MKTG 480/580
Professor Gupta
Group 9: Christopher Rally | Alex Szczepanski | Anthony Woods | Ariana Woo Tan

Executive Summary

The intricate pricing of health insurance premiums plays a vital role in maintaining the viability of the industry. Health insurance companies practice price discrimination, tailoring premiums based on individual risk profiles. Complex statistical models are used to determine premium pricing, as certain attributes of each policyholder guide actuaries in predicting future costs, allowing for an analytical approach to premium setting based on probabilities.

This project aimed to determine physical factors influencing health insurance premiums, realizing their magnitude and directionality through statistical analysis. Using a dataset encompassing 1,300 U.S. insurance policyholders, we conducted rigorous statistical analyses, employing linear regressions with six key variables—age, sex, BMI, children, smoker status, and region. Conclusions highlighted 'smokeryes,' 'age,' 'bmi,' 'children,' 'regionsoutheast,' and 'regionsouthwest' as key determinants, with age, BMI, and smoker status exerting the most pronounced impact. Regional disparities retained significance, reflecting variations in competition, political environments, and the cost of living. Identified limitations included sample size and variable range, lack of diversity and a small number of variables.

Looking ahead, this study's results offer valuable insights for marketing managers, guiding strategic initiatives in collaboration with healthcare providers. Changes that promote preventive care, incentivization for policyholder engagement, and tailored communication strategies position managers as key drivers of positive outcomes in the health insurance

landscape. The study's knowledge extends beyond insurers, impacting policyholders and promoting healthier lifestyle choices.

Table of Contents

Introduction

In countries where healthcare is privatized such as the United States, it is very common for individuals to rely on health insurance. Health insurance serves as a safety net, covering medical and surgical expenses, as well as prescription drugs and medications, taking the financial burden away from the individual. However, this coverage comes at a cost – a monthly premium paid by policyholders to ensure financial protection in case of medical need.

Health insurance premiums are significant when it comes to ensuring the health insurance business remains viable. Companies practice price discrimination, charging each individual customer a different price. This is because Insurance companies profit from making sure that their total insurance payouts are lower than their total income from premiums. To ensure this, those with a higher likelihood of needing a payout must be charged a higher premium than those less likely.

Determining the likelihood of a policyholder requiring a payout is complex and involves various factors. The major factors that insurance companies consider when determining premium price are age, location, tobacco use, and number of dependents (1). Insurance companies cannot, however, take sex into account while deciding the likelihood of requiring a payout. These factors are analyzed through statistical models by actuaries hired by insurance companies to predict future costs and risks for specific characteristics and combinations of factors (2). This data-driven approach allows premiums to be set based on each unique risk profile, allowing companies to maintain a sustainable business model and generate profits.

For our project, we aimed to determine what physical factors impact the prices of health insurance premiums and their weight when effecting that decision. Although we already know what factors are generally used in determining prices of premiums, we do not know the extent to which they have an impact and whether or not it is positive or negative. For our analysis, we used a dataset consisting of over 1,300 individual insurance policy holders from the US and ran multiple regressions on the dataset. Our dependent variable was ‘charges’ which is a float type variable that shows the cost of medical insurance premiums by the individual. Our regression consisted of six independent variables. ‘Age’; an integer variable stating how old the individual is, ‘sex’; a string-type variable indicating the gender of the individual, ‘bmi’; a float type variable showing the body mass index of the individual, ‘children’; an integer type variable that indicates the number of dependants the individual has, ‘smoker; a string-type dummy variable showing weather the individual is a smoker or not, and lastly ‘region’; a string-type variable indicating what geographical sector of the United States the individual resides.

Through statistical analysis, we sought to determine how the interplay between these variables affect premium prices. From our regressions, we were able to yield coefficient estimates, standard error values, and p-values. These results informed us just how much of an impact each independent variable had on price, and whether or not this effect was statistically significant or not.

The conduction of this analysis is important for a number of reasons. By understanding just how much of an effect certain factors have on insurance premium prices, we can promote transparency and fairness in the health insurance industry. With a widespread understanding of premium pricing, disparities in premium pricing practices can be identified and can ensure that policyholders are treated justly. This data can also indirectly address public health concerns. By understanding how much variables such as BMI and smoker status can have on insurance premium, the dangers of obesity and smoking can be truly highlighted.

Methodology & Findings

We first sought to visualize the relationships of the variables with charges by creating graphs. A graph was created for each variable using the graph that told a story about its relationship with charges the best. If there was a more profound question, a more in-depth analysis was done using R, where the graphs were also created. The actual relationship is shown in the regression analysis later on. The graph below (Fig. 1) shows the relationship between the dummy variable sex and average charges. This shows a slightly larger average charge for males compared to females. This may be because this dataset contains more male smokers than females (107 males; 80 females), raising the average male charges. This is shown later in the regression analysis, where sex is an insignificant variable. The following relationship that we visually depict (Fig. 2) is the average charges of smokers and non-smokers. This huge average difference leads to visual differences in the relationship between charges and other variables. The following relationship of age and charges depicts how the variable smoker acts as a moderator between the two. Fig. 3 best represents the relationship between the two

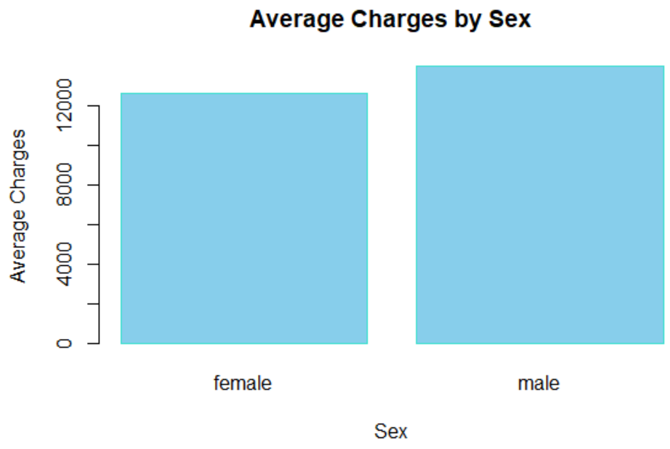


Fig. 1

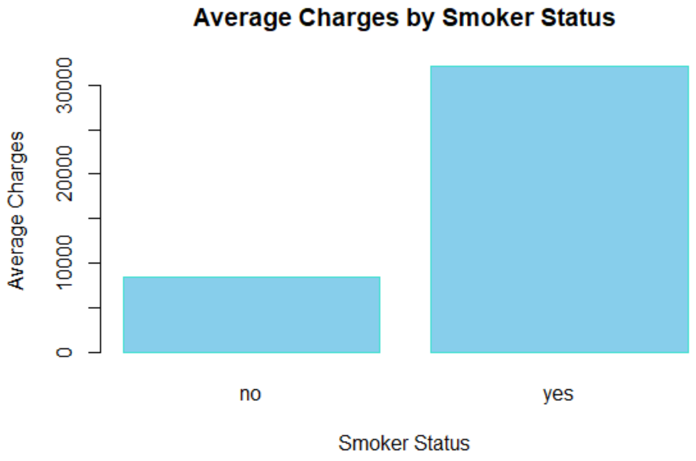


Fig. 2

variables, showing a steady increase in median charges as the age group increases. But, in Fig. 4, we attempt to depict this as a direct relationship using a scatter plot. This figure shows three separate direct correlations between age and charges. We initially hypothesized that this was because there are different health insurance packages the customer can sign up for. But, as shown in Fig. 5 and Fig. 6, the hard cutoff for the relationship with lower charges and the two relationships with higher charges are caused by smokers and non-smokers. Notice that in Fig. 5, data from only non-smokers, most of the data is within charges of 15,000. In Fig 6, data with smokers, all charges are greater than 15,000. We attempted to determine where the cutoff point is for the higher charges by filtering the other variables and plotting them on the same graph. This yielded no apparent findings. The following visualization (Fig. 7) depicts the relationship between median charges and the region of the United States in which the individual charged

Fig. 3

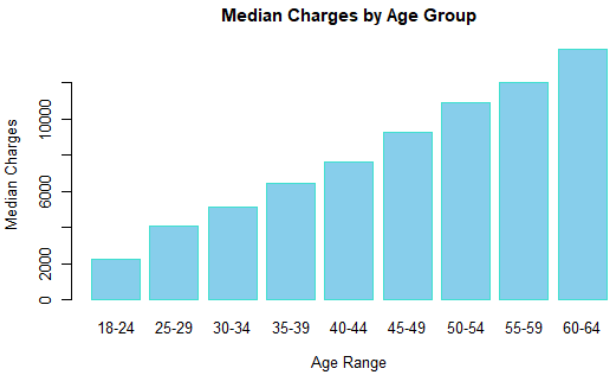


Fig. 4

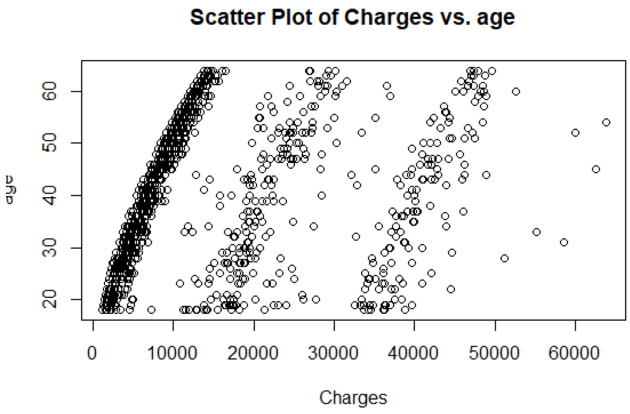


Fig. 6

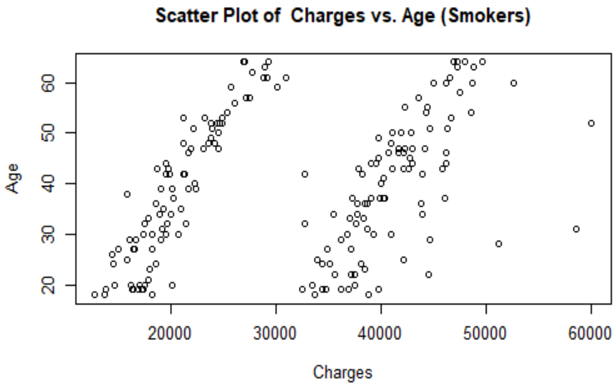
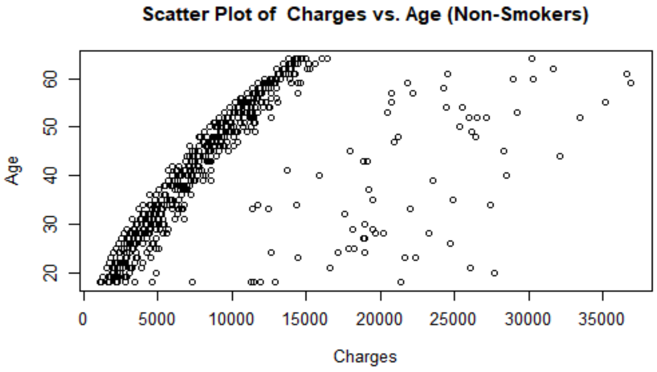


Fig. 5



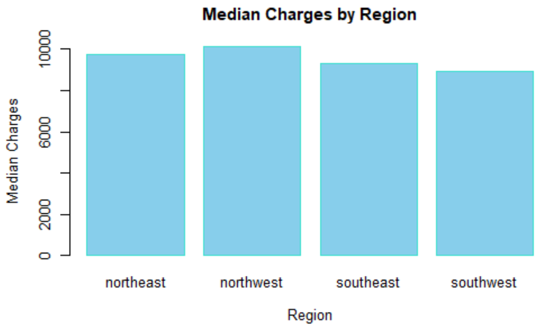


Fig. 7

lives. The chart shows that the median charges in the northwest are slightly higher than the other regions. It also shows that the north generally has higher charges than the south. According to the U.S. government's healthcare website, this is because of differences in competition in these areas, the political environment, and the cost of living in the area (2). Fig. 8 shows a scatterplot of the charges relationship with bmi. This graph produces a Pearson's correlation value of 0.2013. This is a fairly weak correlation, but it is powerful enough to yield significant effects in the linear regression model. The last graph (Fig. 9) depicts the relationship between the policyholder's number of children and their number of children. This shows that the number of children the policyholder has is not the most significant determinant of charges. There is an extensive range of charges for every amount of children, excluding 4 and 5, most likely because of the lack of individuals who fall into this category. But, it seems that charges have much to do with the other variables rather than the number of children.

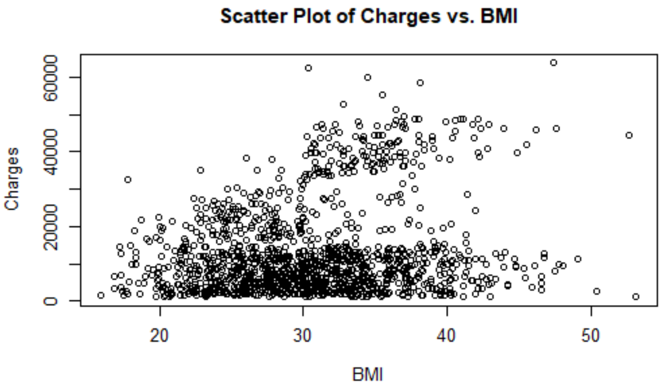


Fig. 8

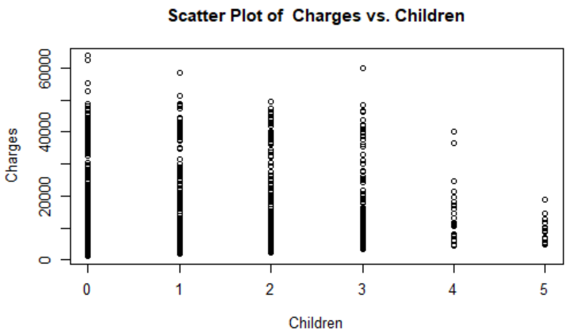
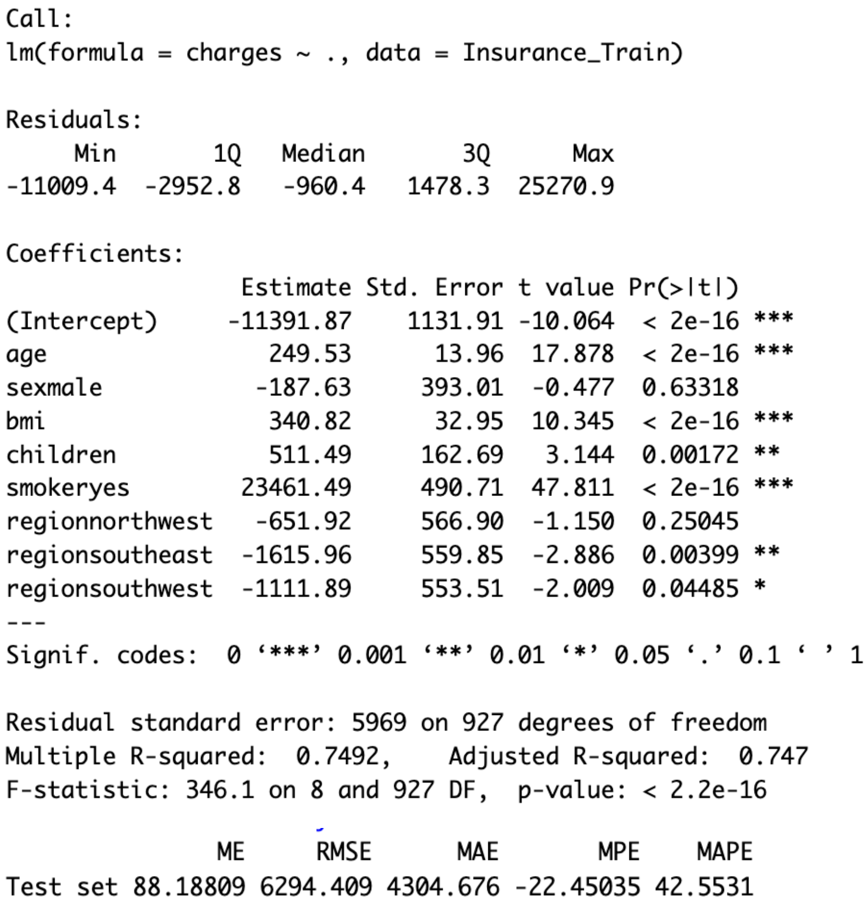


Fig. 9

Based on the relationships in the graphs above, we ran a linear regression using backward, forward, and stepwise selection models to find out which variables affect charges the most. The reason why we decided to use the latter selection models is because linear regression incorporates all predictors simultaneously. In contrast, selection methods include or exclude predictors based on specific criteria to make the model and results simpler and more accurate. Additionally, these selection techniques offer a more methodical way to choose variables compared to linear regression, which lacks a selection process for predictors.

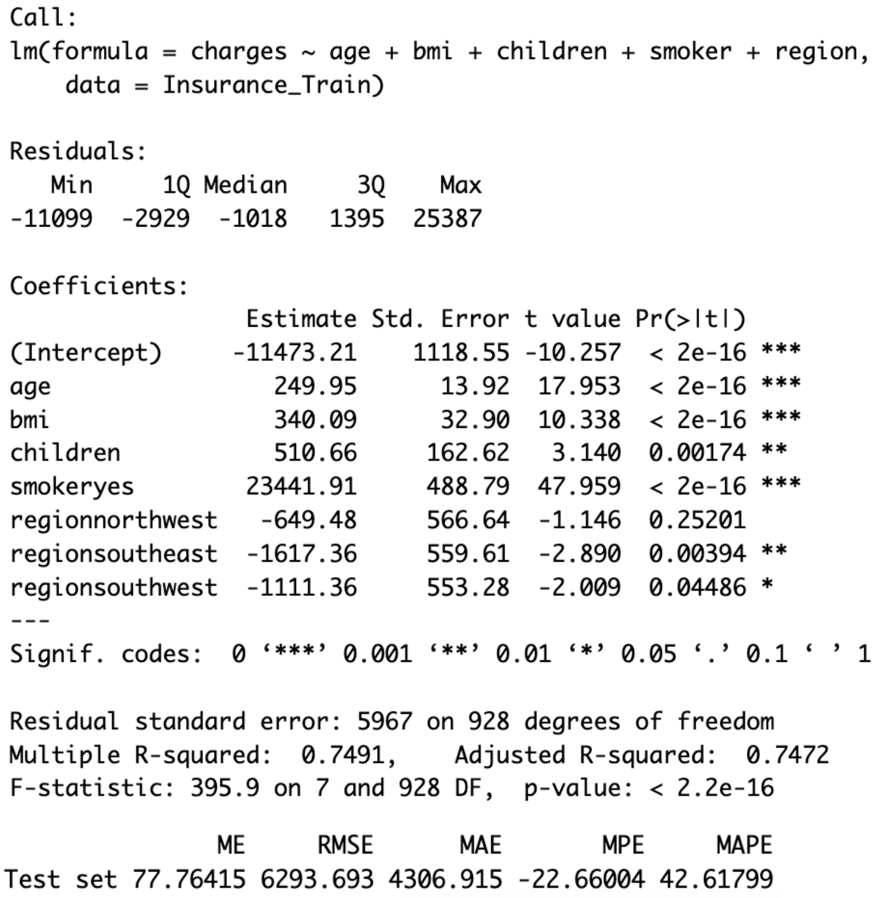
- **Linear Regression Results**:



```
Call:
lm(formula = charges ~ ., data = Insurance_Train)

Residuals:
     Min       1Q   Median       3Q      Max 
-11009.4  -2952.8   -960.4   1478.3  25270.9 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11391.87    1131.91 -10.064  < 2e-16 ***
age                249.53      13.96  17.878  < 2e-16 ***
sexmale           -187.63     393.01  -0.477  0.63318    
bmi                340.82      32.95  10.345  < 2e-16 ***
children           511.49     162.69   3.144  0.00172 ** 
smokeryes        23461.49     490.71  47.811  < 2e-16 ***
regionnorthwest   -651.92     566.90  -1.150  0.25045    
regionsoutheast  -1615.96     559.85  -2.886  0.00399 ** 
regionsouthwest  -1111.89     553.51  -2.009  0.04485 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5969 on 927 degrees of freedom
Multiple R-squared:  0.7492,    Adjusted R-squared:  0.747 
F-statistic: 346.1 on 8 and 927 DF,  p-value: < 2.2e-16

               ME     RMSE      MAE       MPE    MAPE
Test set 88.18809 6294.409 4304.676 -22.45035 42.5531
```

We ran a linear regression model, using 70% of the data as training data and 30% as validation data, with a set seed of 100. ‘sexmale’ and ‘regionsoithwest’ were the only variables that were not relevant due to their p-values exceeding 0.05. The adjusted R-squared of this model was 0.747, which indicates that approximately 74.7% of the changes in the dependent variable can be explained by the independent variables in our regression model. Moreover, we used the accuracy() function in order to evaluate the accuracy and performance of this regression model. Our model yielded an RMSE of 6294.409, a high figure explained by the nature of the dependent variable, 'charges.'

- **Backward Selection Results**:

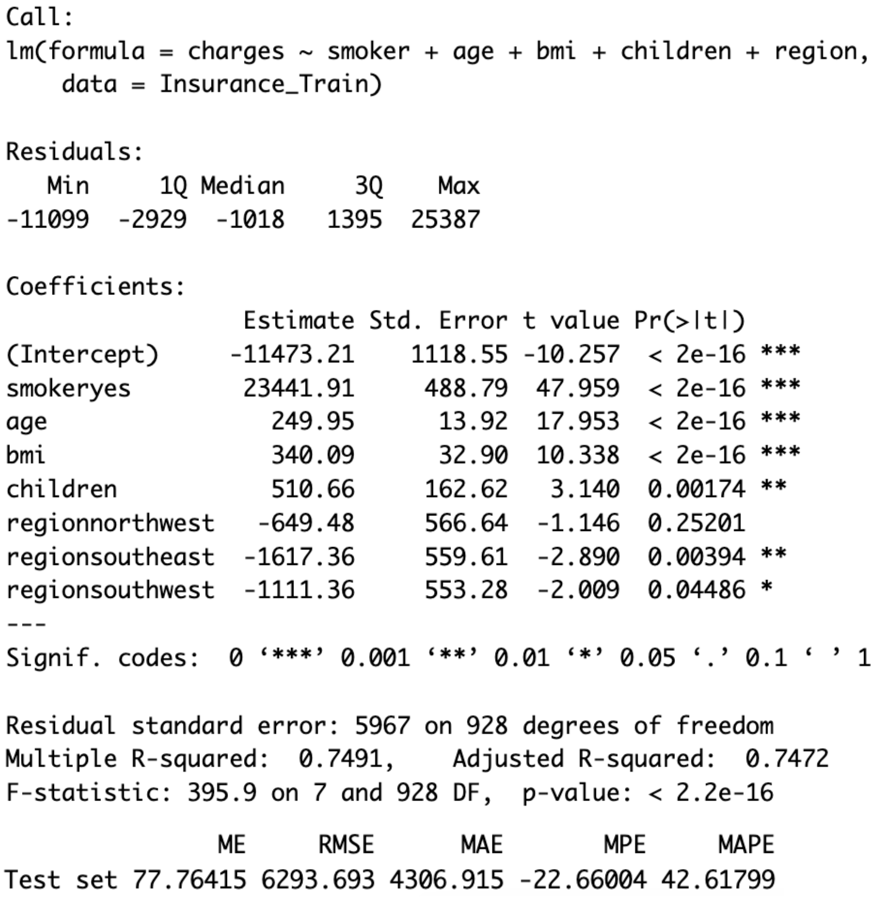


```
Call:
lm(formula = charges ~ age + bmi + children + smoker + region,
    data = Insurance_Train)

Residuals:
   Min     1Q Median     3Q    Max 
-11099  -2929  -1018   1395  25387 

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)    
(Intercept)     -11473.21    1118.55 -10.257  < 2e-16 ***
age                249.95      13.92  17.953  < 2e-16 ***
bmi                340.09      32.90  10.338  < 2e-16 ***
children           510.66     162.62   3.140  0.00174 ** 
smokeryes        23441.91     488.79  47.959  < 2e-16 ***
regionnorthwest   -649.48     566.64  -1.146  0.25201    
regionsoutheast  -1617.36     559.61  -2.890  0.00394 ** 
regionsouthwest  -1111.36     553.28  -2.009  0.04486 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5967 on 928 degrees of freedom
Multiple R-squared:  0.7491,	Adjusted R-squared:  0.7472 
F-statistic: 395.9 on 7 and 928 DF,  p-value: < 2.2e-16

               ME     RMSE      MAE       MPE     MAPE
Test set 77.76415 6293.693 4306.915 -22.66004 42.61799
```

For our Backward Selection model, only ‘sexmale’ was dropped, which isn't unexpected as it had the highest p-value in the linear regression model. However, it's surprising to see that ‘regionnorththwest’ wasn't dropped, despite having a p-value much higher than 0.05 in both the linear regression and backward selection models. Relevant variables for this model include ‘smokeryes’. ‘age’, ‘bmi’, ‘children’, ‘regionsoutheast’, and ‘regionsouthwest’.

Furthermore,The adjusted R-squared of this model was 0.7472, which indicates that approximately 74.72% of the changes in the dependent variable can be explained by the independent variables in our regression model. Assessing accuracy with the accuracy() function, our backward selection model yielded an RMSE of 6293.693, a high figure explained by the nature of the dependent variable, 'charges.'. This RMSE is a little bit lower and better than the Linear Regression model RMSE of 6294.409.

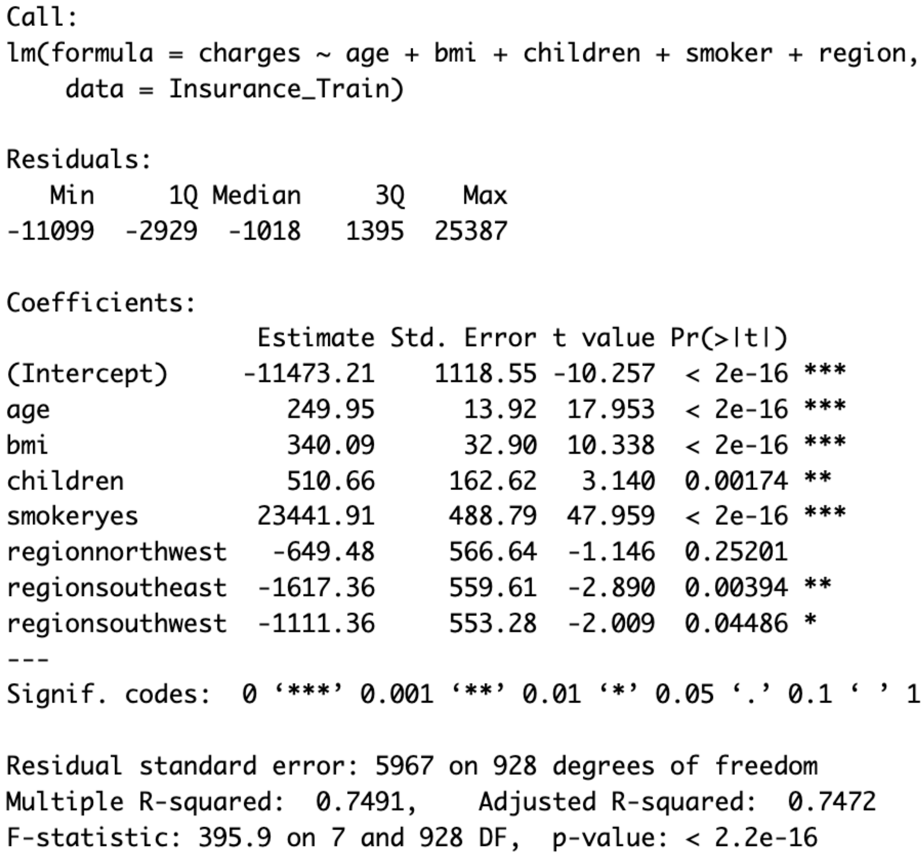
- **Forward Selection**:



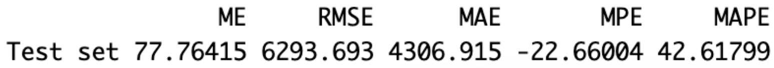
```
Call:
lm(formula = charges ~ smoker + age + bmi + children + region,
    data = Insurance_Train)

Residuals:
   Min     1Q Median     3Q    Max
-11099  -2929  -1018   1395  25387

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11473.21    1118.55 -10.257  < 2e-16 ***
smokeryes        23441.91     488.79  47.959  < 2e-16 ***
age                249.95      13.92  17.953  < 2e-16 ***
bmi                340.09      32.90  10.338  < 2e-16 ***
children           510.66     162.62   3.140  0.00174 **
regionnorthwest   -649.48     566.64  -1.146  0.25201
regionsoutheast  -1617.36     559.61  -2.890  0.00394 **
regionsouthwest  -1111.36     553.28  -2.009  0.04486 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5967 on 928 degrees of freedom
Multiple R-squared:  0.7491,    Adjusted R-squared:  0.7472
F-statistic: 395.9 on 7 and 928 DF,  p-value: < 2.2e-16

               ME     RMSE      MAE       MPE     MAPE
Test set 77.76415 6293.693 4306.915 -22.66004 42.61799
```

The results from our Forward Selection model were very similar to those of our Backward Selection model. The only variable dropped was ‘sexmale’. Relevant variables for this model include ‘smokeryes’. ‘age’, ‘bmi’, ‘children’, ‘regionsoutheast’, and ‘regionsouthwest’. Furthermore, the adjusted R-squared of this model was 0.7472, which indicates that approximately 74.72% of the changes in the dependent variable can be explained by the independent variables in our regression model. Assessing accuracy with the accuracy() function, our Forward Selection model yielded an RMSE of 6293.693, a high figure explained by the nature of the dependent variable, 'charges.' This RMSE is a little bit lower and better than the Linear Regression model RMSE of 6294.409.

- **Stepwise Selection:**



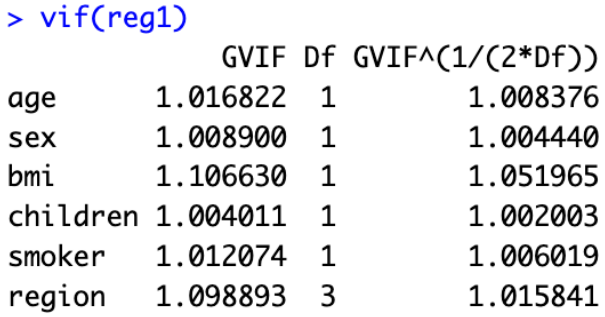
```
Call:
lm(formula = charges ~ age + bmi + children + smoker + region,
    data = Insurance_Train)

Residuals:
   Min     1Q Median     3Q    Max
-11099  -2929  -1018   1395  25387

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11473.21    1118.55 -10.257  < 2e-16 ***
age                249.95      13.92  17.953  < 2e-16 ***
bmi                340.09      32.90  10.338  < 2e-16 ***
children           510.66     162.62   3.140  0.00174 **
smokeryes        23441.91     488.79  47.959  < 2e-16 ***
regionnorthwest   -649.48     566.64  -1.146  0.25201
regionsoutheast  -1617.36     559.61  -2.890  0.00394 **
regionsouthwest  -1111.36     553.28  -2.009  0.04486 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5967 on 928 degrees of freedom
Multiple R-squared:  0.7491,    Adjusted R-squared:  0.7472
F-statistic: 395.9 on 7 and 928 DF,  p-value: < 2.2e-16
```



```
               ME     RMSE      MAE       MPE     MAPE
Test set 77.76415 6293.693 4306.915 -22.66004 42.61799
```

The results from our Stepwise Selection model were pretty much the same as those of our Backward Selection model and Forward Selection model.The only variable dropped was ‘sexmale’. Relevant variables for this model include ‘age’, ‘bmi’, ‘children’, ‘smokeryes’, ‘regionsoutheast’, and ‘regionsouthwest’. Furthermore, the adjusted R-squared of this model was 0.7472, which indicates that approximately 74.72% of the changes in the dependent variable can be explained by the independent variables in our regression model. Assessing accuracy with the accuracy() function, our Stepwise Selection model yielded an RMSE of 6293.693, a high figure explained by the nature of the dependent variable, 'charges.'This RMSE is a little bit lower and better than the Linear Regression model RMSE of 6294.409.

- **VIF ()**



```
> vif(reg1)
             GVIF Df GVIF^(1/(2*Df))
age      1.016822  1        1.008376
sex      1.008900  1        1.004440
bmi      1.106630  1        1.051965
children 1.004011  1        1.002003
smoker   1.012074  1        1.006019
region   1.098893  3        1.015841
```

We used the VIF() function in order to make sure multicollinearity isn't a problem in our models. We can confidently say there is no significant multicollinearity in our models. As we can see from the results, GVIF values are relatively low at 1, none going above the common values of concern (usually 5 or 10, sometimes even 2.5). GVIF values for ‘age’, ‘sex’, ‘children’, ‘smoker’, and ‘region’ are relatively similar at 1, which means there’s almost no correlation among these variables. ‘BMI’ shows the highest GVIF at 1.06, but this value is quite low, suggesting multicollinearity is not a significant concern in our models

Resulting Model

Overall, the results of both our forward selection and backward elimination coincided quite closely to our linear regression model giving us an output of: **Charges** = -11473.21 + 249.95age + 340.09bmi + 510.66children + 2344.91smoker - 649.48northeast - 1617.36southeast - 1111.36southwest. Considering that the two processes led to the same results, even down to the intercept, it indicates that this is the most accurate predictive model when it comes to forecasting general healthcare insurance charges. Despite this scenario not being entirely realistic, logically this model insinuates that for a hypothetical person with an age of 0 years old, 0 bmi, no children, a nonsmoker, and inhabiting the northwest region, then the amount of healthcare insurance expensed would be -$11,473.21.

Conclusions

In our variable selection process, 'sexmale' was the sole variable dropped from our model, while 'regionnorthwest' was not explicitly recognized within the model since it served as our categorical dummy variable, and was not included to avoid multicollinearity. Ultimately, the key determinants of health insurance premium, as indicated by their significant p values (< 0.05), were ‘smokeryes’, ‘age’, ‘bmi’, ‘children’, ‘regionsoutheast’, and ‘regionsouthwest’, with age, bmi and smoker status as the variables with the most pronounced impact on healthcare utilization, as reflected in their strongest p-values. Essentially, an increase in age, BMI, or a positive smoker status would underline heightened health risks, prompting insurers to adjust policyholders’ premiums accordingly. While not as pronounced, regional disparities retained significance within the model, as these variations in variables can be attributed to differences in competition, political environments, and the cost of living when determining premium costs.

Limitations and Recommendations

Some of the identified limitations include small sample size, lack of diversity, overlooking temporal changes as well as our limitation of variables included. To address sample population concerns comprehensively, an approach involves exploring insurance policies of other countries to analyze whether the identified variables play a similar role in those contexts, providing a more expansive and diverse dataset for analysis. In consideration of temporal changes in variable relationships and insurance premiums, future research should involve thorough longitudinal analyses, considering contextual and external factors such as economic conditions. This approach ensures the development of dynamic and responsive insurance pricing models that accurately reflect the evolving nature of insurance dynamics over time. Lastly, the limitation posed by our restricted range of variables, including region, sex, BMI, age, children, and smoker status, prompts the potential for a more nuanced model. Future research endeavors should focus on incorporating additional variables, such as medical history, occupation, and activity status, contributing to a more all-encompassing/robust understanding for assessing healthcare insurance expenses.

Future Research

Looking ahead, marketing managers can play a pivotal role in leveraging research insights for strategic initiatives. Collaborating with healthcare providers to create joint marketing initiatives emphasizes the importance of preventive care and the supportive role of insurance in overall health. Implementing wellness programs and incentivizing policyholder participation through premium reductions can enhance customer engagement. Furthermore, tailoring communication strategies to resonate with specific demographic segments ensures a more personalized approach, illustrating how insurance plans align with unique health profiles and future needs. This proactive stance positions marketing managers as key players in driving positive outcomes in the insurance landscape. Overall, the knowledge gained from this study can be translated not only to insurance companies, but to policyholders with the hopes of guiding them towards healthier lifestyle landscapes.

References

1. Final 2024 payment rule, part 2: Risk adjustment - health affairs. Accessed December 12, 2023. https://www.healthaffairs.org/content/forefront/final-2024-payment-rule-part-2-risk-adjustment.
2. “How Health Insurance Marketplace® Plans Set Your Premiums.” How Health Insurance Marketplace® Plans Set Your Premiums | HealthCare.gov. Accessed December 12, 2023. https://www.healthcare.gov/how-plans-set-your-premiums/.