Final Project Report:

Claudia Pascual, Andrii Dovhaniuk , Arial Tolentino

2023-12-08

Contents

Introduction 5
Objectives and Significance 5
Descriptive Statistics 5

Exploratory Data Analysis **6**

Regression Analysis **7**
Model Diagnostics 7
Model Interpretation 9

Final Model **10**

Summary and Conclusions **10**
Model Evaluation and limitations 10
Concluding thoughts 10

List of Figures

1 Left Figure: Response Variable and Right 2: Dataset 5

2 EDA:Scatter plot of response variable with each independent variables under consideration . 7

3 Top Left Figure: Residuals Plot, Top Right Figure: Normal Q-Q Plot, Bottom Center Figure: Histogram 7

4 Top Left Figure: Residuals Plot, Top Right Figure: Normal Q-Q Plot, Bottom Center Figure: Histogram 8

List of Tables

1 Data Summary Statistics . . . 6
2 Model 1: Multicolllinarity Analysis . . . 8
3 Model Summary Statistics . . . 9
4 Coefficents . . . 9

Introduction

In this study, we explore and analyze different factors that affect medical insurance costs. The primary question guiding our analysis is: Can insurance costs be accurately predicted based on factors such as age, gender, body mass index, number of children, smoking habits, and residential region?

To approach this question, we utilize a dataset as featured in "Machine Learning with R" by Brett Lantz, which provides a comprehensive introduction to machine learning using R. This dataset contains information on factors like age, gender, BMI, number of children, smoking status, and geographical region, all of which are potential influencers of medical insurance costs.

In the article Healthcare Information Management Systems, they state "Among the elderly, prevalent health issues such as heart disease, diabetes, arthritis, dementia, and respiratory conditions are not only common but also chronic in nature, necessitating sustained care and thereby incurring substantial treatment costs." However, this alone does not fully explain the significant disparities in healthcare spending observed between the United States and other wealthy countries.

Objectives and Significance

By looking at this data, we hope to find patterns that can help us understand what factors have an influence on how much individuals spend on healthcare. This leads us to our research hypotheses. The Null Hypothesis is that there is no significant relationship between individual medical costs and factors like age, gender, BMI, number of children, smoking habits, and residential region. The Alternative Hypothesis suggests that significant associations exist between individual medical costs and these variables.

Our goal is to provide useful insights that contribute to the broader discussion about how healthcare is funded and what factors play a role in shaping individual medical expenses. By examining the data, we hope to predict medical charges more accurately but also to enhance the understanding of healthcare analytics and forecasting within the insurance industry. The insights garnered from this study could be pivotal for policy-making, insurance plan design, and individual decision-making in healthcare.

Descriptive Statistics

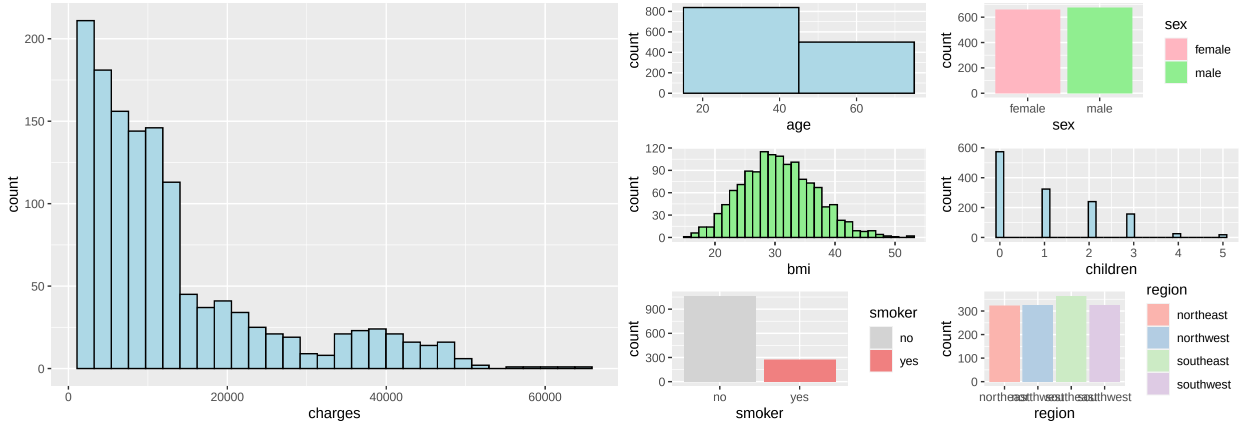

Figure 1: Left Figure: Response Variable and Right 2: Dataset

We aim to investigate the factors that impact medical insurance costs, utilizing the "Medical Cost Personal Datasets." This datase contains samples of individuals collected across regions in the United States, and includes the following key variables:

Variable	Type	Min	Max	Descr.	SD
Age	Num Con	18	64	The age of the main person covered by the insurance..	14.05
Sex	Cat	M	F	The gender of the main person covered by the insurance (male, female).	N/A
BMI	Num Con	15.96	53.13	Body Mass Index	6.10
Children	Num Con	0	5.00	The number of independents covered.	1.21
Smoker	Cat	0	1	If the primary person smokes (yes, no).	N/A
Region	Cat	15.96	53.13	Where the person/family lives (northeast, southeast, southwest, northwest).	N/A
Charges	Num Con	-15537.55	48161.28	Individual cost of the medical bill from the insurance.	11944.732273

Table 1: Data Summary Statistics

Exploratory Data Analysis

The analysis of medical charges in relation to various factors reveals distinct trends and correlations. Firstly, there is a positive correlation between age and medical charges, indicating that as individuals age, their medical expenses tend to rise. This trend underscores the increased healthcare needs typically associated with aging.

In examining the relationship between charges and Body Mass Index (BMI), an interesting pattern emerges. The data seems to bifurcate into two groups, each with different slopes. This suggests that higher medical charges can be incurred across various BMI levels, indicating that BMI alone is not a definitive predictor of medical expenses.

When the charges are analyzed based on sex, the distribution appears similar for both females and males. However, a closer inspection hints that males might incur slightly higher charges on average, suggesting a subtle gender-based difference in medical expenses.

A notable trend is observed when considering the number of children. Here, a negative correlation is evident, with individuals having fewer children tending to incur lower medical charges. This trend might reflect the additional healthcare costs associated with childbearing and childrearing.

One of the most pronounced distinctions is seen in the charges related to smoking status. Non-smokers consistently face lower medical charges compared to smokers, who exhibit significantly higher expenses. This stark difference highlights the substantial financial impact of smoking on healthcare costs.

Finally, the influence of geographical region on medical charges appears to be minimal. The distribution of charges across different regions — northeast, northwest, southeast, and southwest — is relatively uniform, suggesting that regional factors do not play a major role in determining medical expenses. This uniformity across regions indicates that other factors might be more influential in driving medical costs.

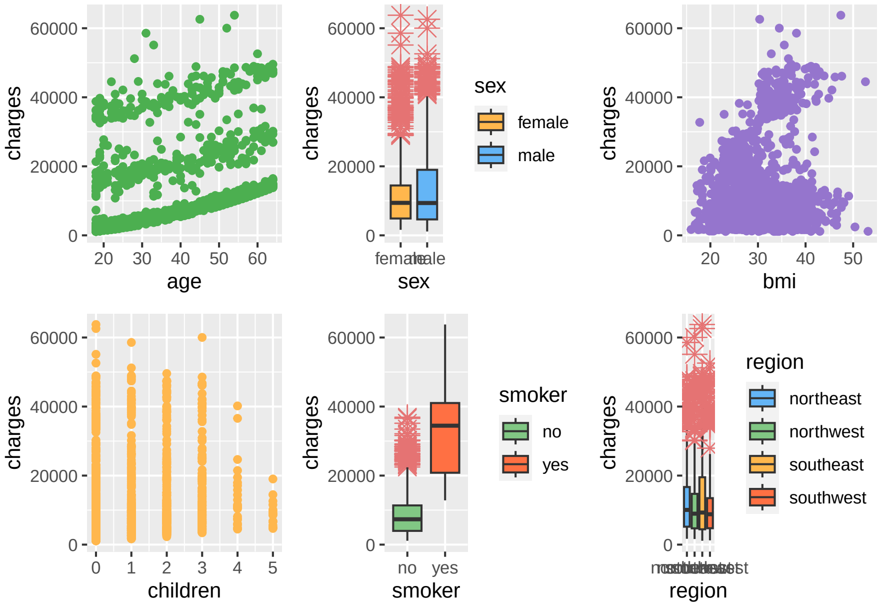

Figure 2: EDA:Scatter plot of response variable with each independent variables under consideration

Regression Analysis

Model Diagnostics

After the analysis of a regression model designed to predict health insurance charges based on various factors. Two models were developed, and Model 2 emerged as the final choice after careful consideration of diagnostic tests and model assumptions.

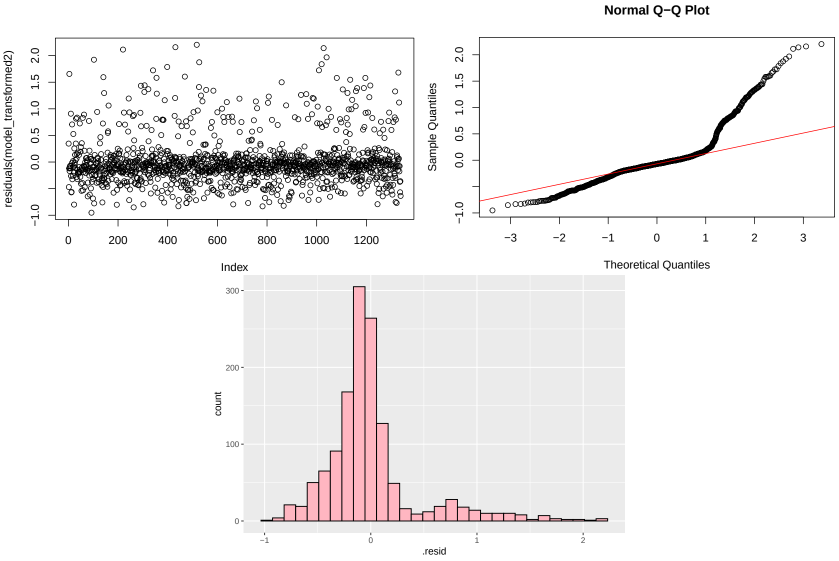

Figure 3: Top Left Figure: Residuals Plot, Top Right Figure: Normal Q-Q Plot, Bottom Center Figure: Histogram

Initially, we did multicollinearity analysis which can be sees in Table 2 and they revealed high Variance Inflation Factors (VIF) for age, age squared, bmi, and bmi squared, indicating multicollinearity. However, children, smoker, sex, and region showed low VIF, suggesting little correlation. Next, we did second-order terms for age and bmi in Table 2, along with interaction terms between smoker and sex, were included initially. Model 1 had a good R^2 but very bad residual standard error.

	GIF	DF	$GIVF^{\frac{1}{2DF}}$
age	47.629207	1	6.9011392
$I(age^2)$	47.578030	1	6.896783
bmi	47.629207	1	7.710307
$I(bmi^2)$	47.629207	1	7.717861
children	47.629207	1	1.050137
smoker	47.629207	1	1.520916
sex	1.260685	1	1.122802
region	1.113884	3	1.018138
smoker: sex	2.650825	1	1.628135

Table 2: Model 1: Multicolllinarity Analysis

However, Model 1 failed to meet key assumptions and displayed poor performance. In Figure 4 as one can see Residuals plot showed no clear patterns, but QQ plot and histogram indicated non-normality. Hence as mentioned before multiple assumptions were not satisfied. To address this issues in Model 1, further transformations were applied to predictors, with particular attention given to region and bmi. Despite improvements, the assumptions were still not fully satisfied. Ultimately, Model 2 incorporated additional transformations on the predictor, leading to better performance.

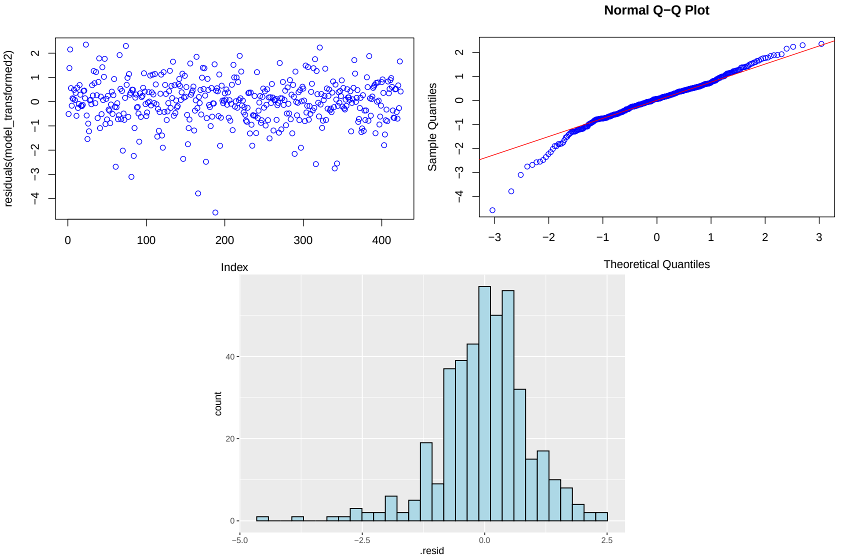

Figure 4: Top Left Figure: Residuals Plot, Top Right Figure: Normal Q-Q Plot, Bottom Center Figure: Histogram

The plots for Model 2 in Figure Number 4 show that the normality, linearity, and independence assumptions are satisfied. To go more in details, Residuals plot satisfied linearity, QQ plot and histogram displayed normality. The only assumption that was not met is constant variance in Residual plot. Our model has a good fit for the majority of the data but may not capture extreme values well. Outliers and wider spread data could make the model less reliable and worse at predicting.

Therefore, after evaluating both models, Model 2 was selected as the final model, addressing issues identified in Model 1. While Model 1 had 0.75 multiple residual R^2-adjusted, it did not meet criteria as the residual standard error was too high as well as the Residual plot showed too much random scatter indicating nonlinearity. While the model has strengths in explaining variance and significant predictors, limitations such as homoscedasticity and multicollinearity should be considered in its application.

Model Interpretation

Residual standard error	0.9157 on 412 degrees of freedom
(914 observations deleted due to missingness)	
Multiple R^2	0.4965
Adjusted R^2	0.4831
F-statistic	36.94 on 11 and 412 DF
p-value	<2.2e-16

Table 3: Model Summary Statistics

Our final model explains the proportion of variance; here, it's about 49.65% (Table3). R^2-adjusted is adjusted for the number of predictors in Model 2. Next, a F-statistic was performed which is a test for the overall significance of the model was high and had a low p-value (less than 2.2e-16). It suggests that the model is statistically significant.

	Estimate	Std. Error	t-value	$\Pr(> \lvert t \rvert)$
(Intercept)	1.7330273	1.0022121	1.729	0.084522
age	-0.1050149	0.0216119	-4.859	1.68e-06 ***
I(age^2)	-0.0013008	0.0002612	-4.979	9.39e-07 ***
bmi	0.2237259	0.0593048	3.772	0.000185 ***
I(bmi^2)	-0.0024347	0.0009381	-2.595	0.009784 **
children	0.0314315	0.0386744	0.813	0.416847
smokeryes	1.4813810	0.1385957	10.689	<2e-16 ***
sexmale	0.1896573	0.1525646	1.243	0.214528
regionnorthwest	0.0769521	0.1278220	0.602	0.547488
regionsoutheast	0.0607494	0.1266427	0.480	0.631701
regionsouthwest	-0.0949230	0.1306582	-0.726	0.467946
smokeryes:sexmale	-0.2085118	0.1900063	-1.097	0.273110

Table 4: Coefficents

From Table 4 the coefficient for age is positive, indicating that as a person's age increases, their insurance charges also tend to increase. The negative coefficient for the age-squared term implies a diminishing return effect. As age increases, the rate at which insurance charges increase slows down. The positive coefficient for BMI implies that individuals with higher BMI values tend to have higher insurance charges. The negative coefficient for the BMI-squared term suggests a diminishing return effect, similar to the age-squared term. The positive coefficient for the number of children suggests that having more children is associated with higher insurance charges. Being a smoker is associated with a significant increase in insurance charges. The significance of this effects are supported by the low p-value. The coefficients for regions represent the differences in charges compared to a baseline region (likely the "northeast"). The negative coefficient for the interaction term suggests that the effect of being a smoker on insurance charges is different for males compared to females. None of these coefficients are statistically significant, indicating that there's no significant difference and caution in interpreting this particular interaction effect.

Final Model

$$\ln(Y) = \beta_0 + \beta_1\text{Age} + \beta_2\text{Age}^2 + \beta_3\text{BMI} + \beta_4\text{BMI}^2 + \beta_5\text{Children} + \beta_{\text{smoker}\times\text{sex}}(\text{Smoker}\times\text{Sex}) + \beta_{\text{region}}\text{Region} + \epsilon$$

Assumptions of ϵ

- $\epsilon \sim N(0, \sigma^2)$
- Constant Variance $\sigma \neq \sigma(X)$

Summary and Conclusions

Analyzing the insurance dataset has revealed significant insights. It confirms that certain factors, notably age, BMI, and smoking status, play a crucial role in determining medical insurance costs. These findings align with the observations noted in "Healthcare Information Management Systems," particularly regarding the chronic nature of diseases and their impact on healthcare spending.

Model Evaluation and limitations

While our final model does well in capturing the general trends and relationships, it has limitations. Particularly, its ability to predict extreme values and outliers is constrained. This limitation is critical, especially in the context of healthcare, where extreme cases can represent significant costs. The issue of homoscedasticity also indicates that our model, despite its strengths, may not be fully appropriate for all levels of independent variables.

Concluding thoughts

In summary, this project adds to the current conversation about understanding and predicting healthcare costs in the insurance world. In concluding our research, we've identified several important factors that significantly affect medical expenses. It's also important that we realize where our prediction methods aren't perfect yet. This awareness helps us move toward developing more refined and effective methods for managing and forecasting healthcare costs.