Table of Contents

[Linear Regression Project for Medical Insurance Forecast 1](#_Toc100925369)

[Introduction: 1](#_Toc100925370)

[Exploratory Data Analysis: 1](#_Toc100925371)

[Regression Analysis: 1](#_Toc100925372)

[Discussion: 2](#_Toc100925373)

[Assumption of Linear Regression: 3](#_Toc100925374)

[F-Test: 4](#_Toc100925375)

[Multicollinearity: 4](#_Toc100925376)

[Leverages, Outliers, and influential points: 4](#_Toc100925377)

[Prediction Using Model: 4](#_Toc100925378)

[Limitations: 5](#_Toc100925379)

[Conclusions 5](#_Toc100925380)

# **Linear Regression Project for Medical Insurance Forecast**

## **Introduction:**

To earn a profit, insurance firms must calculate premiums based on demographic trends, despite knowing little about the insured population. Our goal is to predict the amount needed to pay for medical insurance based on a dataset with row 1338 and the following details.

|  |  |  |
| --- | --- | --- |
| Variable Name | Variable Type | Description |
| age | Independent | Age of the insured |
| Sex | Independent/Categorical | Gender of the insured |
| BMI | Independent | BMI of the insured |
| Children | Independent/Categorical | Number of children |
| Region | Independent/Categorical | Region where the insured belong |
| Charges | Dependent | Total charge to be paid insured |
| Smoker | Independent/Categorical | If the Insured is smoking or not |

Table 1 Variable Type

## **Exploratory Data Analysis:**

There are three categorical variables in the data, and observing the data, we find one pattern: when the number of children is high, the smoking nature of the parents is less likely to be present.

Graphical user interface, text

Description automatically generated

Figure 1 Exploratory Data Analysis

## **Regression Analysis:**

Our first step in the regression analysis was to convert the categorical variables (Sex, Region, children, smokers) we have in the model using the function available in R:

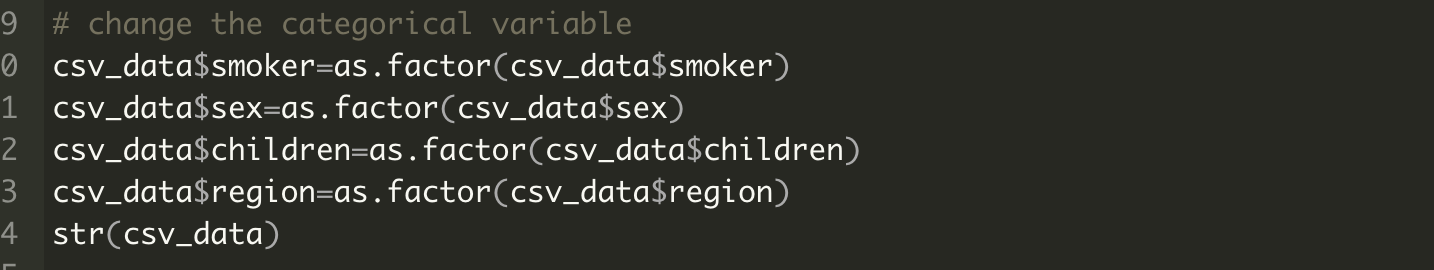


Figure 2- Changing Categorical Variables

And then we finally set up our model using the function available in R:

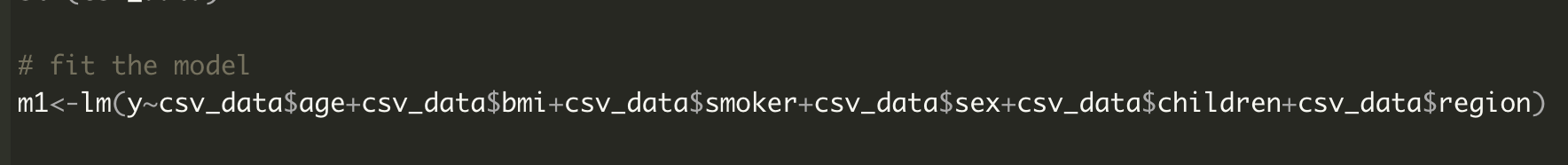


Figure 3 Fitting the Model for Data

On checking the summary of the model, we have the following result:

The R-squared value of 0.7516 shows that there is a significant relationship between the variables. The variables age, BMI, smokers, and children 2 are strongly contributing to the charges, yet regions and children 4 and children 5 are not significantly contributing, which may not be true, so we need to check and use some diagnostic tests for that section.

Also, we have found the following regression model with coefficients.

**charges= -11927.17 + 257.19(age) + 336.91(bmi) + 23836(smoker-yes) +1635.78(children2)**

## **Discussion:**

For the model we found the above we have checked different coefficients and found the following result:

Text

Description automatically generated

Figure 4 Summary of The Initial Model

**Assumption of Linear Regression:**  
In addition to this we also checked the assumption of the Linear Regression and found the following results:

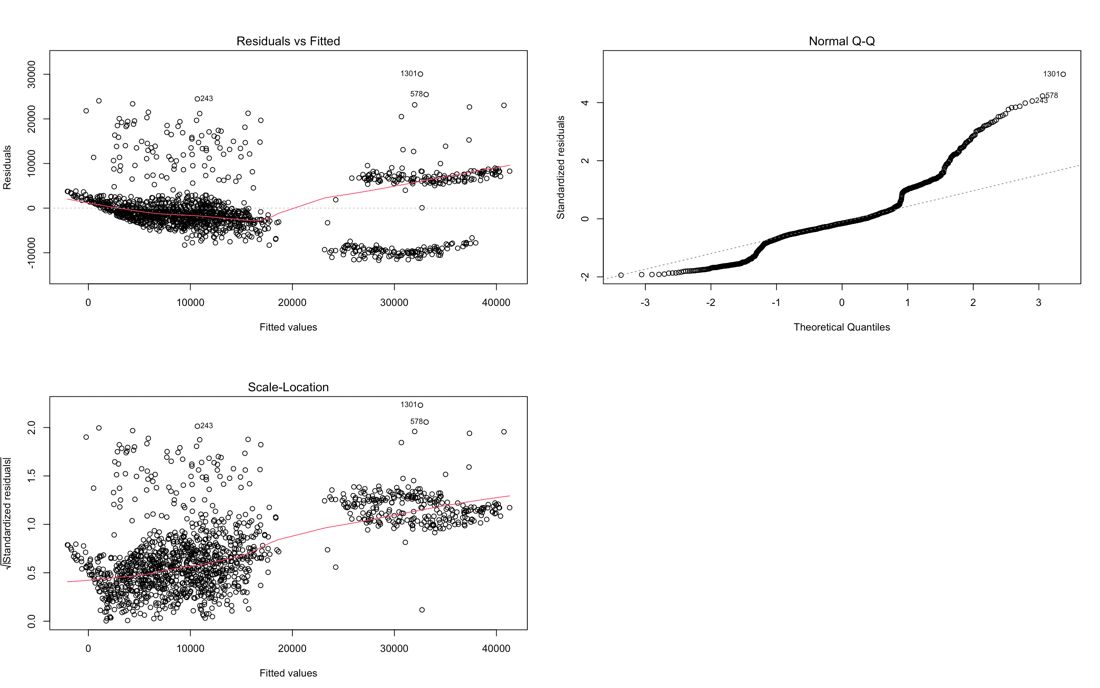


Figure 5 Assumptions of the Linear Regression

The plot shows that linearity and normality are not violated extremely whereas the graph also shows that the data are heteroscedastic. For this, we try to move it using logarithmic method on dependent variables and we have following model:

**log(charges)= 6.9692685+0.0348147\*age+0.0104155\*bmi+**

**(1.5437176)smoker+children1\*0.1423780+children2\*0.2792741+children3\*0.2494404+**

**(children4\*0.5197718)+children5\*0.3979236**

With following summary:

Graphical user interface, text

Description automatically generated

Figure 6 Summary of Model with no Heteroscedasticity

As a result of this model, the r-squared value for the model has increased by around 1 %, and all other variables are now equally accounted for in the model to predict the charges. Heteroscedasticity has also been removed from the model.

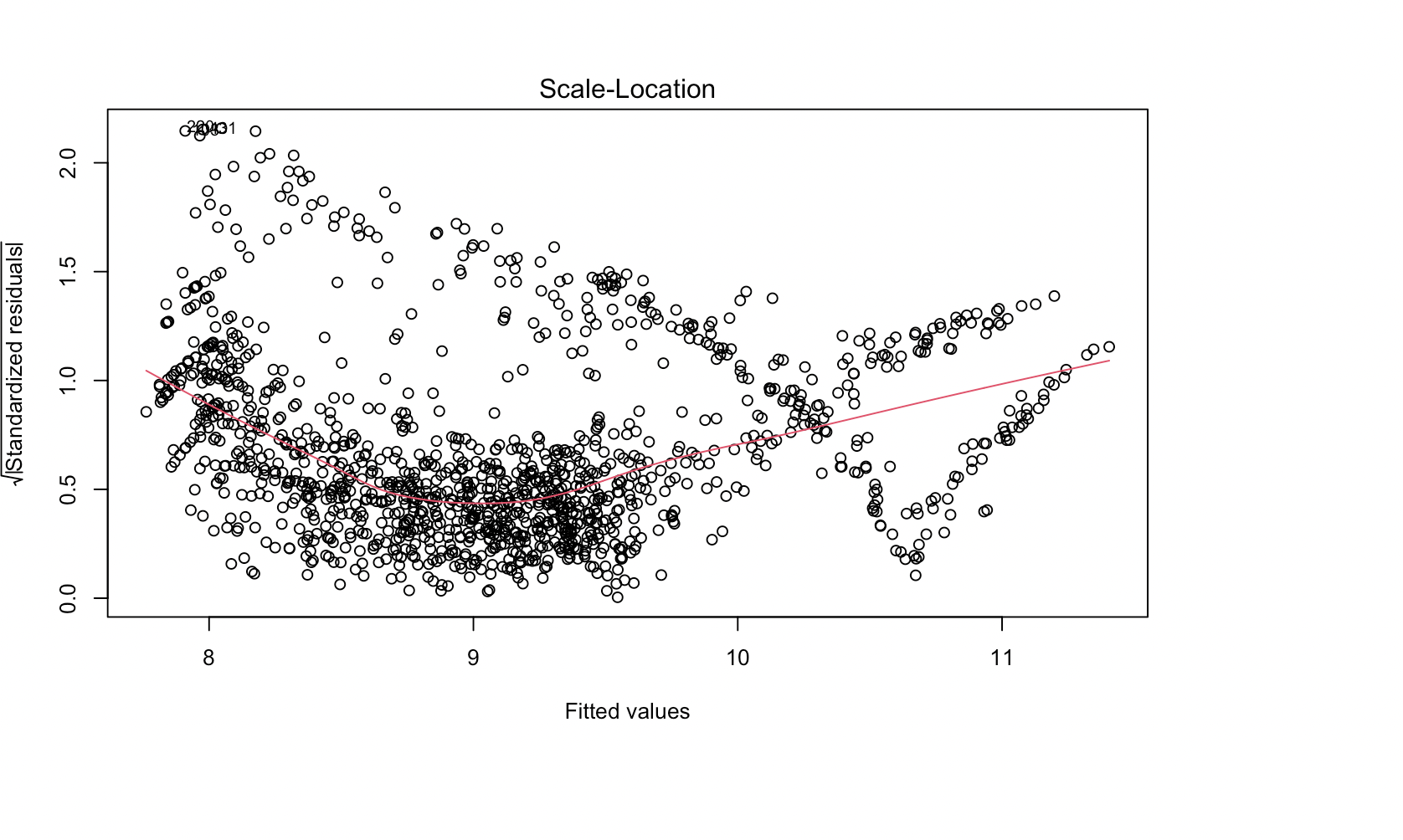


Figure 7 Heteroscedasticity Removed

After Heteroscedasticity removed from the model, we setup the hypothesis testing for the model with 97.5 confident level and with 1333 degree of freedom.

### **F-Test:**

H0: β=0, β1=0, β2=0 (the slope is equal to zero)

H1: β¹0(the slope is not equal to zero)

At the given condition the tabulated value for the F-value is 1.961745 whereas the value from the calculation is 828.82 which is greater than the given value at the 97.5% confident level, therefore the null hypothesis is rejected, and the alternate hypothesis is accepted. This means that the slope is not equal to zero.

### **Multicollinearity:**

For the multicollinearity we’ve tested using VIF method which show there is no multicollinearity in between the variable.

### **Leverages, Outliers, and influential points:**

In our testing for Leverages, we discovered that the dataset is not free with leverage points. About 39 rows contain leverage values. Furthermore, the data set contains some outlier values, which may affect the model and the value we predict. Some of these values have been confirmed to have an impact on our model.

### **Prediction Using Model:**

We can use data with different criteria for which we've already fitted the model to predict values in the future based on the different criteria.

For example:

Prediction value for the following data is:

**age=27, bmi=35.06, smoker="yes”, children="3" is 10.06759 which is e­­10.06759 =23566.700526888** as we’ve use log for predicted value to remove the heteroscedasticity in the data.

Other Sample Data we used for the prediction is as follows:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SN | Age | BMI | Smoker | Children | Predicted Value | Actual Predicted Value |
| 1 | 35 | 22.09 | Yes | 2 | 10.240855 | 28025.077126578 |
| 2 | 25 | 28.90 | No | 4 | 8.660417 | 5769.9402259822 |
| 3 | 36 | 32.60 | Yes | 3 | 10.355303 | 31423.239197901 |
| 4 | 48 | 35.00 | No | 3 | 10.798077 | 48926.624715447 |

Table 2 Predicted value for some Data

## **Limitations:**

Based on the model we have, the R-value is 75.15%, which means that there is a significant relationship between variables, but there isn't a strong one. Another factor that affects the model is that there are some categorical variables (e.g., sex). In addition, there are outliers and leverage points in the dataset and the number of data points is limited.

## 

## **Conclusions:**

We have fitted a linear regression model to the dataset using R and predicted the values for different conditions using the model. The model has a precision of 76.44 %, which can be improved by adding additional datasets or by removing a few categorical variables from the dataset.