively minor effects on insurance charges.

4.3 Ridge Regression Results

The Ridge regression analysis demonstrated how regularization affects coefficient estimates:

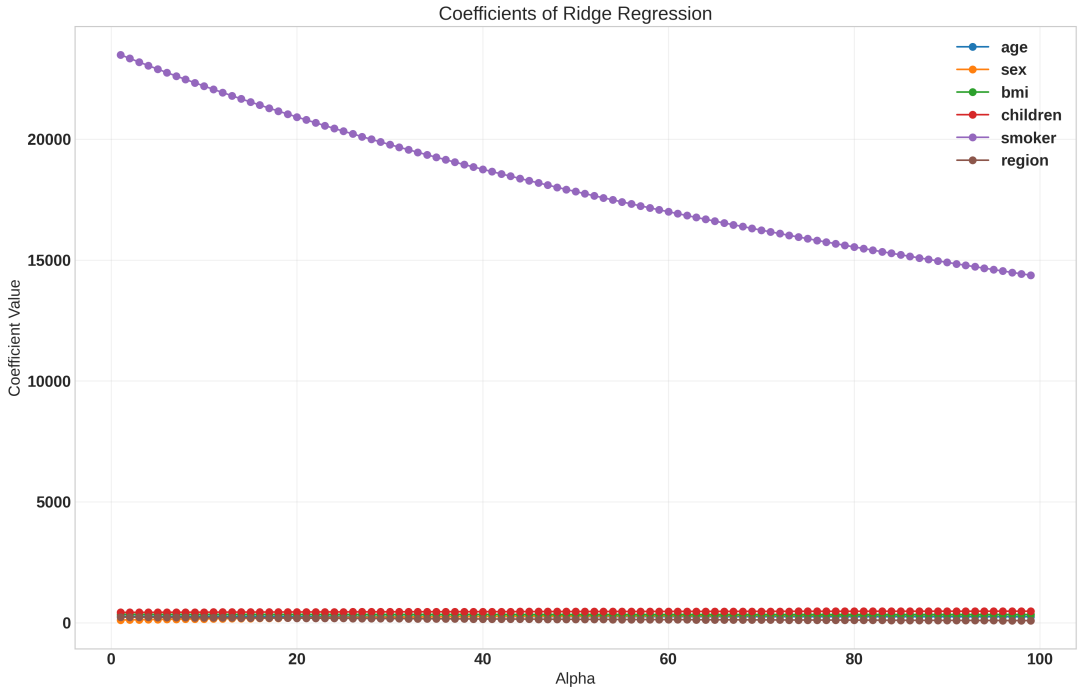

Figure 10: Ridge regression coefficients across different alpha values showing regularization paths

- At $\alpha = 99$: Smoker coefficient reduced to 14,374.39, showing the regularization effect
- Other coefficients showed proportional shrinkage while maintaining relative rankings
- The regularization path plots revealed stable coefficient patterns across different α values

4.4 Model Diagnostics Results

4.4.1 Linearity

The actual vs. predicted values plot showed a strong linear relationship, confirming the appropriateness of linear modeling.

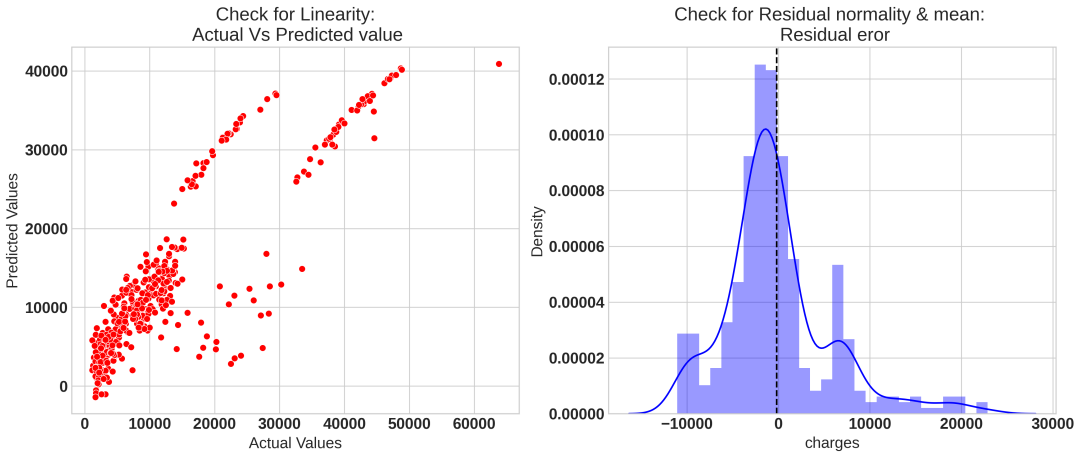

Figure 11: Model diagnostics: actual vs predicted values (left) and residual distribution (right)

4.4.2 Residual Analysis

Residuals showed approximately normal distribution with mean near zero, satisfying model assumptions.

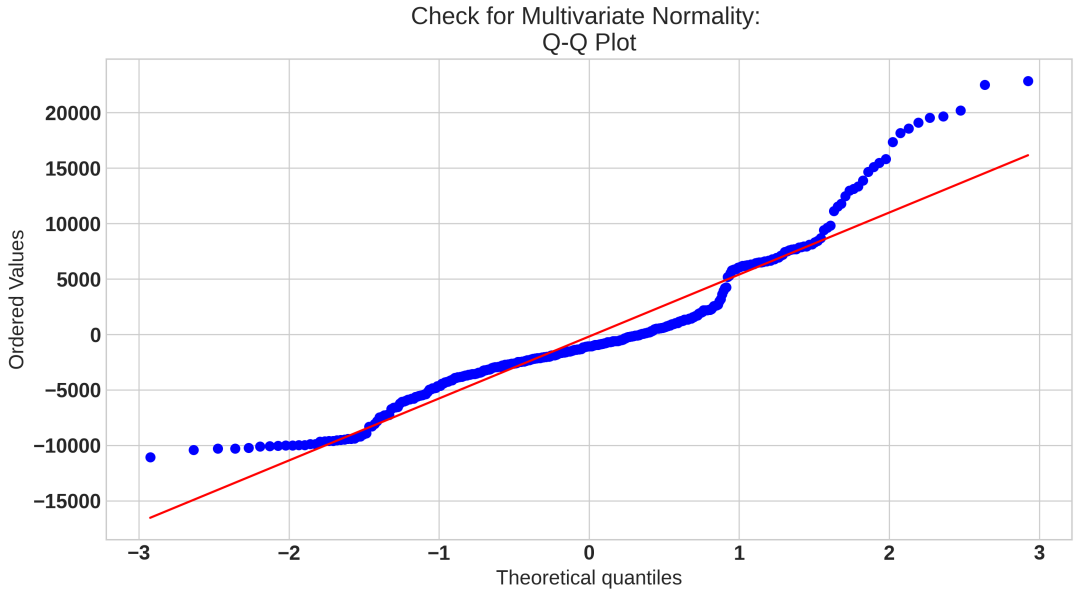

Figure 12: Q-Q plot for normality check

4.4.3 Interaction Effects

The analysis revealed important interaction effects, particularly:

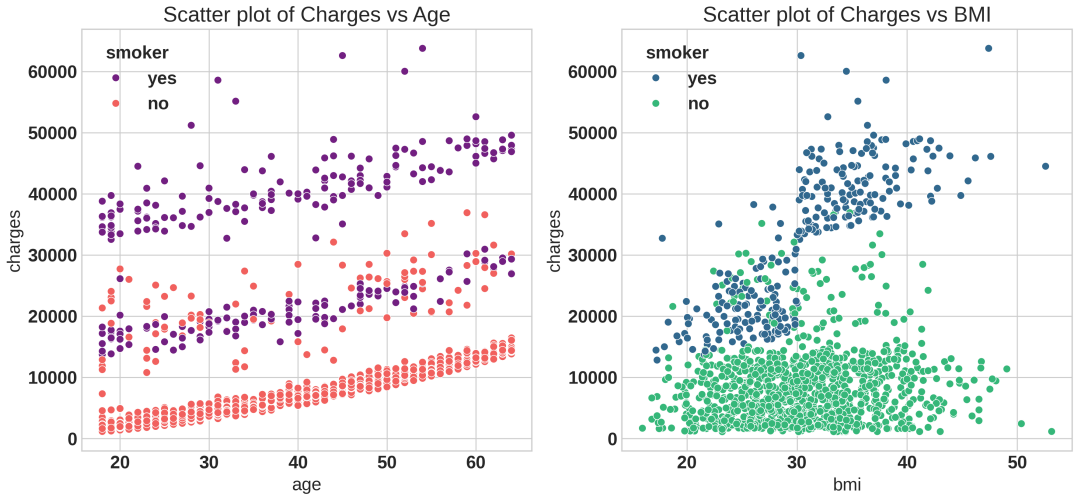

Figure 13: Scatter plots showing interactions between age/BMI and smoking status on charges

- **BMI × Smoking**: The impact of BMI on charges is significantly amplified for smokers
- **Age × Smoking**: Three distinct charge levels were identified based on age and smoking status combinations

These findings suggest that the simple additive model may not fully capture the complexity of insurance charge determination.

5 Code and Data Availability

The complete source code, data analysis scripts, and datasets used in this study are available in the GitHub repository: `https://github.com/pleyva2004/Statistical-Learning-Capstone`. The repository includes Jupyter notebooks for data analysis, visualization scripts, and all generated figures referenced in this report.

6 Conclusion

This comprehensive statistical analysis of insurance charges has yielded several important conclusions:

6.1 Primary Findings

6.1.1 Smoking Status Dominance

The analysis unequivocally demonstrates that smoking status is the most critical factor in determining insurance charges, with an effect magnitude far exceeding all other variables combined. This finding aligns with medical evidence regarding smoking-related health risks and associated healthcare costs.

6.1.2 Model Performance

The linear regression model achieved satisfactory performance with an R-squared of 0.769, indicating that the selected demographic and health-related variables explain approximately 77% of the variance in insurance charges. This level of explanatory power suggests that the model captures the major determinants of insurance pricing.

6.1.3 Feature Hierarchy

A clear hierarchy of feature importance emerged: smoking status » children > BMI > age > region ≈ sex. This ranking provides valuable insights for both insurance companies and policyholders regarding factors that most significantly impact premiums.

6.2 Methodological Insights

6.2.1 Linear Model Adequacy

Despite the complexity of insurance pricing, linear regression proved effective for this dataset, though the presence of interaction effects suggests that more sophisticated models might capture additional variance.

6.2.2 Regularization Effects

Ridge regression demonstrated how regularization can provide more stable coefficient estimates, particularly valuable when deploying models in production environments where robustness is crucial.

6.3 Practical Implications

For Insurance Companies:

- Smoking status should be weighted heavily in risk assessment models
- The identified feature importance hierarchy can inform pricing strategy development
- The model provides a foundation for actuarial analysis and rate setting

For Policyholders:

- Smoking cessation represents the single most impactful action for reducing insurance premiums
- BMI management can provide meaningful cost savings
- Demographic factors (sex, region) have minimal impact on charges

6.4 Limitations and Future Research

Model Limitations:

- Interaction effects are not fully captured in the simple additive model
- The dataset may not include all relevant predictors (e.g., pre-existing conditions, lifestyle factors)

Future Research Directions:

- Investigation of non-linear models and interaction terms
- Incorporation of additional health and lifestyle variables
- Time-series analysis to understand temporal trends in insurance pricing
- Advanced machine learning approaches for improved prediction accuracy

6.5 Final Remarks

This study successfully demonstrates the application of statistical learning techniques to understand insurance charge determination. The findings provide actionable insights for stakeholders in the insurance industry while highlighting the paramount importance of smoking status in healthcare cost prediction. The developed models serve as a foundation for more sophisticated insurance pricing systems and provide a framework for future research in this domain.

The results underscore the critical role of lifestyle factors, particularly smoking, in healthcare costs and insurance pricing, reinforcing the importance of public health initiatives aimed at smoking prevention and cessation. From a statistical modeling perspective, this study illustrates the effectiveness of linear regression for interpretable insurance pricing models while identifying areas for methodological enhancement in future research.