Health insurance forecasting based on multiple linear regression

Jinwei Zhuang

Introduction

Purpose:

The cost of health insurance has always been an important concern for individuals, families and policy makers. Understanding the factors that contribute to these costs can help us make informed decisions about choosing insurance plans, pricing strategies, and health care policy reform. The report aims to identify key variables affecting health insurance costs and provide insight into the drivers behind cost changes.

Methodology:

The report uses multiple linear regression as the main method to study the relationship between various independent variables and health insurance costs. It provides greater insight into how different factors affect health insurance costs and provides tools to predict future costs based on individual characteristics. The method involves collecting a comprehensive data set of socioeconomic factors related to personal attributes, including age, sex, bmi, child, smoker and charges. The data set "Medical Cost Personal Datasets" [1] contains

valuable information for analyzing insurance trends. All in all, the report aims to enhance understanding of cost projections and contribute to more effective.

Exploratory data analysis

We will conduct exploratory data analysis on health insurance data sets. We will describe each variable, visualize their distribution, and evaluate the relationship between the variables.

1. Variable description

The target variable here is charges and remaining five variables such as age, sex, bmi, children, smoker are independent variable. We will use multiple linear regression to fit these data. Assuming the following:

$$y_i = \beta_0 + \beta_1 * x_{i1} + \beta_2 * x_{i2} + \beta_3 * x_{i3} + \beta_4 * x_{i4} + \beta_5 * x_{i5} + \epsilon$$

Where:

x_1 Age: The age of the individual.

x_2 Sex: The sex of the individual (male or female).

x_3 BMI: Body mass index, a measure of body fat based on height and weight.

x_4 Children: Number of children/dependents covered by insurance.

x_5 Smokers: Whether the individual smokes (yes or no).

y Charges: Individual health insurance costs.

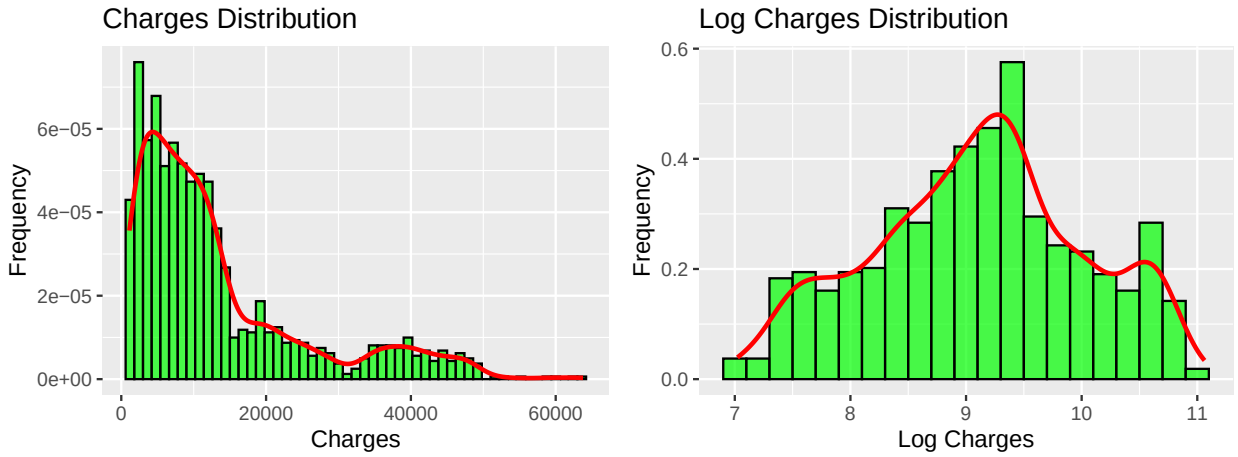
β_0 Intercept

$\beta_i, i = 1, 2, 3, 4, 5$ The influence of corresponding independent variable x_i on dependent variable y_i

ϵ Error term

2. Determine the distribution of charge

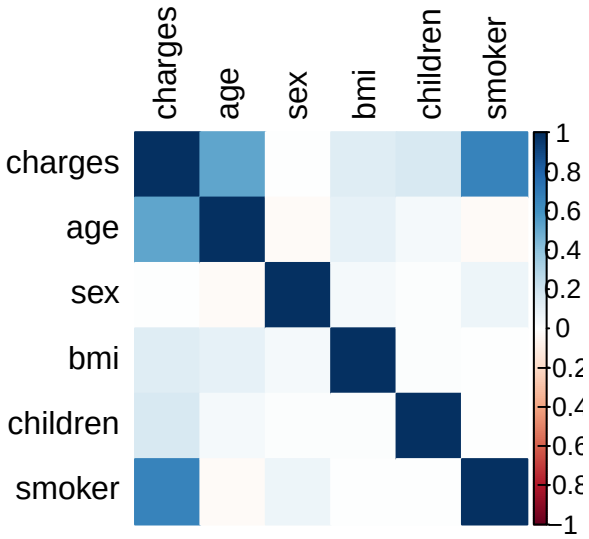
The left plot is right skewed. In right plot we apply natural log, then plot approximately tends to normal. We can consider using the log-transformed data in our analysis for modeling and inference. This may help to improve the degree to which the assumptions of the model are satisfied, thus improving the reliability and interpretability of the model.



3. Label Encoding

We need to process non-numerical changes in the data. Turn words into numeric values for fitting operations. Here we will use binary numbers (0 and 1) to represent male and female, and non-smoking and smoking states.

4. Label Encoding



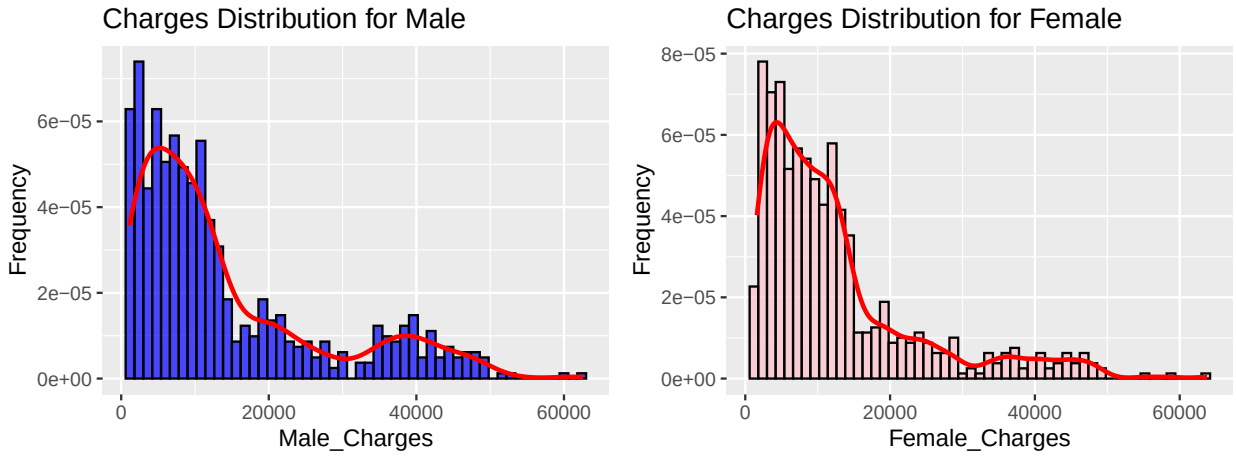
As can be seen from the figure:

1. The correlation coefficient between independent variables is small, which may indicate that there is no strong linear correlation between these independent variables, that is, they are not inclined to change together in value.

2. They are relatively independent when changing and are not affected by strong common changes, which helps to avoid multicollinearity problems in multiple regression analysis.

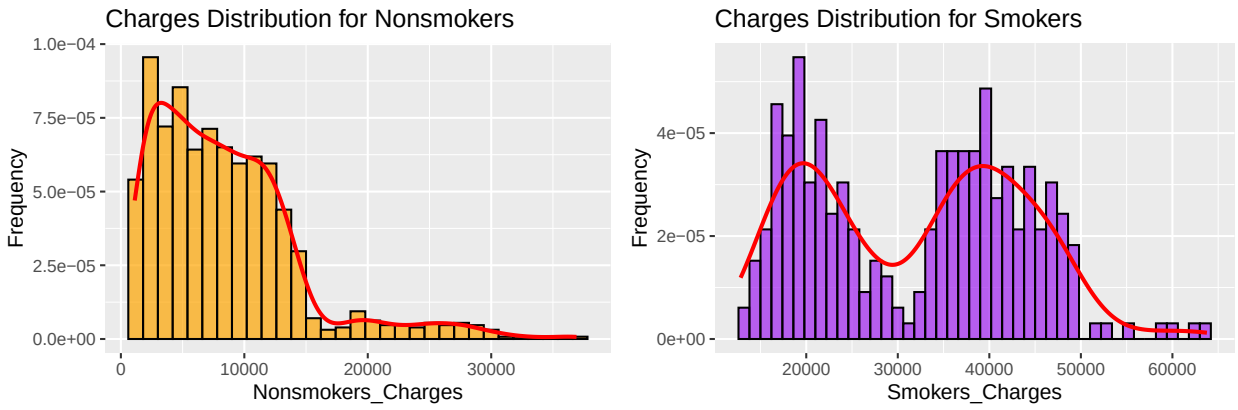
3. Different independent variables provide different information without excessive redundancy.

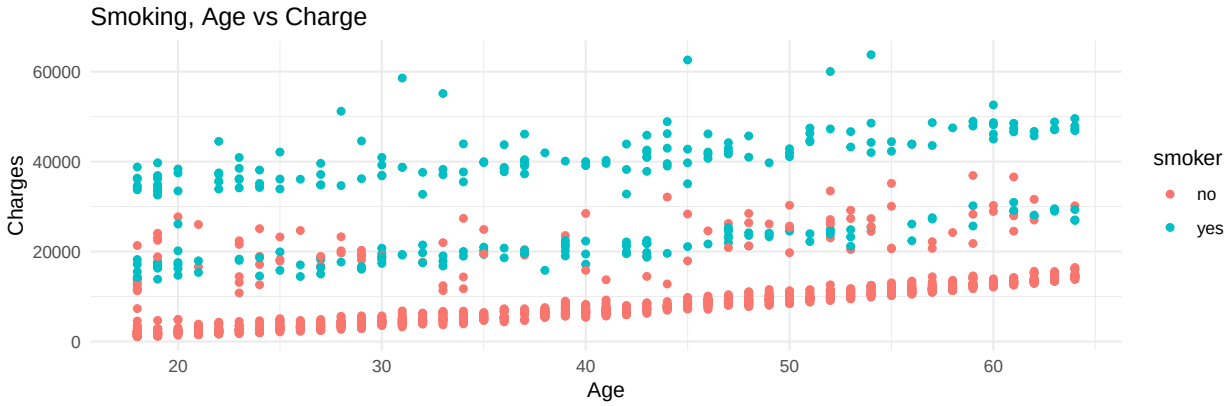
5. The effect of sex on charges

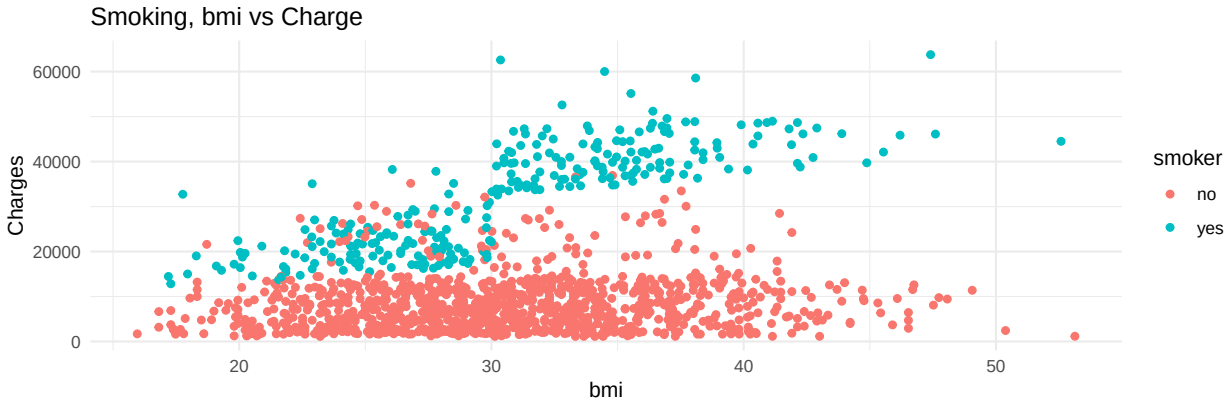


According to the histogram drawn, the distribution of male and female insurance costs is analyzed. The insurance costs of male and female are roughly in the same range, and their distribution curves are similar. It can be concluded that gender may not have much influence on insurance costs.

6. The effect of smoke, bmi and age on charges







According to the histogram and scatter plot, the insurance cost distribution of non-smokers and smokers is analyzed. The insurance cost of smokers is significantly higher than that of non-smokers, and their distribution curves are not similar. It can be concluded that whether smoking is an important factor in insurance cost.

In addition, the age scatter plot shows that with the increase of age, the insurance cost also has a significant upward trend, which indicates that age is also an important factor in the insurance cost.

According to the bmi scatter plot, it can also be found that for smokers, the larger the bmi (that is, the fatter they are), the higher the health insurance costs they bear. Looking at the whole picture, the greater the bmi, the higher the cost of health insurance.

Model Development

Generate model

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker, data = insurance_dummy)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.08241 -0.20315 -0.05185  0.07057  2.11173
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.0121103  0.0701685  99.932  < 2e-16 ***
## age          0.0347158  0.0008781  39.536  < 2e-16 ***
## sex         -0.0750088  0.0245899  -3.050  0.00233 **
## bmi          0.0109087  0.0020225   5.394 8.16e-08 ***
## children     0.1017275  0.0101688  10.004  < 2e-16 ***
## smoker       1.5502366  0.0304293  50.946  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4477 on 1332 degrees of freedom
```

```
## Multiple R-squared:  0.7638, Adjusted R-squared:  0.7629
```

```
## F-statistic: 861.5 on 5 and 1332 DF,  p-value: < 2.2e-16
```

We use multiple regression analysis to get the fitting function

$$ln(\hat{y}) = 7.012 + 0.035 * x_1 - 0.075 * x_2 + 0.011 * x_3 + 0.101 * x_4 + 1.550 * x_5 + \epsilon$$

$$R^2 = 0.7638$$

$$F = 861.5$$

The P-values of the coefficients are all statistically significant. R^2=0.7638, indicating that the model can explain about 76.38% of insurance cost variability. Overall, the statistical significance and goodness of fit indicators of the model justify its use to predict insurance costs based on the selected variables.

Test and correction of multicollinearity

```
##      age      sex      bmi children   smoker
## 1.015129 1.008878 1.014578 1.002242 1.006457
```

VIF values of all independent variables are within the range of close to 1, which indicates that there is no obvious multicollinearity problem in the model.

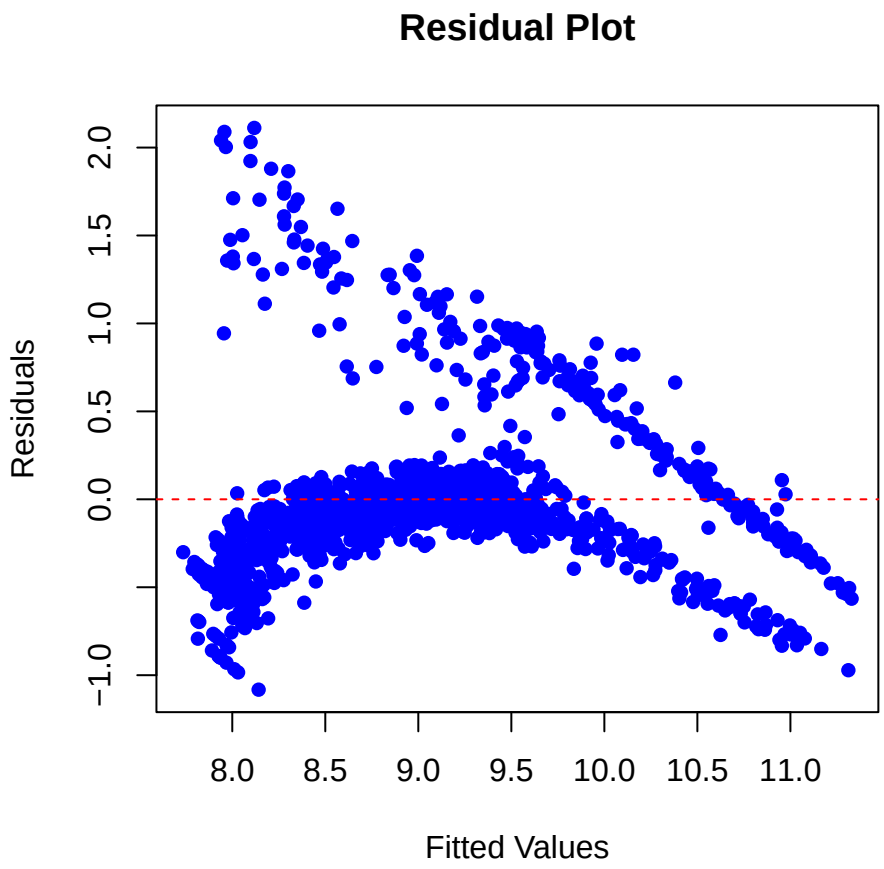
Autocorrelation test

```
##
##  Durbin-Watson test
##
## data:  multi.fit
## DW = 2.0519, p-value = 0.8288
## alternative hypothesis: true autocorrelation is greater than 0
```

The Durbin-Watson statistic (DW) has a value of 2.0519 and a p-value of 0.8288. The value of the DW statistic is close to 2, while the p-value is high, which means that the probability of autocorrelation between the residuals is low in this particular model. The residuals may be independent and do not have a significant negative effect on the validity of the regression model.

Heteroscedasticity test

```
##
##  studentized Breusch-Pagan test
##
## data:  multi.fit
## BP = 79.147, df = 5, p-value = 1.266e-15
```



The points in residual plot are not evenly distributed, which indicates the existence of heteroscedasticity. The Breusch-Pagan test of multiple linear regression model is carried out. The test statistic (BP) is 79.147 with 5 degrees of freedom, and the p-value is very close to zero (1.266e-15). We will reject the null hypothesis and conclude that heteroscedasticity is a problem in this model.So we're going to perform Weighted Least Squares Regression.

```
##
## Call:
## lm(formula = log_charges ~ age + sex + log_bmi + children + smoker,
##     data = insurance, weights = weights)
```

```
##
## Weighted Residuals:
##       Min        1Q    Median        3Q       Max
## -0.034698 -0.006303 -0.002449  0.002374  0.110854
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.0693696  0.1904263  31.873  < 2e-16 ***
## age          0.0351324  0.0008749  40.156  < 2e-16 ***
## sexmale     -0.0795042  0.0241677  -3.290  0.00103 **
## log_bmi      0.3781385  0.0578989   6.531 9.27e-11 ***
## children     0.1020972  0.0100109  10.199  < 2e-16 ***
## smokeryes    1.4391541  0.0298512  48.211  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0153 on 1332 degrees of freedom
## Multiple R-squared:  0.7591, Adjusted R-squared:  0.7582
## F-statistic: 839.7 on 5 and 1332 DF,  p-value: < 2.2e-16

##
##  studentized Breusch-Pagan test
##
```

```
## data:  wls.fit
## BP = 0.29311, df = 5, p-value = 0.9978
```

Using Weighted Least Squares Regression, the Breusch-Pagan test performed here shows that the null hypothesis cannot be rejected based on the value of p-value (0.9978), which means that for a given level of significance, The data do not provide sufficient evidence for heteroscedasticity in the model's residuals.

Finally, we get fitting function:

$$ln(\hat{y}) = 6.06 + 0.035 * x_1 - 0.079 * x_2 + 0.378 * ln(x_3) + 0.102 * x_4 + 1.439 * x_5 + \epsilon$$

$$R^2 = 0.7591$$

$$F = 839.7$$

Multiple R-squared is 0.7591, which is a relatively high value, indicating that the model has a strong ability to interpret the data. The F statistic is 839.7, and the corresponding p-value is less than 2.2e-16, which indicates that the model as a whole is significant. In the end, we passed Multicollinearity test, Autocorrelation test, and Heteroscedasticity test. These results increase the reliability of model analysis and prediction results.

Conclusion

In the conclusion part, the results of the whole multiple regression analysis will be summarized and explained. We have developed a multiple regression model to predict an individual's health insurance charges. The model is based on a number of major explanatory variables, including age, sex, bmi, number of children, and whether or not they smoke.

This model has wide application value in the real world. Through this model, we can better understand how insurance costs are affected by factors such as age, gender, BMI, etc. This helps insurance companies price more accurately and provide customers with more reasonable insurance costs. In addition, we can use the model for forecasting to predict the cost of insurance needed by individuals with different characteristics, helping clients make more informed decisions. Even though we built a predictive model, there are still some limitations. For example, a model may perform poorly in a particular subset. In further research, consider introducing more explanatory variables or trying more complex models to capture more data variability and optimize the model's predictive performance.

Reference

1. Kaggle. (2018). "Medical Cost Personal Datasets" Retrieved August 13, 2023, from https://www.kaggle.com/datasets/mirichoi0218/insurance

Appendix

```r
insurance=read.csv('insurance.csv', header=TRUE)

attach(insurance)

age=insurance$age

sex=insurance$sex

bmi=insurance$bmi

children=insurance$children

smoker=insurance$smoker

charges=insurance$charges

log_charges <- log(charges)

library(ggplot2)

library(cowplot)

plot1 <- ggplot(insurance, aes(x = charges, y = after_stat(density))) +

  geom_histogram(binwidth = 1200, fill = "green", color = "black", alpha = 0.7) +

  geom_density(color = "red", alpha = 1, linewidth = 1) +

  labs(title = "Charges Distribution", x = "Charges", y = "Frequency")

plot2 <- ggplot(insurance, aes(x = log_charges, y = after_stat(density))) +

  geom_histogram(binwidth = 0.2, fill = "green", color = "black", alpha = 0.7) +

  geom_density(color = "red", alpha = 1, linewidth = 1) +

  labs(title = "Log Charges Distribution", x = "Log Charges", y = "Frequency")


plot_grid(plot1, plot2, nrow = 1)

sex_dummy <- as.numeric(insurance$sex == "male")
```

```r
smoker_dummy <- as.numeric(insurance$smoker == "yes")

insurance_dummy <- data.frame(

  charges = log_charges,

  age = age,

  sex = sex_dummy,

  bmi = bmi,

  children = children,

  smoker = smoker_dummy)

cor_matrix <- cor(insurance_dummy)

corrplot(cor_matrix, method = "color", tl.col = "black")



male_charges <- insurance$charges[insurance$sex == "male"]

female_charges <- insurance$charges[insurance$sex == "female"]

plot_male <- ggplot(data = data.frame(charges = male_charges), aes(x = charges, y = after_stat

  geom_histogram(binwidth = 1200, fill = "blue", color = "black", alpha = 0.7) +

  geom_density(color = "red", alpha = 1, linewidth = 1) +

  labs(title = "Charges Distribution for Male", x = "Male_Charges", y = "Frequency")

plot_female <- ggplot(data = data.frame(charges = female_charges), aes(x = charges, y = after_

  geom_histogram(binwidth = 1200, fill = "pink", color = "black", alpha = 0.7) +

  geom_density(color = "red", alpha = 1, linewidth = 1) +

  labs(title = "Charges Distribution for Female", x = "Female_Charges", y = "Frequency")

plot_grid(plot_male, plot_female, nrow = 1)



nonsmoke_charges <- insurance$charges[insurance$smoker == "no"]
```

```r
smoke_charges <- insurance$charges[insurance$smoker == "yes"]


plot_nonsmoker <- ggplot(data = data.frame(charges = nonsmoke_charges), aes(x = charges, y = a
  geom_histogram(binwidth = 1200, fill = "orange", color = "black", alpha = 0.7) +
  geom_density(color = "red", alpha = 1, linewidth = 1) +
  labs(title = "Charges Distribution for Nonsmokers", x = "Nonsmokers_Charges", y = "Frequency


plot_smoker <- ggplot(data = data.frame(charges = smoke_charges), aes(x = charges, y = after_st
  geom_histogram(binwidth = 1200, fill = "purple", color = "black", alpha = 0.7) +
  geom_density(color = "red", alpha = 1, linewidth = 1) +
  labs(title = "Charges Distribution for Smokers", x = "Smokers_Charges", y = "Frequency")
plot_grid(plot_nonsmoker, plot_smoker, nrow = 1)
ggplot(insurance, aes(x = age, y = charges, color = smoker)) +
  geom_point() +
  labs(title = "Smoking, Age vs Charge", x = "Age", y = "Charges") +
  theme_minimal()
ggplot(insurance, aes(x = bmi, y = charges, color = smoker)) +
  geom_point() +
  labs(title = "Smoking, bmi vs Charge", x = "bmi", y = "Charges") +
  theme_minimal()


multi.fit = lm(charges~age + sex + bmi + children + smoker, data=insurance_dummy)
summary(multi.fit)
vif_values <- vif(multi.fit)
```

```r
print(vif_values)



library(lmtest)

dwtest(multi.fit)

bptest(multi.fit)

residuals <- residuals(multi.fit)

fitted_values <- fitted(multi.fit)

plot(fitted_values, residuals,

     main = "Residual Plot",

     xlab = "Fitted Values",

     ylab = "Residuals",

     col = "blue",

     pch = 16)

abline(h = 0, col = "red", lty = 2)

library(lmtest)

insurance$log_bmi <- log(insurance$bmi)

weights <- 1 / (insurance$bmi^2)

wls.fit <- lm(log_charges ~ age + sex + log_bmi + children + smoker, data = insurance, weights

summary(wls.fit)

bptest(wls.fit)

residuals <- residuals(wls.fit)

fitted_values <- fitted(wls.fit)
```