Medical Insurance Cost Prediction and Determinant Analysis

ISyE 6414 Regression Analysis Course Project

Kexin Wu

GTID: 904167062

kwu424@gatech.edu

Georgia Institute of Technology

November 2025

Abstract

This project analyzes medical insurance charges for 1,338 individuals to identify key cost drivers and develop an interpretable predictive model. Exploratory analysis reveals strong right-skewness in charges and clear differences across smoking status, BMI, and age. A baseline multiple linear regression explains about 75% of the variation but violates normality and homoscedasticity assumptions. Introducing a BMI × smoker interaction captures an important nonlinear effect, while Box–Cox analysis supports a log transformation of the response. The final log-linear model, selected using forward selection, includes age, sex, region, number of children, and the BMI–smoking interaction. It provides improved residual behavior and meaningful insights: smoking and high BMI jointly drive the largest cost increases, while age and family size yield predictable moderate effects.

1 Introduction

Health insurance pricing is increasingly data-driven: premiums depend on demographic, behavioral, and regional characteristics that jointly determine expected medical costs. Prior research shows that factors such as age, smoking behavior, and body composition are among the strongest predictors of health expenditures [1]. In practice, insurers must design pricing rules that are both economically sound and explainable to consumers and regulators, making regression analysis a central tool for quantifying these relationships.

In this project, we analyze a publicly available medical insurance dataset to address the following questions:

- Which demographic and lifestyle factors are most strongly associated with annual medical insurance charges?
- How well can a multiple linear regression model predict medical charges for individuals with different risk profiles?
- What model refinements (e.g., interaction terms, transformations) are necessary to satisfy regression assumptions and improve model performance?

The main challenges in this problem include the heavy-tailed distribution of medical costs, potential nonlinear effects (particularly of BMI and smoking), and the presence of influential observations. The analysis therefore proceeds iteratively: starting from an interpretable baseline model, diagnosing assumption violations, and then refining the specification via interaction terms and transformations.

The rest of the report is organized as follows. Section 2 describes the data source and variables. Section 3 introduces the modeling methodology. Section 4 presents the analysis, model diagnostics, and key findings. Section 5 summarizes conclusions, discusses practical implications, and reflects on lessons learned.

2 Data Source and Problem Statement

This section provides an overview of the dataset used in the project and clarifies the objectives of the modeling task. It first summarizes the variables available in the data and their relevance to predicting medical insurance charges. It then highlights initial patterns observed in the dataset, including skewed distributions, group differences, and potential nonlinear effects that motivate further methodological refinement. Finally, it formalizes the main problem statement by outlining the goals of the analysis and explaining why regression modeling is appropriate for addressing the research questions.

2.1 Data Description

The data are obtained from the "Medical Insurance Cost" dataset hosted on Kaggle [2]. The dataset consists of $n = 1338$ observations, each representing an individual policyholder. For each individual, the dataset includes annual medical insurance charges as well as several predictors:

- **age** (numeric): age of the insured individual (18–64).
- **sex** (categorical): gender of the insured (female, male).
- **bmi** (numeric): body mass index (BMI), a proxy for body fat.
- **children** (integer): number of dependents covered by the insurance (0–5).
- **smoker** (categorical): smoking status (no, yes).
- **region** (categorical): region in the U.S. (northeast, northwest, southeast, southwest).
- **charges** (numeric): annual medical insurance billing amount (USD).

Categorical variables are coded as factors, with the following reference levels:

- **sex**: female,
- **smoker**: no,
- **region**: northeast.

This coding allows straightforward interpretation of regression coefficients as differences relative to baseline groups.

2.2 Preliminary Exploration

Basic summaries of the continuous variables indicate that age is roughly uniformly distributed across working ages, BMI exhibits mild right-skewness, and the number of children is concentrated at lower values. Medical charges show pronounced right-skewness with a wide range. These patterns are illustrated in the exploratory plots provided in Appendix A, which further motivate the need for transformation and careful model diagnostics in later sections.

- **Age** is approximately uniformly spread over working ages (mean $\approx$ 39.2, SD $\approx$ 14.0).
- **BMI** is slightly right-skewed (mean $\approx$ 30.7, SD $\approx$ 6.3).
- **Children** is concentrated at 0–2 dependents (mean $\approx$ 1.09).
- **Charges** is highly right-skewed with a wide range (mean $\approx$ \$13,270, SD $\approx$ \$12,111; minimum around \$1,100 and maximum around \$63,770).

Scatter plots of charges versus age and BMI show positive trends with increasing dispersion at higher charge levels. Boxplots of charges by smoking status reveal much higher costs for smokers than non-smokers, indicating that smoking is likely a major driver of cost. Regional and gender differences appear weaker but still noticeable.

No missing values or duplicate records are detected. Therefore, no imputation or deletion of observations is required in preprocessing.

3 Proposed Methodology

This section summarizes the modeling strategy used to analyze medical insurance charges. We begin with a baseline multiple linear regression and evaluate its adequacy using standard diagnostics. Guided by these results, we refine the model by adding an interaction term and applying a Box–Cox transformation to stabilize variance. Finally, forward selection yields a parsimonious and interpretable log-linear model.

3.1 Baseline Multiple Linear Regression

The initial model regresses raw charges on all main effects:

$$Y_i = \beta_0 + \beta_1 X_{i,\text{age}} + \beta_2 X_{i,\text{sex}} + \beta_3 X_{i,\text{bmi}} + \beta_4 X_{i,\text{children}} + \beta_5 X_{i,\text{smoker}} + \beta_6 X_{i,\text{region}} + \varepsilon_i, \tag{1}$$

with coefficients estimated by ordinary least squares,

$$\hat{\boldsymbol{\beta}} = (X^\top X)^{-1} X^\top \mathbf{Y}.$$

3.2 Assumption Checking and Diagnostic Tools

Model adequacy is evaluated using:

- **Residuals vs. fitted:** linearity via $e_i = Y_i - \hat{Y}_i$.
- **Scale–Location:** homoscedasticity via $r_i = e_i / (\hat{\sigma}\sqrt{1 - h_{ii}})$.
- **Normal Q–Q:** normality via $r_{(i)} \approx \Phi^{-1}((i - 0.5)/n)$.
- **Leverage and Cook's distance:**

$$h_{ii} = X_i (X^\top X)^{-1} X_i^\top, \quad D_i = \frac{e_i^2}{p\hat{\sigma}^2} \frac{h_{ii}}{(1 - h_{ii})^2}.$$

- **GVIF:** collinearity via $\text{GVIF}^{1/(2d)} = 1/\sqrt{1 - R^2}$.

Diagnostics reveal heteroscedasticity, right-skewness, and influential observations, motivating model refinement.

3.3 Model Refinement via Interaction Terms

Because BMI effects differ for smokers and non-smokers, we add the interaction:

$$Y_i = \beta_0 + \cdots + \beta_7(X_{i,\text{bmi}} \cdot X_{i,\text{smoker}}) + \varepsilon_i. \tag{2}$$

The interaction substantially improves fit and reduces curvature in residual plots.

3.4 Response Transformation Using Box–Cox Analysis

Residual diagnostics still show heteroscedasticity and heavy tails, motivating a Box–Cox transformation:

$$Y_i^{(\lambda)} = \begin{cases} \dfrac{Y_i^\lambda - 1}{\lambda}, & \lambda \neq 0, \\ \log(Y_i), & \lambda = 0. \end{cases} \tag{3}$$

Maximizing the profile likelihood yields $\hat{\lambda} \approx 0.05$, supporting the log transformation. This conclusion is also consistent with standard practice in health econometrics, where log-linear specifications are commonly recommended for positively skewed cost data [3]. The resulting model is:

$$\log(Y_i) = \beta_0 + \beta_1 X_{i,\text{age}} + \beta_2 X_{i,\text{sex}} + \beta_3 X_{i,\text{bmi}} + \beta_4 X_{i,\text{children}} + \beta_5 X_{i,\text{smoker}} + \beta_6 X_{i,\text{region}} + \beta_7(X_{i,\text{bmi}} \cdot X_{i,\text{smoker}}) + \varepsilon_i. \tag{4}$$

On the log scale, variance stabilizes, Q–Q deviations diminish, and influence decreases.

3.5 Variable Selection for Model Parsimony

Forward selection using partial F-tests,

$$F = \frac{(\text{RSS}_{\text{reduced}} - \text{RSS}_{\text{full}})/(p_{\text{full}} - p_{\text{reduced}})}{\text{RSS}_{\text{full}}/(n - p_{\text{full}})}, \tag{5}$$

identifies age, sex, children, region, and the BMI × smoker interaction as key predictors. The BMI main effect remains negligible for non-smokers. The resulting model is both parsimonious and statistically well-behaved.

4 Analysis and Results

This section presents the empirical findings obtained from applying the proposed modeling strategy to the medical insurance dataset. We first summarize the performance of the baseline regression model and examine the extent to which its assumptions are violated. We then evaluate the impact of incorporating interaction terms and applying the Box–Cox transformation, followed by an assessment of the final log-linear model selected through forward selection. The results highlight both the statistical improvements achieved through model refinement and the substantive insights gained regarding the determinants of medical insurance charges.

4.1 Baseline Model on Raw Charges

We begin by fitting a multiple linear regression model using the raw insurance charges as the response and including all main effects as predictors. The model is implemented in R (version 4.3.1) using the `lm()` function from the `stats` package [4], which computes the ordinary least squares estimator

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y.$$

The estimated regression equation is:

$$\begin{aligned}\hat{Y}_{\text{charge}} =& -11938.5 + 256.9\,X_{\text{age}} - 131.3\,X_{\text{sex}} + 339.2\,X_{\text{bmi}} + 475.5\,X_{\text{children}} \\ &+ 23848.5\,X_{\text{smokeryes}} - 353.0\,X_{\text{regionnorthwest}} - 1035.0\,X_{\text{regionsoutheast}} - 960.0\,X_{\text{regionsouthwest}}.\end{aligned}$$

The model explains a substantial proportion of the variability in charges. The residual standard error (RSE) is

$$\text{RSE} = 6062,$$

and the coefficient of determination and the adjusted coefficient of determination are

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2} \approx 0.751,$$

$$R^2_{\text{adj}} = 1 - (1 - R^2)\frac{n-1}{n-p-1} \approx 0.749,$$

indicating that approximately 75% of the variation in medical charges is explained by the predictors. The overall model is highly significant, with

$$F = \frac{(\text{TSS} - \text{RSS})\,/p}{\text{RSS}/(n-p-1)} = 500.8, \qquad p < 2.2 \times 10^{-16}.$$

Interpretation of Coefficients. The estimates indicate that several predictors have strong and statistically significant effects on annual medical charges:

- **Age:** Each additional year increases expected charges by approximately \$256.9, reflecting higher utilization with age.
- **BMI:** A one-unit BMI increase raises expected charges by approximately \$339.2, indicating higher costs associated with elevated body mass.
- **Children:** Each additional dependent increases charges by \$475.5, consistent with expanded coverage and family-based medical expenses.
- **Smoking status:** Smokers incur dramatically higher expenses, with an expected increase of \$23,848.5 relative to non-smokers. This is by far the largest single effect in the model.
- **Sex and Region:** The main effect of sex is not statistically significant ($p = 0.693$), suggesting limited cost differences between males and females after adjusting for other factors. Regional effects are modest, with the Southeast and Southwest regions showing slightly lower expected charges relative to the Northeast.

4.2 Diagnostic Observations.

Residual diagnostics for the baseline model, generated using a combination of the `plot()` functions from the `stats` package and custom visualization tools from `ggplot2`, along with the `influencePlot()` and `vif()` functions from the `car` and `tidyverse` package [4, 5, 6, 7], reveal several violations of classical regression assumptions (see Appendix B for full plots):

- **Residuals vs. fitted values:** Noticeable curvature and increasing spread indicate violations of linearity and homoscedasticity.
- **Normal Q–Q plot:** Strong right-tail deviations reflect heavy-tailed residuals driven by extreme-charge observations.
- **Scale–Location plot:** Increasing residual variance across fitted values further confirms heteroscedasticity.
- **Residuals vs. leverage:** Although no point exceeds Cook's $D = 1$, several moderately influential observations contribute to model instability.

Overall, these diagnostics indicate that modeling raw charges on the original scale does not satisfy linearity, normality, or constant-variance assumptions, motivating refinement in later sections. By contrast, GVIF values near 1, computed using the `car::vif()` function, confirm that multicollinearity is not a concern.

4.3 Interaction Model Results

To capture the substantial behavioral differences between smokers and non-smokers, an interaction term between BMI and smoking status is added to the model. The interaction model is estimated using the `lm()` function from the `stats` package [4], which allows the slope of BMI to vary by smoking category rather than assuming a uniform linear effect across all individuals.

The interaction coefficient is large and highly statistically significant:

$$\hat{\beta}_{\text{bmi:smokeryes}} = 1443.10, \qquad p < 2 \times 10^{-16},$$

indicating that the effect of BMI on medical charges differs dramatically between smokers and non-smokers. Specifically:

- For **non-smokers**, the BMI slope is small and statistically insignificant, suggesting limited marginal effect of BMI on expected charges.
- For **smokers**, the BMI slope increases by roughly \$1443 per BMI unit, producing a much steeper cost gradient.

This pattern is illustrated using an interaction plot generated with `ggplot2` [6] (see Appendix C), which shows an almost flat BMI–cost relationship for non-smokers but a strong positive trend for smokers.

The inclusion of the interaction term substantially improves model performance:

$$R^2_{\text{interaction}} = 0.8409, \qquad R^2_{\text{adj}} = 0.8398,$$

representing a significant increase from the baseline model ($R^2 \approx 0.751$). The residual standard error decreases from 6062 to 4846, indicating greater explained variability and improved predictive accuracy.

Although the interaction reduces curvature in the residuals–fitted plot and produces a more linear trend, heteroscedasticity and heavy-tailed residuals remain present. Thus, while the interaction captures a key behavioral relationship, transforming the response is still necessary to fully satisfy linear model assumptions. This motivates the Box–Cox transformation discussed in Section 4.4.

4.4 Log-Transformed Final Model

Based on the Box–Cox analysis implemented using the `boxcox()` function from the `MASS` package [8], the representative fitted model on the log scale is:

$$\begin{aligned}\log(\widehat{Y}) = 7.34 &+ 0.035\,X_{\text{age}} - 0.087\,X_{\text{sex}} + 0.0034\,X_{\text{bmi}} + 0.103\,X_{\text{children}} + 0.156\,X_{\text{smokeryes}} \\ &- 0.071\,X_{\text{regionnorthwest}} - 0.163\,X_{\text{regionsoutheast}} - 0.138\,X_{\text{regionsouthwest}} + 0.046\,X_{\text{bmi}} \times X_{\text{smokeryes}}.\end{aligned} \tag{6}$$

Diagnostic plots for the log-transformed model, generated using the `plot()` functions from the `stats` package and the leverage/influence tools available in the `car` package [4, 5], show substantial improvement compared with the raw-charge model (see Appendix D). The Residuals vs. fitted plot displays no visible curvature and much more uniform variance, indicating that the log transformation effectively stabilizes heteroscedasticity. The Normal Q–Q plot exhibits only mild upper-tail deviations, suggesting that residual normality now holds reasonably well. The Scale–Location plot confirms homoscedasticity, and influential observations are greatly reduced in the residuals–leverage diagram. Overall, the transformed model satisfies the linearity, normality, and constant-variance assumptions to a far greater degree than the untransformed baseline model.

4.5 Variable Selection via Forward Selection

Forward selection on the log-transformed response, implemented using iterative `lm()` model fitting and partial F-tests via the `anova()` function from the `stats` package [4], confirms that a parsimonious and interpretable model can be constructed while retaining strong explanatory power. Starting from the null model, predictors were added sequentially only when their entry p-value fell below 0.05. The selection sequence shows that the BMI × smoker interaction enters first with overwhelming significance, followed by age, number of children, region, and sex. Smoking status and the main BMI effect never enter after adjusting for the interaction term, indicating that BMI has minimal marginal effect for non-smokers.

The resulting final model includes:

$$\{\text{age}, \ \text{sex}, \ \text{children}, \ \text{region}, \ \text{bmi} \times \text{smoker}\}.$$

Model performance metrics produced using `stats::summary()` demonstrate that the final log-linear model fits well, with a residual standard error of approximately 0.41, an R^2 near 0.80, and a large AIC reduction from roughly 27,115 (raw-scale model) to about 1,546. This reflects substantial gains in both parsimony and explanatory power after transformation and variable selection.

Overall, forward selection highlights the central role of the BMI × smoker interaction in explaining cost heterogeneity, while age, sex, number of children, and geographic region provide essential demographic adjustments. The negligible BMI main effect for non-smokers aligns with earlier diagnostic evidence and the interaction plot (see Appendix C).

The final fitted model for the log-transformed response takes the following form:

$$\begin{aligned}\log(\widehat{Y}) = 7.372 &+ 0.035\, X_{\text{age}} - 0.088\, X_{\text{sex}} + 0.103\, X_{\text{children}} \\ &- 0.072\, X_{\text{regionnorthwest}} - 0.163\, X_{\text{regionsoutheast}} - 0.139\, X_{\text{regionsouthwest}} \\ &+ 0.0023\, X_{\text{bmi}} \times X_{\text{smokerno}} + 0.0528\, X_{\text{bmi}} \times X_{\text{smokeryes}}. \end{aligned} \tag{7}$$

5 Conclusion

5.1 Summary

This project shows that smoking status is the dominant driver of medical insurance charges, with high-BMI smokers forming the most expensive subgroup. Age and number of children have steady positive effects, while sex and region contribute moderate adjustments. A Box–Cox analysis supports log-transforming charges, which greatly improves variance stability and diagnostics. The final log-linear model with the BMI × smoker interaction offers a clear, interpretable structure that aligns well with linear regression assumptions.

These findings suggest that premiums should account for the elevated risk of smokers, particularly those with high BMI, while age, family size, and region inform baseline rate adjustments. The log-scale formulation also facilitates percentage-based interpretation for non-technical audiences.

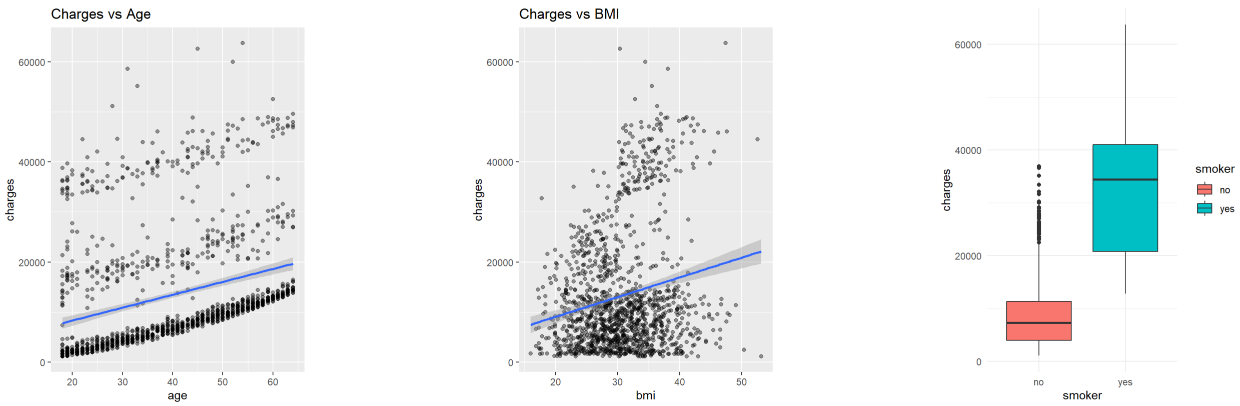
5.2 Limitations and Future Work

The current analysis relies on a streamlined set of predictors and includes only a single interaction term within a linear modeling framework. Future work could expand the specification by incorporating a broader set of demographic, behavioral, or contextual covariates, as well as alternative data sources. Methodologically, exploring flexible nonlinear structures—such as generalized additive models, tree-based ensembles, or modern machine learning approaches—may reveal additional patterns in the data and provide valuable benchmarks for evaluating the stability, interpretability, and predictive performance of the log-linear formulation.

Lessons Learned

Key lessons from the project and ISyE 6414 include: the importance of diagnostics, the effectiveness of transformations and interactions, the value of parsimony in applied modeling, and the need for clear communication of statistical results. The project as a whole strengthened my understanding of how theoretical regression concepts inform practical data analysis.

A Appendix: Exploratory Data Analysis Figures

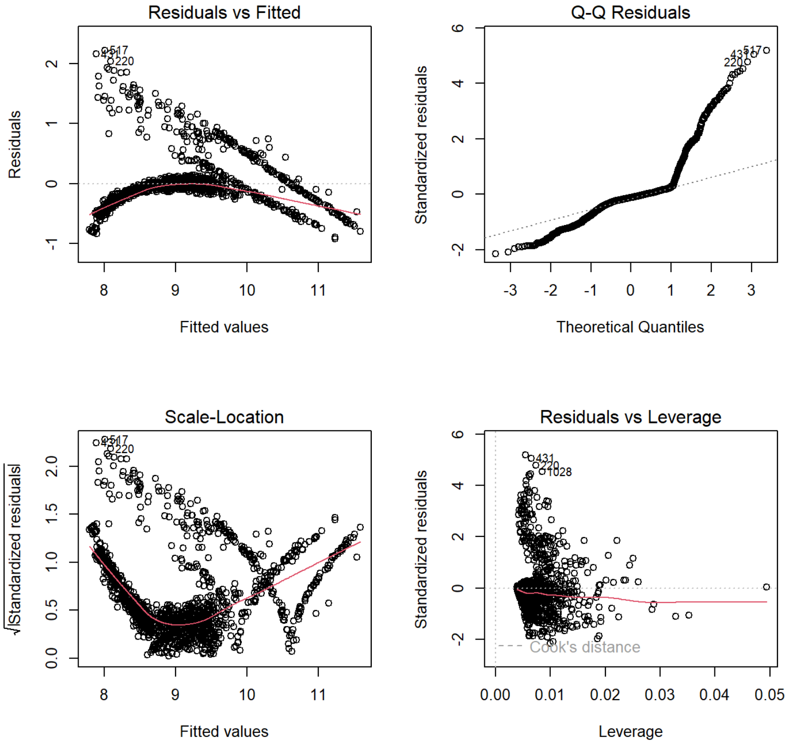


(a) Scatter plot for age

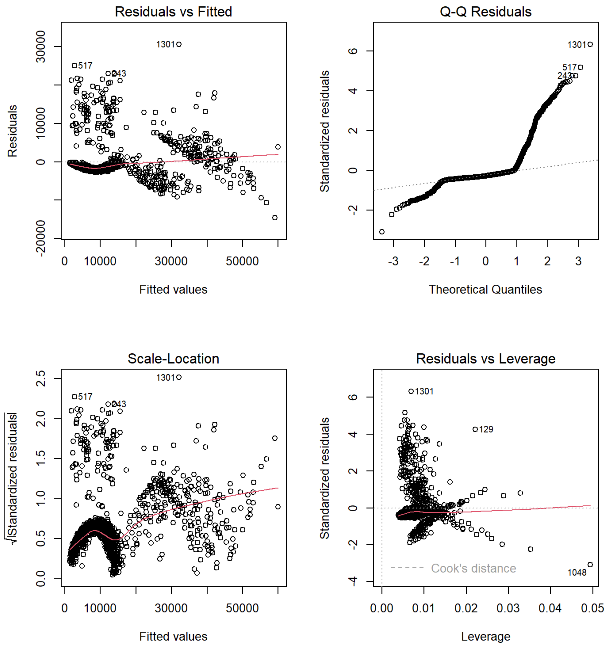
(b) Scatter plot for bmi

(c) Box plot for smoker

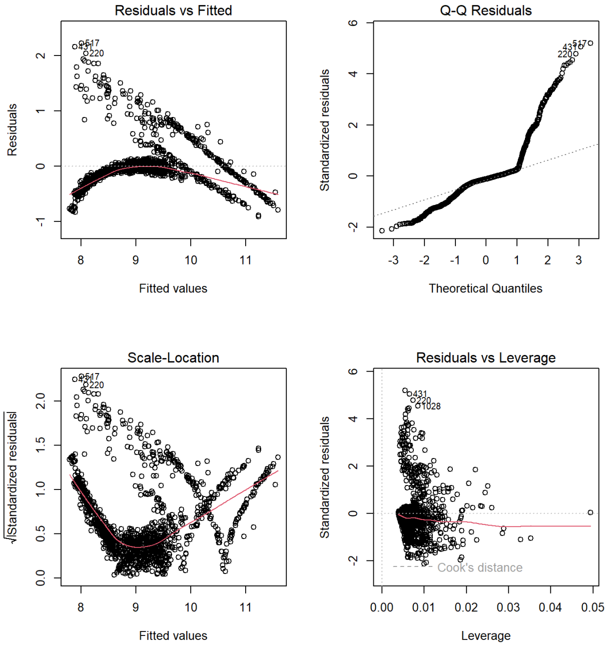
B Appendix: Baseline Model Diagnostic Plots



C Appendix: Interaction Plot: BMI × Smoker



D Appendix: log-transformed Plot



E Appendix: R Code Snippets

```
library(tidyverse)
library(car)
library(ggplot2)
library(MASS)

insurance <- read.csv("insurance.csv")
str(insurance)
summary(insurance)

insurance$sex     <- factor(insurance$sex,     levels = c("female", "male"))
insurance$smoker  <- factor(insurance$smoker,  levels = c("no", "yes"))
insurance$region  <- factor(insurance$region)

# Exploratory summaries for key continuous variables
summary(insurance[, c("age", "bmi", "children", "charges")])

ggplot(insurance, aes(age, charges)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  labs(title = "Charges vs. Age")

ggplot(insurance, aes(bmi, charges)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm") +
  labs(title = "Charges vs. BMI")

ggplot(insurance, aes(smoker, charges, fill = smoker)) +
  geom_boxplot() +
  theme_minimal()

# Fit full multiple linear regression with all main effects
fit_full <- lm(charges ~ age + sex + bmi + children + smoker + region, data = insurance)
summary(fit_full)

# Baseline diagnostic plots (base R)
```

```
par(mfrow = c(2, 2))
plot(fit_full)

# Influence diagnostics
influencePlot(fit_full)

# Multicollinearity check
vif(fit_full)

# Remove influential observations identified in diagnostics
insurance_clean <- insurance[-c(1301, 578, 544), ]
fit_clean <- lm(charges ˜ age + sex + bmi + children + smoker + region, data = insurance_clean)
summary(fit_clean)

# Compare coefficients before/after influential point removal
summary(fit_full)$coefficients
summary(fit_clean)$coefficients

# Diagnostic plots for cleaned model
plot(fit_clean)

# Add interaction term: BMI × smoker
fit_interact <- lm(charges ˜ age + sex + bmi * smoker + children + region, data = insurance)
summary(fit_interact)

# Compare full model vs. interaction model
anova(fit_full, fit_interact)
AIC(fit_full, fit_interact)

# Visualization of BMI × smoker interaction
ggplot(insurance, aes(x = bmi, y = charges, color = smoker)) +
  geom_point(alpha = 0.4) +
  geom_smooth(method = "lm", se = FALSE) +
  labs(title = "Interaction: BMI × Smoker",
       x = "BMI", y = "Insurance Charges (USD)")
```

```
# Diagnostic plots for interaction model
plot(fit_interact)

# Box{Cox transformation analysis
bc <- boxcox(fit_interact, lambda = seq(-2, 2, 0.1))

# Fit log-transformed model using Box{Cox recommendation
fit_log <- lm(log(charges) ~ age + sex + bmi * smoker + children + region, data = insurance)
summary(fit_log)

# Compare models using AIC
AIC(fit_full, fit_log)

# Diagnostic plots for log-transformed model
plot(fit_log)



# Forward selection
vars <- c("age", "sex", "bmi", "children", "smoker", "region", "bmi:smoker")

current_model <- lm(log(charges) ~ 1, data = insurance)  # null model
remaining <- vars

while (length(remaining) > 0) {

  pvals <- c()

  for (v in remaining) {
    # Construct candidate model formula
    f <- as.formula(
      paste("log(charges) ~",
            paste(c(attr(terms(current_model), "term.labels"), v),
                  collapse = " + "))
    )

    m <- lm(f, data = insurance)
    # Extract p-value from partial F-test
    p <- anova(current_model, m)$‘Pr(>F)‘[2]

    if (!is.na(p)) {
      pvals[v] <- p
    }
  }

  if (length(pvals) == 0) break

  best <- names(which.min(pvals))
  best_p <- min(pvals)

  cat("Candidate:", best, "p =", best_p, "\n")

  if (!is.na(best_p) && best_p < 0.05) {
    message(paste("Adding:", best))
    current_model <- update(current_model, paste(". ˜ . +", best))
    remaining <- setdiff(remaining, best)
  } else {
    message("No further predictors meet the entry criterion; stopping.")
    break
  }
}

# Final selected model
summary(current_model)
```

References

[1] Ellis, R. P., and McGuire, T. G. Predictability and Predictiveness in Health Care Spending. *Journal of Health Economics*, 26(1): 25–48, 2007.

[2] Kaggle. Medical Cost Personal Dataset. Available at: https://www.kaggle.com/datasets/mirichoi0218/insurance.

[3] Jones, A. M., and Rice, N. *Applied Health Economics* (4th ed.). Routledge, 2020.

[4] R Core Team. *stats: The R Statistics Package*. Available at: https://www.rdocumentation.org/packages/stats.

[5] Fox, J., and Weisberg, S. *An R Companion to Applied Regression* (3rd ed.). Sage, 2019.

[6] Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. Springer, 2016.

[7] Wickham, H., Averick, M., Bryan, J., *et al.* *Welcome to the tidyverse*. Journal of Open Source Software, 4(43), 1686, 2019.

[8] Venables, W. N., and Ripley, B. D. *Modern Applied Statistics with S* (4th ed.). Springer, 2002.