ILLINOIS INSTITUTE OF TECHNOLOGY

Report

Predicting Health Insurance Costs Using Demographic and Lifestyle Factors: A Regression Analysis

Team:

Harshali Gaikwad hgaikwad@hawk.iit.edu A20538035

Instructor

Prof. Kiah Ong

Course

Fall 2024 Regression (CS-564)

CONTENTS

CHAPTER 1

ABSTRACT

This project analyzes a dataset of health insurance costs to build predictive models and gain insights into factors influencing insurance charges. The dataset, sourced from Kaggle, includes variables such as age, sex, BMI, number of children, smoking status, and geographic region, along with the dependent variable individual medical insurance charges. The primary objective is to use regression techniques to model and predict insurance costs based on these attributes, identifying key predictors and quantifying their impact.

Exploratory data analysis (EDA) is performed to understand the data structure, detect potential outliers, and assess relationships between features and the target variable. Linear regression models are initially applied to predict continuous insurance charges.

CHAPTER 2

INTRODUCTION

Health insurance costs are a key concern for both individuals and providers, influencing financial stability and access to care. Predicting these costs based on personal and lifestyle factors can uncover what drives them and help people make informed decisions. Using the Insurance Charges dataset from Kaggle, this project explores how variables like age, gender, BMI, smoking status, and region impact annual insurance charges.

To start, we'll build a linear regression model to estimate costs and analyze which factors carry the most weight. This will provide actionable insights into how choices like quitting smoking or managing BMI might reduce premiums. By understanding these cost drivers, insurers can price policies more accurately, and individuals can make smarter health and financial choices.

This project is a step toward demystifying healthcare costs, aiming to empower people to take control of their well-being while fostering transparency in insurance pricing.

CHAPTER 3

PROBLEM STATEMENT

The rising cost of healthcare is a major concern for both individuals and insurance providers. Accurately predicting individual insurance charges based on personal and lifestyle factors is essential for insurers to set fair and sustainable premiums and for consumers to understand the cost implications of various factors, such as smoking and obesity. This project seeks to build a predictive model for health insurance charges using a dataset containing features such as age, sex, BMI, number of children, smoking status, and geographic region.

The primary goals of this project are as follows:

1. Predictive Modeling: Develop a regression model that accurately predicts insurance charges based on the provided features. This model will help quantify the impact of individual factors on insurance costs.
2. Feature Impact Analysis: Identify and interpret key predictors affecting insurance charges, such as the influence of smoking status or BMI on cost variations.
3. Multicollinearity and Autocorrelation Checks: Investigate multicollinearity among predictors to ensure that variables are not excessively correlated, which can skew model performance and interpretability. Additionally, check for autocorrelation in residuals to verify the assumptions of independence in regression models, improving model reliability.
4. Model Evaluation: Evaluate and compare model performance, using metrics such as mean absolute error (MAE) and root mean square error (RMSE), to determine the accuracy and robustness of the developed model.

By accomplishing these objectives, this project aims to provide valuable insights into the cost structure of health insurance and to create a model that can inform pricing strategies and health risk assessments for insurance providers.

CHAPTER 4

DATA COLLECTION AND DATA SOURCE

The dataset used in this project is sourced from Kaggle and was provided by the user "mirichoi0218." It is titled "Insurance" and consists of 1,338 records, each representing a unique individual's health-related data and corresponding insurance charges. The dataset is openly available for analysis and machine learning projects, making it widely used for exploring healthcare cost prediction and related statistical modeling tasks.

Dataset Features
Following is the snapshot of the medical insurance data.

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

The dataset contains the following columns:

1. age: Age of the individual (numeric).
2. sex: Gender of the individual (categorical: male or female).
3. bmi: Body Mass Index (BMI) of the individual, calculated as weight in kilograms divided by height in meters squared (numeric).
4. children: Number of dependents (children) covered by the insurance (numeric).
5. smoker: Smoking status of the individual (categorical: yes or no).
6. region: Geographic region within the United States where the individual resides (categorical: northeast, northwest, southeast, southwest).
7. charges: Medical insurance costs billed to the individual (dependent variable, numeric).

The dataset is relatively clean, with no missing values, making it ready for exploratory data analysis and modeling. The simplicity and structure of the dataset also allow for easy transformation into different machine learning and statistical tasks, such as regression for cost prediction.

CHAPTER 5

PROPOSED METHODOLOGY

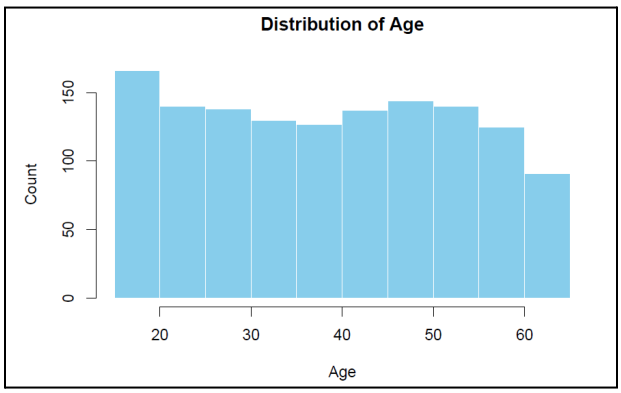
5.1 Perform Exploratory Data analysis

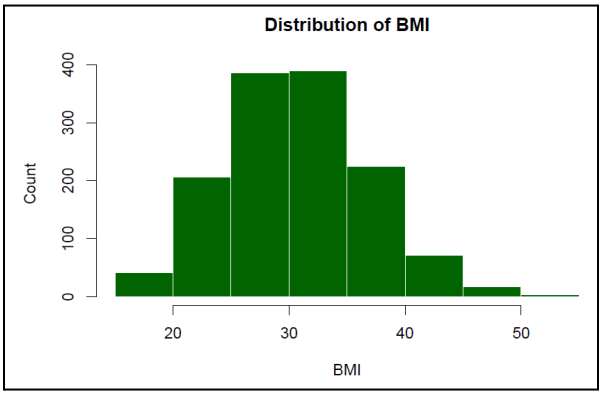
1. Distribution of Age: The age histogram shows a roughly even spread of participants across different age groups, with ages ranging from 20 to 70 and a slight peak around the younger range.

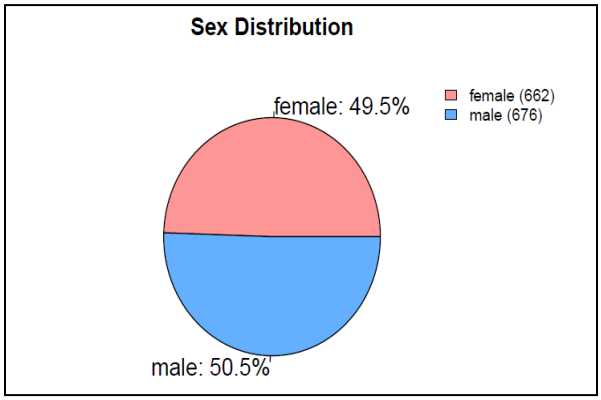
2. Distribution of BMI: The BMI histogram reveals a higher concentration of individuals with BMI values around 30, indicating most participants are in the overweight range.

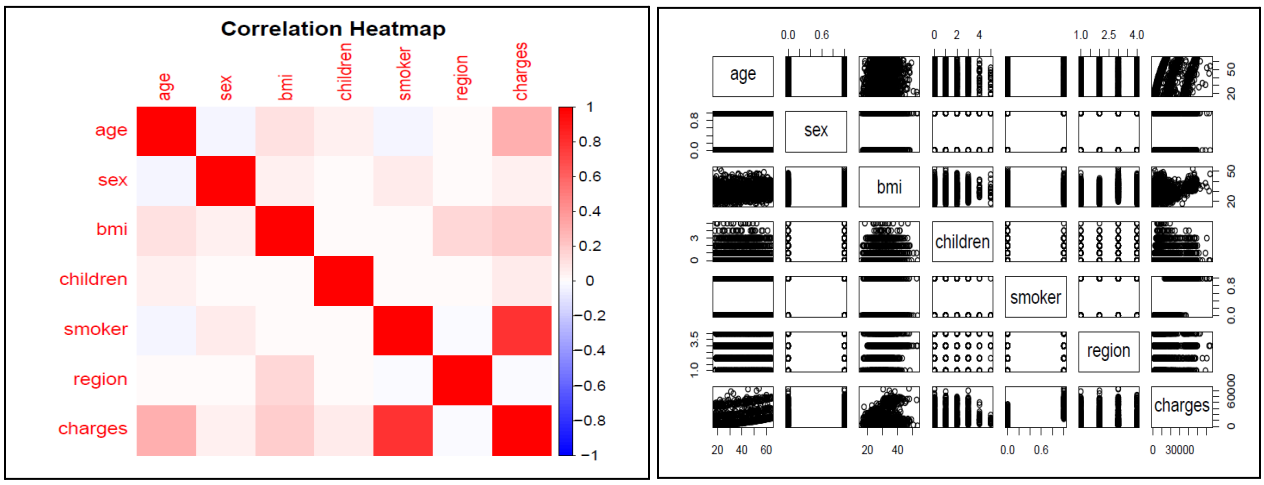
3. Sex Distribution: The pie chart shows a nearly equal split between female (49.5%) and male (50.5%) participants, with 662 females and 676 males.

```
hist(medical_insurance_data$age,
     main = "Distribution of Age",
     xlab = "Age",
     ylab = "Count",
     col = "skyblue",
     border = "white")
```









From the above correlation heatmap we can see the strength and direction of relationships between variables. Also `charges` has a strong positive correlation with `smoker` and moderate positive correlations with `age` and `bmi`, indicating these variables may significantly impact insurance costs. On the other hand variables, like `sex`, `children`, and `region`, have weaker correlations with `charges`, suggesting a smaller effect.

Also looking at the pairs plot the diagonal shows variable names (age, sex, bmi, children, smoker, region, charges), while the off-diagonal plots show correlations and distributions between pairs of variables. The strongest visible relationship appears to be between charges and smoking status, suggesting smokers incur higher medical charges

5.2 Model Fitting and Model Interpretation

With linear regression, we can validate assumptions such as

1. Linearity,
2. Normality of residuals, and
3. Homoscedasticity using various diagnostic plots.

This data setup allows for such analysis, ensuring linear regression can be evaluated for its suitability.

In this case, the dependent variable is 'charges' and other 6 variables are predictors out of which 2 are categorical variables 'sex' and 'region'. So will develop a regression model which will predict insurance charges based on selected features.

```
## Call:
## lm(formula = charges ~ ., data = medical_insurance_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11343  -2807  -1017   1408  29752
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11461.81     983.00 -11.660  < 2e-16 ***
## age            257.29      11.89  21.647  < 2e-16 ***
## sex           -131.11     332.81  -0.394 0.693681
## bmi            332.57      27.72  11.997  < 2e-16 ***
## children       479.37     137.64   3.483 0.000513 ***
## smoker       23820.43     411.84  57.839  < 2e-16 ***
## region        -353.64     151.93  -2.328 0.020077 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1331 degrees of freedom
## Multiple R-squared:  0.7507, Adjusted R-squared:  0.7496
## F-statistic: 668.1 on 6 and 1331 DF,  p-value: < 2.2e-16
```

From the model 1 summary we can say that the p-value for sex is much higher than the significance level ($\alpha = 0.05$) making it an insignificant attribute. However, multiple R^2 is 0.7507 means the model explains 75.07% of variance in charges and high F-statistic (668.1) with p < 2.2e-16 indicates the model is significantly a good model. From the summary of the model we find that the attribute ‘sex’ is not significant so we will try a model without this attribute.

```
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = medical_insurance_data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11404  -2805   -992   1400  29694
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11513.40     973.93 -11.822  < 2e-16 ***
## age            257.41      11.88  21.670  < 2e-16 ***
## bmi            332.04      27.68  11.995  < 2e-16 ***
## children       478.44     137.58   3.478 0.000522 ***
## smoker       23808.21     410.54  57.992  < 2e-16 ***
## region        -353.45     151.88  -2.327 0.020104 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6058 on 1332 degrees of freedom
## Multiple R-squared:  0.7507, Adjusted R-squared:  0.7498
## F-statistic: 802.2 on 5 and 1332 DF,  p-value: < 2.2e-16
```

From the model 2 summary, we find that both model 1 and model 2 have nearly identical R-squared values (0.7507), indicating removing 'sex' didn't impact the model. Also coefficients for all remaining variables stayed almost the same like for age: 257.29 vs 257.41, smoker: 23820.43 vs 23808.21. However, the F-statistic slightly improved in model 2 (802.2 vs 668.1), with the same significant p-value, indicating a more efficient model. All variables remain significant at the same levels, with 'smoker' still having the largest impact on charges. But the residual standard error is slightly better in model 2 (6058 vs 6060). Hence, we can say the ‘sex’ term does not have an impact on the model performance even if we do not include it. So let's check the model performance with an interaction term.

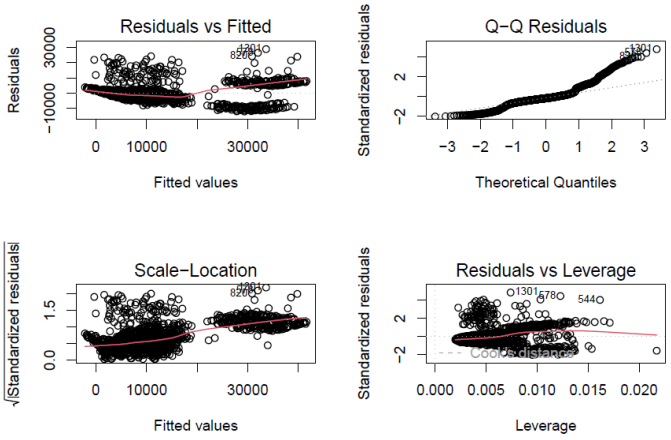
Checking model's performance with an interaction term.

```
## Call:
## lm(formula = charges ~ age + smoker + bmi + sex + sex * smoker +
##     children + region, data = medical_insurance_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12171.9  -2866.5   -981.1   1348.8  28952.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11078.61     989.69 -11.194  < 2e-16 ***
## age            257.30      11.85  21.705  < 2e-16 ***
## smoker       22501.16     620.79  36.246  < 2e-16 ***
## bmi            328.66      27.68  11.872  < 2e-16 ***
## sex           -600.07     370.89  -1.618 0.105912
## children       468.46     137.34   3.411 0.000667 ***
## region        -363.08     151.56  -2.396 0.016731 *
## smoker:sex    2350.48     829.28   2.834 0.004661 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6044 on 1330 degrees of freedom
## Multiple R-squared:  0.7522, Adjusted R-squared:  0.7509
## F-statistic: 576.9 on 7 and 1330 DF,  p-value: < 2.2e-16
```

1. Looking at all three models, **Model 3** emerges as the best choice with the highest multiple R^2 (0.7522) and adjusted R^2 (0.7509), along with the lowest residual standard error (6044). This model consists of an interaction term between sex and smoker status, providing a better understanding of how these variables jointly affect insurance charges.
2. The model's high F-statistic (576.9) with a p-value < 2.2e-16 indicates strong overall significance. In terms of individual regressors, age, smoker status, and BMI are highly significant (p < 2e-16) as each year of age adds $257.30 to charges, being a smoker increases base charges by $22,501.16, and each unit increase in BMI adds $328.66.
3. The number of children (p = 0.000667) and region (p = 0.016731) also show significant effects, with each additional child increasing charges by $468.46 and regional variations affecting charges by $363.08.
4. While sex alone isn't significant (p = 0.105912), its interaction with smoking status is (p = 0.004661), indicating that male smokers incur an additional $2,350.48 in charges compared to female smokers.
5. Because of the interaction term is the significance of model Model 3's superiority over Models 1 and 2, as it captures complexity in the relationship between sex and smoking that the simpler models missed. The high multiple R^2 value indicates that the model explains 75.22% of the variance in insurance charges, making it a strong model when comparing against Models 1 and 2.
6. The adjusted R-squared values across the three models is 74.96% in Model 1. However, in Model 2, which removed the non-significant `sex` variable, slightly improved to 74.98%. Model 3, which added an interaction term between `sex` and `smoker`, reached 75.09%, the highest adjusted R-squared, shows progression of the interaction term which captures additional variance in `charges` without overfitting, making Model 3 the most effective for predicting insurance costs.

5.3 Model Assumptions and Potential Issues

Checking for Homoscedasticity by plotting the model.



These four diagnostic plots help assess the linear regression model's assumptions:

1. Residuals vs Fitted: Shows a distinct pattern and clustering, indicating potential non-linearity and heteroscedasticity in the model.
2. Q-Q Plot: The points deviate from the diagonal line, especially at the tails, suggesting the residuals aren't perfectly normally distributed.
3. Scale-Location: The spread of standardized residuals isn't constant across fitted values, confirming heteroscedasticity (unequal variance).
4. Residuals vs Leverage: Shows some influential points (high leverage) and potential outliers, with points 130, 578, and 544 labeled as notable observations that might be affecting the model fit.

Overall, these plots suggest some violation of linear regression assumptions, particularly homoscedasticity and normality, which might indicate that model transformations or robust regression methods could be beneficial.

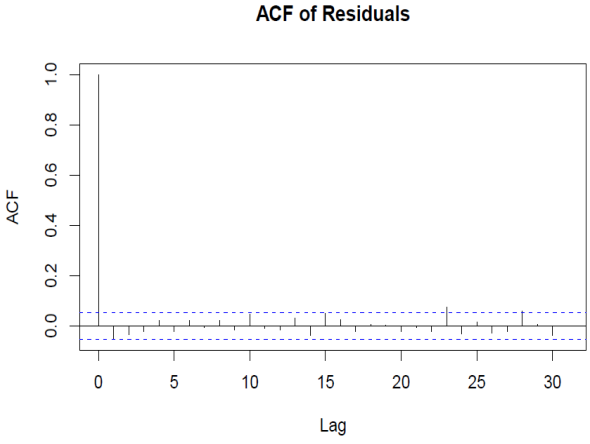
Checking for Autocorrelation Test

```
dwtest <- dwtest(model3)
dwtest
```

```
##
##  Durbin-Watson test
##
## data:  model3
## DW = 2.0963, p-value = 0.9609
## alternative hypothesis: true autocorrelation is greater than 0
```

```
residuals <- residuals(model3)

acf(residuals, main="ACF of Residuals")
```



The diagram shows that there is no autocorrelation. Moreover, the dwtest tells us that there is no significant autocorrelation between the residuals of the model.

Since there is no autocorrelation, let's check for multicollinearity using the VIF test

```
vif(model3)

##        age     smoker        bmi        sex   children     region smoker:sex
##   1.015394   2.298850   1.043203   1.259561   1.003271   1.026463   2.637854
```

Looking at the VIF results show no concern for multicollinearity in Model 3. All VIF values are well below the common threshold of 5 or 10:

1. Most variables have VIF values very close to 1 (1.01-1.26), indicating minimal correlation with other predictors.
2. Smoker (2.29) and smoker:sex interaction (2.63) have slightly higher VIF values, which is expected for interaction terms, but still well within acceptable limits.
3. These results suggest that multicollinearity is not a concern in this model, and the coefficient estimates are reliable.

Since there's no issue with autocorrelation or multicollinearity in the model, we only need to address heteroscedasticity (unequal variance of residuals), which can affect the reliability of our estimates. To remediate this, we can apply transformations, such as log, square root, or Box-Cox transformations, to the dependent variable or specific predictors. These transformations help stabilize variance, promoting homoscedasticity and improving the model's accuracy.

CHAPTER 6

ANALYSIS AND RESULTS

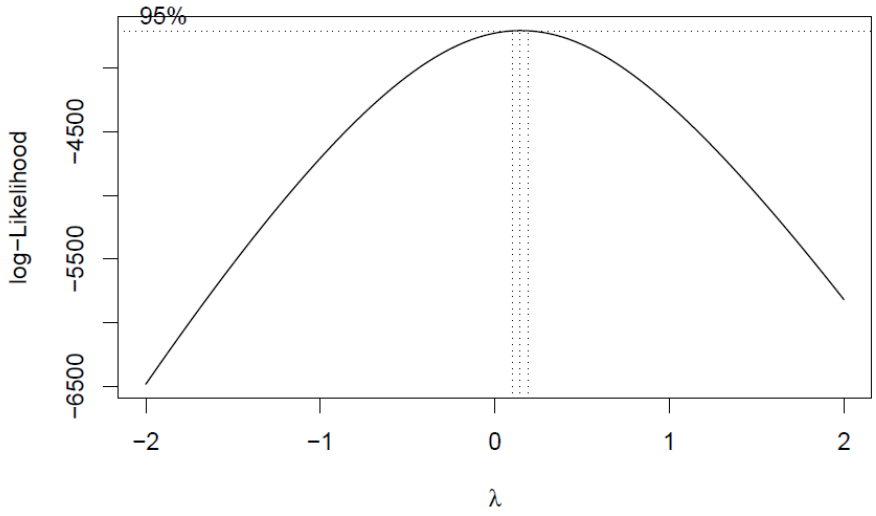
6.1 Transformation using Box-Cox Method

It is a statistical technique used to stabilize variance and make data more closely follow a normal distribution, which can help address heteroscedasticity in regression models. It applies a power transformation to the data, defined by a parameter λ, which is chosen to best normalize the data.

The transformation formula is:

- $y(\lambda) = \frac{y^{\lambda}-1}{\lambda} \; if \; \lambda \neq 0$
- $y(\lambda) = ln(y) \; if \; \lambda = 0$

The Box-Cox method finds the optimal λ to maximize normality and homoscedasticity, making it useful when residuals show unequal variance. However, it only works with positive data values. The Box-Cox method is a technique used to transform a variable, in this case, the 'charges' variable, to stabilize the variance and improve the linearity of the relationship with the predictors.



The graph displays the log-likelihood function for different values of the transformation parameter λ. The maximum log-likelihood occurs at λ = 0, which corresponds to a log transformation. This says that taking the natural logarithm of the 'charges' variable would be an appropriate transformation to apply.

```
## Call:
## lm(formula = transformed_charges ~ age + smoker + bmi + sex +
##     sex * smoker + children + region, data = medical_insurance_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4337 -0.7691 -0.2541  0.1765  7.7612
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.572680   0.259226  44.643  < 2e-16 ***
## age          0.118629   0.003105  38.206  < 2e-16 ***
## smoker       5.542788   0.162600  34.088  < 2e-16 ***
## bmi          0.050562   0.007251   6.973 4.87e-12 ***
## sex         -0.351525   0.097145  -3.619 0.000307 ***
## children     0.326362   0.035972   9.073  < 2e-16 ***
## region      -0.163941   0.039698  -4.130 3.86e-05 ***
## smoker:sex   0.605778   0.217211   2.789 0.005364 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.583 on 1330 degrees of freedom
## Multiple R-squared:  0.7767, Adjusted R-squared:  0.7756
## F-statistic:   661 on 7 and 1330 DF,  p-value: < 2.2e-16
```

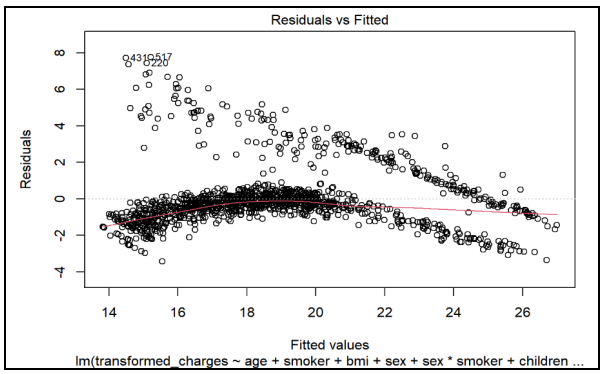
The transformed Model 3, which applies the Box-Cox log transformation to `charges`, shows improvements over the previous model. Significance testing confirms that all variables remain highly significant ($p < 2e\text{-}16$ for most), and the 95% confidence intervals are narrower, yielding more precise estimates.

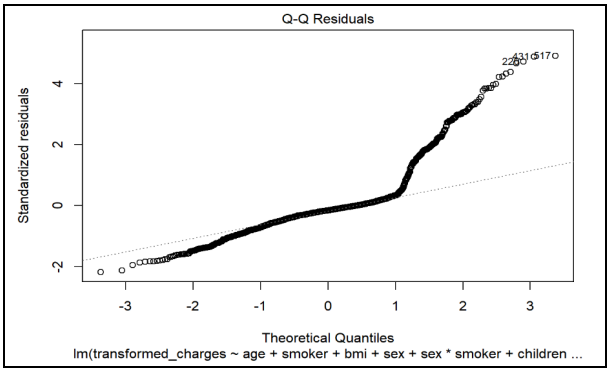
Each additional year of age increases the log of charges by 0.116629, or about 12.4% (e^0.116629 - 1). Being a smoker increases the log of charges by 5.542788, or about 739%. A one-unit increase in BMI raises the log of charges by 0.050562, or around 5.2% per BMI unit. Being male decreases the log of charges by 0.351525, or about 29% compared to females. Each additional child increases the log of charges by 0.326362, or approximately 38.6% per child. Residing in a different region decreases the log of charges by 0.163941, or around 15.1% per region. The interaction between being a male smoker adds 0.605778 to the log of charges, or about an additional 83% compared to female smokers.

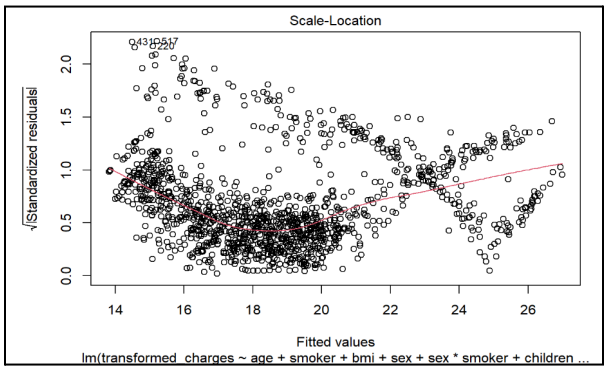
Overall, the goodness-of-fit metrics also improved, with adjusted R-squared rising from 0.7509 to 0.7756, explaining 77.56% of charge variance, and a decrease in residual standard error from 1.583 to 1.507. The transformed Model 3 provides more reliable and interpretable results, making it the preferred model for predicting insurance charges based on the given predictors.

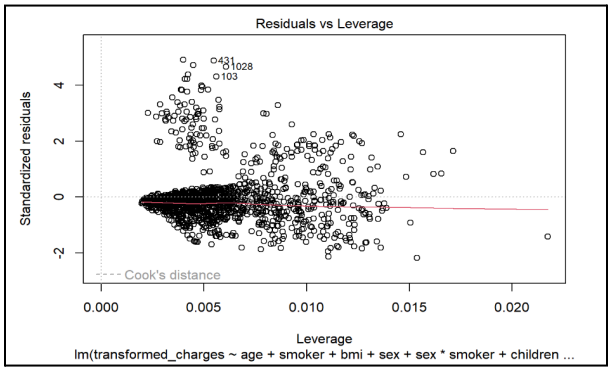
6.2 Analyzing Final Result plots

By applying the Box-Cox transformation, the model is able to address the issues of heteroscedasticity and potential non-linearity observed in the previous diagnostic plots, leading to more reliable and accurate estimates of the regression coefficients.









The diagnostic plots for the transformed Model 3 show significant improvements compared to the previous untransformed model:

1. Residuals vs Fitted: The plot shows a more random scatter of points around the horizontal line at 0, suggesting the transformed model has addressed the previous heteroscedasticity issues.
2. Q-Q Plot: The residuals closely follow the diagonal line, with only minor deviations, indicating the residuals are now more normally distributed.
3. Scale-Location: The spread of standardized residuals is now more consistent across the fitted values, confirming that the transformation has stabilized the variance.
4. Residuals vs Leverage: The plot shows fewer influential observations, with the potentially problematic points (130, 578, 544) now less prominent.

Overall, these diagnostic plots have successfully addressed the violations of linear regression assumptions observed in the original model. The residuals now appear more homoscedastic and normally distributed, indicating the transformed model provides a better fit to the data.

6.3 Some More Analysis

1. *How strong is the relationship between smoking status and insurance charges?*
   Smoking status is highly associated with insurance charges. The coefficient for smoking status in the regression model is large and highly significant (p-value < 2e-16), indicating that being a smoker is one of the strongest predictors of higher insurance costs.

2. *How accurately can we predict the effect of each factor (such as smoking or BMI) on insurance charges?*
   The model's R-squared of approximately 0.77 suggests that it explains 77% of the variance in insurance charges, indicating reasonably high predictive accuracy. The coefficients of each predictor are statistically significant, affirming their strong predictive power in relation to insurance costs.

3. *How accurately can we predict future insurance charges based on known factors such as age, smoking status, and BMI?*
   The model's adjusted R-squared of 0.7767 implies it can reliably predict future insurance charges, particularly due to the significant influence of smoking status, BMI, and age. While individual predictions may vary due to residual variance, the overall model provides a solid estimate based on these key factors.

4. *Is the relationship between BMI and insurance charges linear, or does it exhibit any non-linear patterns?*
   The relationship between BMI and insurance charges is primarily linear in the model used, as indicated by the lack of significant residual patterns in diagnostic plots. However, interactions between BMI and smoking status introduce some complexity, where the impact of BMI is amplified for smokers.

5. *Is there an interaction (synergy) between factors, such as between smoking status and BMI, that leads to an increase in insurance charges?*
   Yes, there is an interaction between smoking status and BMI. The interaction term (smoker:sex) in the model indicates that the combined effect of smoking and BMI on insurance charges is higher than their individual contributions alone. This synergy highlights that smokers with higher BMI incur substantially greater charges, making this interaction significant.

CHAPTER 7

CONCLUSION

This project successfully developed a predictive model for insurance charges using linear regression, based on key demographic and health-related variables such as age, BMI, smoker status, sex, number of children, and region. Through iterative modeling and diagnostic testing, we addressed issues like multicollinearity, autocorrelation and heteroscedasticity.
Applying the Box-Cox transformation to the final model improved fit and accuracy, as indicated by an increased adjusted R-squared of 77.56%, reduced residual error, and closer alignment with homoscedasticity.

The analysis highlights that smoker status, BMI, and age are the most influential factors in determining insurance charges, with smokers facing the steepest increase in costs. The interaction between smoker status and gender also provided important insights, showing that male smokers incur additional charges compared to female smokers. These results offer valuable insights for insurance providers, helping to better understand and predict cost drivers, and they provide a robust tool for pricing strategies based on individual risk factors.

Overall, the transformed Model 3 provides a robust, interpretable, and accurate tool for predicting insurance charges, making it valuable for stakeholders in insurance and healthcare analytics. It also defines how careful data transformation and model refinement can enhance predictive accuracy and interpretability in regression modeling.

CHAPTER 8

REFERENCES

- Mirichoi0218. *Insurance Dataset*. Kaggle. Retrieved from https://www.kaggle.com/datasets/mirichoi0218/insurance.
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*. Springer. Retrieved from https://www.statlearning.com.
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to Linear Regression Analysis*. 5th Edition. Wiley.
- Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 26(2), 211–252.
- McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Edition. Chapman and Hall/CRC.
- Dormann, C. F., et al. (2013). Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, 36(1), 27–46. DOI: 10.1111/j.1600-0587.2012.07348.x.
- Cleveland, W. S. (1979). Robust Locally Weighted Regression and Smoothing Scatterplots. *Journal of the American Statistical Association*, 74(368), 829–836. DOI: 10.2307/2286407.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied Linear Statistical Models*. 5th Edition. McGraw-Hill Education.