Exploring the Factors Influencing Health Insurance Charges

with Multiple Linear Regression Analysis

Yanting Hu, Shuxin Zhang, Liuyixin Shao, Yi Su, Dongfeng Li

1 Introduction

Health insurance plays a crucial role in safeguarding individuals and families against the financial burdens associated with healthcare expenses. Especially after Covid-19, "health insurance has become one of the most prominent areas of research." [1] As medical costs continue to rise and healthcare needs evolve, health insurance charges also increase gradually, after a 4% increase in health insurance in 2023, "American families will face another 4% increase in the cost of private health insurance." [2] People can't help asking after reading these exposures — What are the key factors that affect the health insurance charges? Can I predict next year's charges? How good/credible are these predictive models?

This report sets out on an exploratory voyage in an effort to solve the mystery surrounding health insurance charges and then give prediction models on the charges. The rest of the report is divided into data, the initial model and its analysis, the alternative model and their analysis, conclusion, and discussion. By delving into these topics, we aim to unravel the factors that influence insurance charges and to fit the regression models capable of accurately predicting health insurance charges.

2 Data

2.1 About The Data

The dataset is collected by Kaggle. [3] The dataset contains 1338 observations across 7 columns of customer personal information: age, sex(female or male), bmi(body mass index), children(the number of children), smoker(yes or no), regions(customers' residential area), and charges (insurance charge bill). Specifically, there are 3 character columns: `sex`, `smoker`, `region`, and 4 numeric columns: `age`, `children`, `bmi`, and `charges`, providing a comprehensive foundation for analysis and exploration of health insurance-related factors.

2.2 Initial Data Prepocessing and Spliting

To address the issue of having multiple intercepts due to the presence of categorical variables in regression analysis, we changed `sex`, `smoker`, and `region` into factors in R. To facilitate the development and evaluation of predictive models, we divided the dataset into 80% training and 20% testing (validation) subsets.

[1](Health Insurance Cost Prediction Using Machine Learning, n.d.)
[2](ValuePenguin, 2023)
[3](Health Insurance Dataset, 2020)

2.3 Data Processing for the Alternative Solution

We realized that splitting the dataset based on smoking status may alter the distribution of the response variable: `charges`. Therefore, we transformed the dataset separately for each smoke status group to ensure our modeling was appropriate for these differences and then produced a reliable analysis. After the box-cox analysis, we chose to do a square root transformation to the charges for the smoker dataset and a log transformation to the charges for the non-smoker dataset.

2.3.1 Smoker

The potential issue regarding the dataset is the variable representing the number of children. To be more specific, the dataset tends to be a higher frequency of individuals with fewer children (0, 1, or 2) compared to those with a larger number of children (3, 4, or 5). This can lead to imbalanced data distribution and potentially affect the accuracy of statistical analyses, such as regression modeling.

In order to address this issue and ensure a more balanced representation of the data, we transformed the numerical variable representing the number of children into a categorical variable with distinct groups: "no child," and "have child." This categorization helps to address the potential skewness in the data distribution and improves the interpretability of regression analysis results.

2.3.2 Non-Smoker

Since the non-smoker uses the bootstrap method which will be introduced below, it is robust and is resistant to the imbalanced data distribution. Therefore we still consider the children's number to be numerical.

3 Model for the Initial Thought: One Full Regression Model

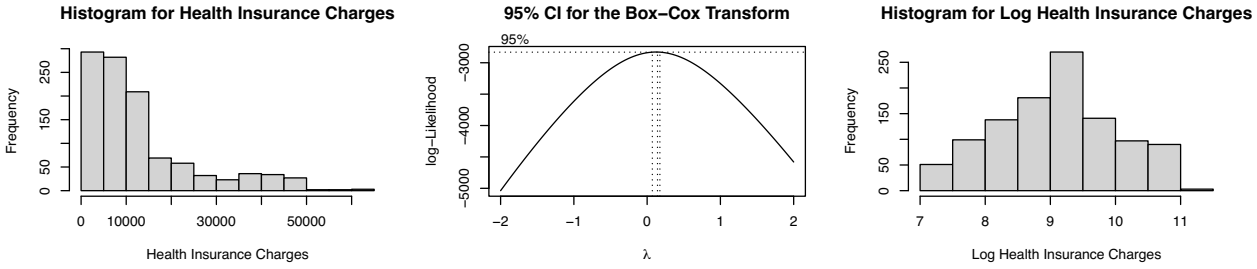

Figure 1: Log Transformation on Health Insurance Charges

Our initial thought for this analysis was to find the best-fitted linear regression model for the entire dataset. Before fitting the full regression model, we noticed the distribution of Health Insurance Charges is very right skewed (Figure 1, first plot). In order to fit the full regression model, we need a Box-Cox transformation. Here, we took $\lambda = 0$ and applied a log transformation to the charges (Figure 1, second plot). We can see that the distribution of Health Insurance Charges after log transformation is normal.

3.1 Assumption Checks and Multicollinearity Checks

In the following analysis, we performed Assumption Checks and Multicollinearity Checks several times. To streamline our analysis, we used the list of conditions under which each assumption holds and referred to them as Assumption 1, 2, 3, and (not) having serious Multicollinearity issues.

- Check Assumption 1: linearity ($E[\epsilon_i] = 0$), using Residual vs. Fitted plot. Assumption holds if the points are randomly scattered around the horizontal line at zero.
- Check Assumption 2: normality ($\epsilon_i \sim N(0, \sigma^2)$), using Q-Q Residuals plot. Assumption holds if the points are staying on the line. (That is, the ordered residuals correspond linearly with quantiles of a standard normal distribution.)
- Check Assumption 3: homoscedasticity ($Var[\epsilon_i] = \sigma^2$), using Standard Residuals vs. Fitted plot since we transformed the response variable - `charges`. Assumption holds if the spread of points is constant across all values of the fitted values.
- Check Multicollinearity: enabling the precision of the coefficients, using Residuals vs. Leverage plot. Check holds if there isn't any point pass the 0.5 cook distance curve.

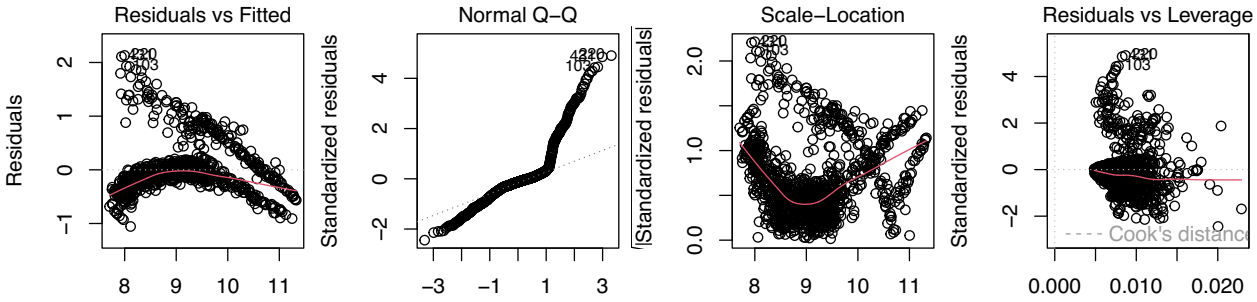

Figure 2: Assumption Check For Full Model

Here, we fitted a full model as `log_charges ~ age + sex + bmi + children + smoker + region` to see how personal information is related to the charges.

According to Figure 2, Assumption 1 is not satisfied, with clear trends in points. Assumption 2 is not satisfied as the last part of the quantile deviates a lot. Assumption 3 is not satisfied and there is no serious Multicollinearity issue. Therefore, we see that all of the assumptions are violated even though all the explanatory variables are normally distributed.

Since this model does not conform to the assumptions of linear regression, if we continue to use this model for our analysis, our results will be inaccurate. To ensure the reliability of the models that we made, we need to do some analysis.

3.2 Solution - Further Spliting the Dataset into Smokers and Nonsmokers

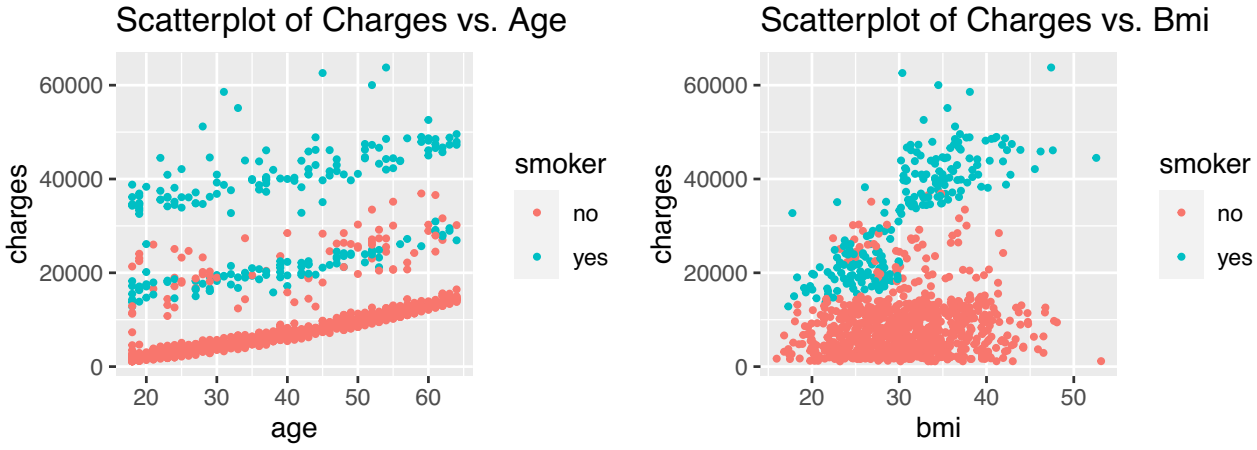

Figure 3: Scatter Plot of Age VS Charges and BMI VS Charges

According to Figure 3, we can see that charges have many outliers. we tried to separate these outliers, but there is no variable that separates outliers completely. There are far more factors that affect the cost of health insurance than the number of variables we have. We noticed that smoking status could have a significant impact on the cost of insurance since smoking can cause many diseases and those diseases can cause the rise of the health insurance charges. Plus, this is the only variable that could separate some of the outliers. Therefore, we decided to divide the dataset into two groups, smokers and nonsmokers, to make a more accurate analysis.

4 Models for the Alternative Thought: Split The Dataset Based on Smoke Status

4.1 Model for Smoker

4.1.1 Model Selection

Table 1: Model Selection Criteria

Mallows_Cp	AIC	BIC
1161.372	2139.16	2156.036

The Mallow's Cp statistic, Akaike Information Criterion (AIC), and Bayesian Information Criterion (BIC) are used for the appropriate model selection. While Mallow's Cp assesses the trade-off between model fit and complexity in linear regression, AIC and BIC extend to various modeling contexts, penalizing complexity differently based on information theory and sample size, respectively. Here, we chose to use these three model selection criteria for all the possible models (including the interactions of all variables) that we can build from the dataset. The surprising result is that all the optimal model selection criteria ultimately point to the same model with parameters `bmi`, `age`, and `sex` with no interaction term being selected. The corresponding Mallow's Cp statistic, AIC, and BIC for this selected model are displayed in Table 1.

4.1.2 Analysis of the Selected Linear Model

For the selected model with predictors `bmi`, `age`, and `sex`:

Here, the intercept is not worthy of explanation because it suggests that when both bmi and age are zero, the square root of charges for females is expected to be approximately 42.8299, and for males is expected to be 41.0109. However, no one will have a bmi or age of zero.

According to the summary, keeping other factors the same, for every one-unit increase in bmi, the estimate of the average square root of charges increases by 8.1623. For every increase in age, the estimate of the average square root of charges increases by 1.5053. Based on the p-values, the significant predictors are `bmi` and `age`, which implies that the effect of age and bmi on the square root of charges is evident from the data.

The coefficient for `sexmale` is -1.8190, indicating that, keeping other factors the same, being male is associated with a decrease in the square root of charges of -1.8190 compared to being female. However, this effect is not statistically significant, as indicated by the p-value, which suggests that sex may not have a strong influence on the square root of charges within this model.

The F-statistic 220.5 with a significantly small p-value suggests that the overall model is statistically significant. The residual standard error is 33.75. The R-squared value is 0.7573, meaning that the model explains approximately 75.73% of the variability in charges, which is quite high. The adjusted R-squared value is 0.7539, which is also quite high and close to the R-squared value, suggesting that the model is not being overly penalized for having unnecessary predictors.

```
Call:
lm(formula = charges ~ bmi + age + sex, data = datasmoker)

Residuals:
    Min      1Q  Median      3Q     Max
-83.393 -24.618  -5.279  23.034 141.815

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  42.8299    12.2847   3.486 0.000595 ***
bmi           8.1623     0.3581  22.796  < 2e-16 ***
age           1.5053     0.1676   8.984  < 2e-16 ***
sexmale      -1.8190     4.6456  -0.392 0.695783
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.75 on 212 degrees of freedom
Multiple R-squared:  0.7573,    Adjusted R-squared:  0.7539
F-statistic: 220.5 on 3 and 212 DF,  p-value: < 2.2e-16
```

4.1.3 Assumption Checks and Multicollinearity Checks

According to Figure 4, Assumption 1 is satisfied, though the points are not perfectly scattered. Assumption 2 is roughly satisfied with some exception of the points on the edges. Assumption 3 is satisfied and there is no serious Multicollinearity issue.

We can see that the standard error for all the parameters in the selected model is smaller. Thus, there should not be a serious multicollinearity problem for this model. According to the Variance Inflation Factor (VIF) that we have calculated for the model, we can see that all of the VIF are smaller than 5 (Table 2), meaning that there isn't serious multicollinearity for our parameters.

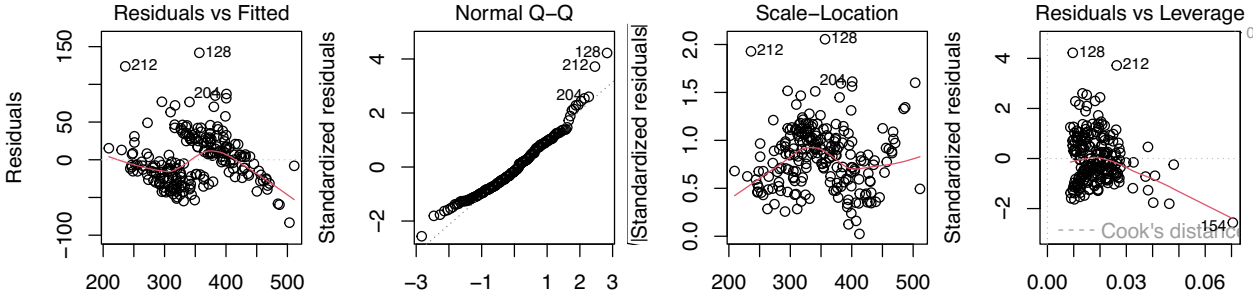

Figure 4: Assumption Check For Selected Model

Table 2: VIF for Multicollinearity Analysis

bmi	age	sexmale
1.028692	1.013127	1.015946

4.1.4 Prediction Results

Figure 5 shows the relationship between the actual charge and the charge predicted by the selected linear regression model, represented by the red dots. The data points are distributed around the

dashed line, which represents the perfect prediction line, where the actual value is equal to the predicted value. The linear regression model seems to perform reasonably well, as many of the points are clustered around the perfect fit line. However, there are some deviations from this line, indicating some degree of prediction error. Overall, the model generally performs well despite some discrepancies between predicted and actual charges.

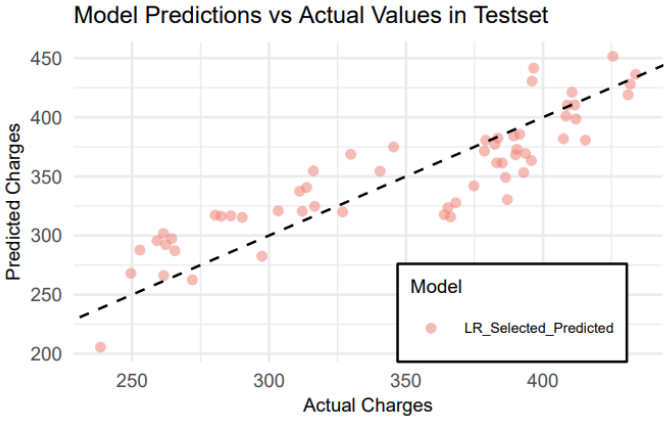

Figure 5: Comparison of Model Predictions with Actual Values

4.2 Non-Smoker

4.2.1 Assumption Checks and Outliers Filtering

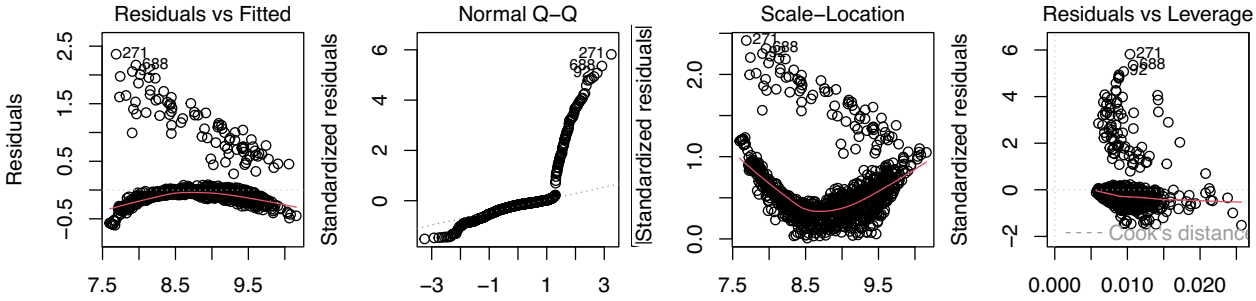

Figure 6: Assumption Check For Initial Model

When checking the assumptions by fitting the model `log_charges ~ age + sex + bmi + children + region`, as we can see from Figure 6, no assumptions are satisfied. The data of non-smokers shows a clear pattern of residuals that is dense on the bottom of the residual plots and also some outliers that lie above it. These outliers cannot be distinguished by any of the variables in our database. We believed that the cause of the presence of these outliers may be related to the presence of some disease for those persons, but this part of the information is missing in our dataset. Therefore, we would consider extracting this pattern out of the data and consider those floating data to be outliers in order to make a linear regression model. Consider the residuals that

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i.$$

If we compare the residuals to $\hat{\epsilon}_i$ in the regression model:

$$\hat{Y}_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \Longleftrightarrow \quad \epsilon_i = Y_i - \beta_0 - \beta_1 X_i.$$

Essentially, each e_i mimics the role of $\hat{\epsilon}_i$ when the fitted coefficients $\hat{\beta}_0, \hat{\beta}_1$ are close to β_0, β_1. Therefore, we would consider to use $\hat{\epsilon}_i$ of the linear model to capture the distance of the data point

to the assumed model and therefore only keep data that are not outliers. In this case we choose 85% quantile of the empirical residuals that $\hat{Y} = \{\hat{Y}_i : \hat{\epsilon}_i < Q_{0.85\hat{\epsilon}}\}$.

Figure 7 shows the extraction of the data point and the outliers:

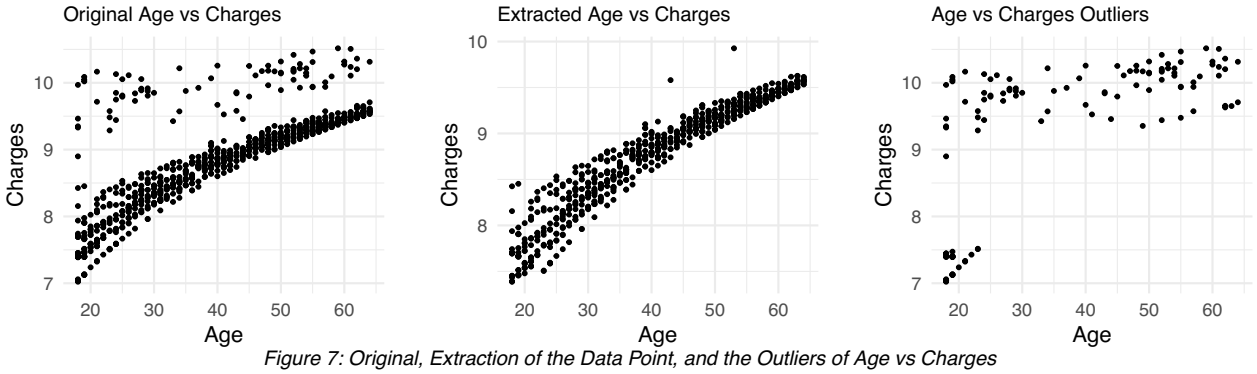

Figure 7: Original, Extraction of the Data Point, and the Outliers of Age vs Charges

```
Call:
lm(formula = charges ~ age + sex + bmi + children + region, data =
data_with_small_residuals)

Residuals:
     Min       1Q   Median       3Q      Max
-0.26794 -0.04187  0.02345  0.05786  0.41018

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.9540758  0.0188836 368.260  < 2e-16 ***
age              0.0432018  0.0002415 178.897  < 2e-16 ***
sexmale         -0.0840325  0.0063880 -13.155  < 2e-16 ***
bmi              0.0002688  0.0005611   0.479 0.632009
children         0.1202463  0.0026797  44.873  < 2e-16 ***
regionnorthwest -0.0328302  0.0089478  -3.669 0.000261 ***
regionsoutheast -0.0985473  0.0094504 -10.428  < 2e-16 ***
regionsouthwest -0.1006027  0.0090966 -11.059  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.08524 on 718 degrees of freedom
Multiple R-squared:  0.9793,    Adjusted R-squared:  0.9791
F-statistic:  4860 on 7 and 718 DF,  p-value: < 2.2e-16
```

Taking the 85% quantile as a threshold cannot yield those exact outliers in the dataset, but that's what we want because we're not aiming to fit a perfect model in the training dataset but in the unseen testing dataset. Therefore, we would consider that we have successfully deleted the outliers that are irrelevant to our model. The linear regression model is presented above.

We then checked the assumptions by fitting the model `log_charges ~ age + sex + bmi + children + region` again. As we can see from the table, all the factors except `bmi` are significant. The R-squared value is 0.9793 and the adjusted R-squared value is 0.9791, which are quite high.

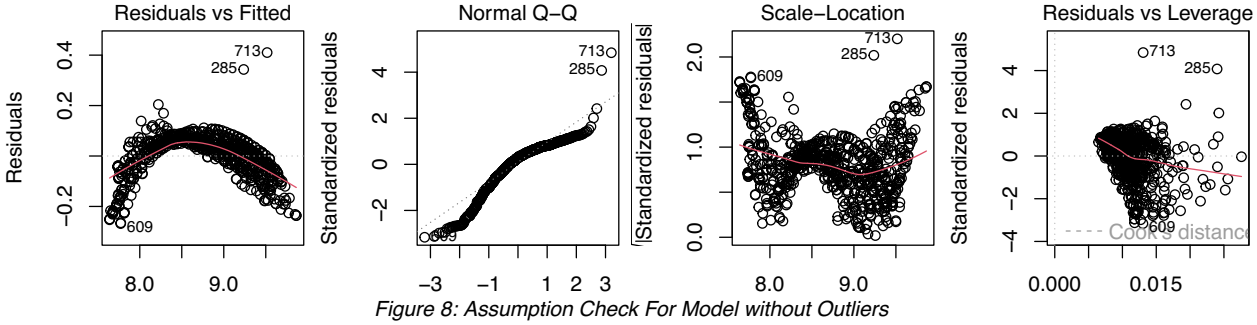

Figure 8: Assumption Check For Model without Outliers

According to Figure 8, Assumption 1 is not satisfied with a trend. Assumption 2 is roughly

satisfied. Assumption 3 is not satisfied with points with non-constant variance and there is no serious Multicollinearity issue.

4.2.2 Residual Bootstrap

To improve our model performance and better satisfy our model assumption, we would consider to use residual bootstrap [4].

The residual bootstrap first generates IID $\hat{\epsilon}_1^*, \ldots, \hat{\epsilon}_n^*$ such that for each $\hat{\epsilon}_i^*$,

$$P(\hat{\epsilon}_i^* = \hat{\epsilon}_i) = \frac{1}{n}, \quad \forall n = 1, \ldots, n.$$

And then generates a new bootstrap sample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$ via $X_i^* = X_i, \quad Y_i^* = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\epsilon}_i^*$.

Namely, we fixed the covariate X_i but generate a new value of Y_i using the fitted regression function and the 'noise' from sampling the residuals with replacement.

Now we would repeat the process B times, we would have:

$$\begin{aligned}
&(X_1^{*(1)}, Y_1^{*(1)}), \ldots, (X_n^{*(1)}, Y_n^{*(1)}) \\
&(X_1^{*(2)}, Y_1^{*(2)}), \ldots, (X_n^{*(2)}, Y_n^{*(2)}) \\
&\vdots \\
&(X_1^{*(B)}, Y_1^{*(B)}), \ldots, (X_n^{*(B)}, Y_n^{*(B)}).
\end{aligned}$$

For each bootstrap sample, say $(X_1^{*(\ell)}, Y_1^{*(\ell)}), \ldots, (X_n^{*(\ell)}, Y_n^{*(\ell)})$, we fit the linear regression, leading to a bootstrap estimate of the fitted coefficients $\hat{\beta}_0^{*(\ell)}, \hat{\beta}_1^{*(\ell)}$. Thus, the B bootstrap samples leads to B sets of fitted coefficients. We then estimate the variance and construct our confidence interval by:

$$\widehat{\mathrm{Var}}(\hat{\beta}_0) = \frac{1}{B}\sum_{\ell=1}^{B}(\hat{\beta}_0^{*(\ell)} - \bar{\hat{\beta}}_0)^2, \qquad C.I.(\hat{\beta}_0) = \hat{\beta}_0 \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_0)}, \qquad \bar{\hat{\beta}}_0 = \frac{1}{B}\sum_{\ell=1}^{B}\hat{\beta}_0^{*(\ell)},$$

$$\widehat{\mathrm{Var}}(\hat{\beta}_1) = \frac{1}{B}\sum_{\ell=1}^{B}(\hat{\beta}_1^{*(\ell)} - \bar{\hat{\beta}}_1)^2, \qquad C.I.(\hat{\beta}_1) = \hat{\beta}_1 \pm z_{1-\alpha/2} \cdot \sqrt{\widehat{\mathrm{Var}}(\hat{\beta}_1)}, \qquad \bar{\hat{\beta}}_1 = \frac{1}{B}\sum_{\ell=1}^{B}\hat{\beta}_1^{*(\ell)}.$$

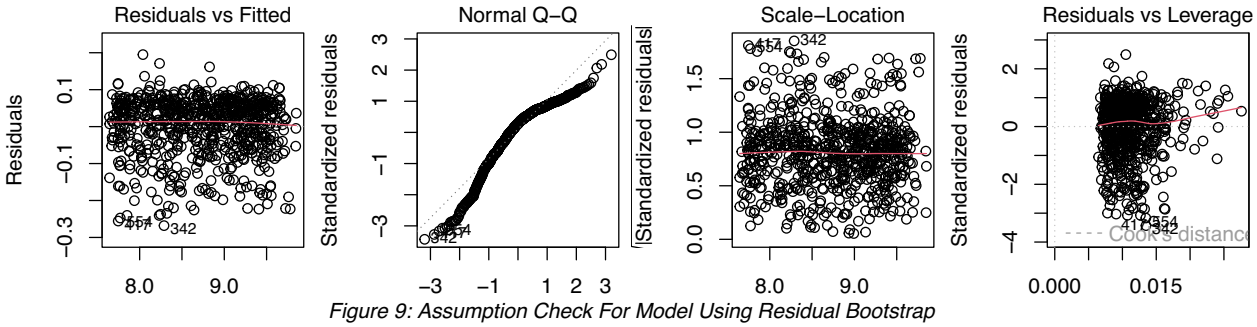

Figure 9: Assumption Check For Model Using Residual Bootstrap

According to Figure 9, Assumption 1 satisfied. Assumption 2 may have some violation but we treat

[4](@Chen2017)

it as satisfied. Assumption 3 is satisfied and there is no serious Multicollinearity issue.

4.2.3 Prediction Results

Figure 10 shows confidence intervals for model coefficients from a regression analysis for non-smoker data by using the bootstrap method. In the model, the predictors of `age` and `children` have positive coefficients and their confidence intervals do not include zero, suggesting they are statistically significant predictors in the model. While predictors `regionsoutheast` and `regionsouthwest` have negative coefficients and their confidence intervals do not include zero, suggesting they are statistically significant predictors in the model. The coefficient of `bmi` and `regionnorthwest` has a small confidence interval and it contains 0, which indicates it is not a significant predictor in the model.

The result of the statistically significant predictors is slightly inconsistent with the linear model using `lm()` directly. This might be because the standardized residuals are still not constant across fitted values and therefore might include some violations. Bootstrap methods can overcome this situation by resampling from residuals.

For predictions, we would consider in the test set that we still use the 85% quantile of the empirical residuals from the initial regression model in the training set as the threshold to judge if the given data is an outlier or not. That is to say, $\hat{Y} = \{\hat{Y}_i : \hat{\epsilon}_i < Q_{0.85\hat{\epsilon}_{train}}\}$.

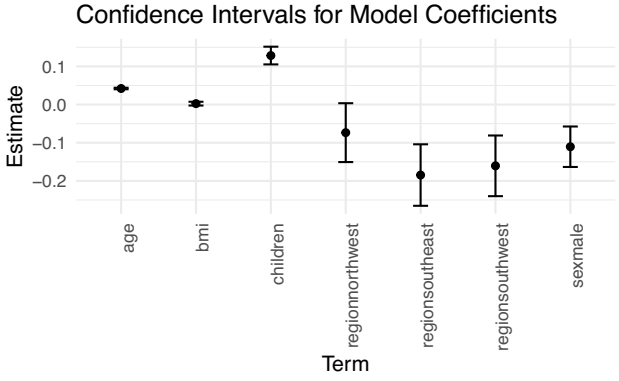

Figure 10: Confidence Intervals for Model Coefficients

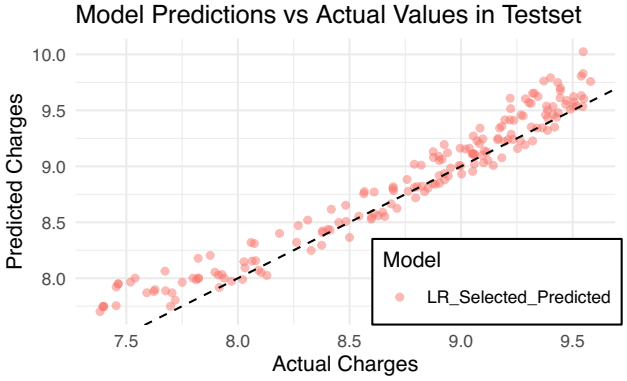

Figure 11: Comparison of Model Predictions with Actual Values

According to Figure 11, the linear regression model appears to predict the charges well, as indicated by the close clustering of the data points around the line of perfect prediction, where the actual value is equal to the predicted value. The data points are represented by red dots, where the x-axis shows the actual charges and the y-axis shows the charges predicted by the model. Overall, the prediction aligns well with the expected outcomes, with no apparent outlier present.

5 Conclusion

From the above analysis, for the smokers, the result showed that `bmi` and `age` are strong predictors of smokers' insurance charges. The prediction model shows a linear relationship between the actual and predicted charges and it performs well for the testing dataset. For nonsmokers, our improved method, which included removing outliers and residual scaling, found that `age`, `children`, and `region` are statistically significant predictors, while `bmi` is not a statistically significant predictor. Using the 85% quantile of the empirical residuals in the model, it appears to predict the charges well. We were unable to predict insurance costs for nonsmoker outliers. We need more information

to accomplish this task. These results show that health insurance costs are complicated and that it's important for prediction models to take into account personal traits.

6 Discussion

This is truly a hard dataset for linear regression analysis because: 1) most of the explanatory variables in the dataset are categorical or ordinal; 2) The dataset contains a lot of outliers that cannot be separated easily by its existing explanatory variables, suggesting that we might need more information for the person (especially the person's medical history) to do a more accurate prediction.

Based on what we learned from this research, we should consider more potentially effective factors on health insurance charges, use more advanced predictive modeling methods, and keep improving the models to better reflect how healthcare use and prices change over time. The clear effect that smoking has on insurance rates shows how important it is to have custom models that take living factors into account. As healthcare needs and insurance policies change, it will be important to use machine learning algorithms and keep models up to date with new data to make predictions more accurate and make sure they stay relevant to current trends.

7 References

- Bhatia, K., Gill, S. S., Kamboj, N., Kumar, M., & Bhatia, R. K. (2022). Health Insurance Cost Prediction using Machine Learning. In *2022 3rd International Conference for Emerging Technology (INCET)* (pp. 1-5). Belgaum, India. doi: 10.1109/INCET54531.2022.9824201
- Chen, Y. (n.d.). Lecture 6: Bootstrap for regression. https://faculty.washington.edu/yenchic/17Sp_403/Lec6-bootstrap_reg.pdf
- Health Insurance dataset. (2020, December 18). https://www.kaggle.com/datasets/shivadumnawar/health-insurance-dataset/data
- ValuePenguin. (2023, December 19). Private health insurance premiums reach a record high of $7008/year in 2024. *PR Newswire: press release distribution, targeting, monitoring and marketing.* https://www.prnewswire.com/news-releases/private-health-insurance-premiums-reach-a-record-high-of-7008year-in-2024-302019327.html

8 Work Assignment

8.1 First Author

- Section 1, 2, 3: Yanting Hu
- Section 4.1: Liuyixin Shao and Yi Su
- Section 4.2: Dongfeng Li and Shuxin Zhang
- Section 5, 6: All

We worked together to review, refine, and organize the final report, streamlining it from 28 pages in the original draft to 10 pages. Everyone's contribution to the project was the same.