Name: Anupam Vanjari

Linear Model

Problem Statement:

To develop a simple linear regression model and to predict the relationship between the current premium and effect of fixed expenses on it.

Objective:

To establish a reliable relationship between the fixed expenses of the insurance company and its respective current premiums, to help insurance companies better understand their own pricing models in benefit of both the company and its customers.

Here, X: Fixed expenses (Independent variable)

Y: Current premium (Dependent variable)

Dataset Link: https://www.kaggle.com/datasets/thedevastator/insurance-companies-secret-sauce-finally-exposed?select=cgr-premiums-table.csv

The Dataset contains insurance rates data from across the United States, providing insights into the premiums charged by insurers, the underlying factors that affect those rates. It includes information on premiums, underlying factors, current premium prices, selected premiums prices, indicated premium prices, fixed expenses and more.

Pre-analysis using Excel: https://docs.google.com/spreadsheets/d/1-ZClgvmWqIVYgb0AIxDlkzMwQsgAuobi/edit?usp=sharing&ouid=107511735992376290897&rtpof=true&sd=true

R-code:

```
y=final.data$current_premium

x=final.data$fixed_expenses

model=lm(y~x)

summary(model)

plot(model)

shapiro.test(y)

# Create an index for data partitioning

install.packages("caret")
```

```
library(caret)

#Set a seed for reproducibility
Set.seed(123)

# Create an index for data partitioning
index <- sample(1:nrow(final.data), 0.8 * nrow(final.data))
train_data <- final.data[index, ]
train_data

# Create testing set
test_data <- final.data[-index, ]
test_data

# Fit the simple linear regression model using the training data
model1 <- lm(y ~ x, final.data = train_data)
model1
summary(model1)

# Make predictions on the test set
predictions <- predict(model1, newdata = test_data)
predictions

# Calculate Mean Squared Error
# Example for regression
mse <- mean(( final.data$current_premium - predictions)^2)
mse
```

Output:

lm(formula = y ~ x, final.data = train_data)

Residuals:

Min 1Q Median 3Q Max

-870.8 -434.6 -157.1 263.4 4744.1

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 643.8358 76.1819 8.451 < 2e-16 ***

x 3.1319 0.4731 6.621 5.85e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 622.3 on 994 degrees of freedom

Multiple R-squared: 0.04223, Adjusted R-squared: 0.04127

F-statistic: 43.83 on 1 and 994 DF, p-value: 5.845e-11

Shapiro-Wilk normality test

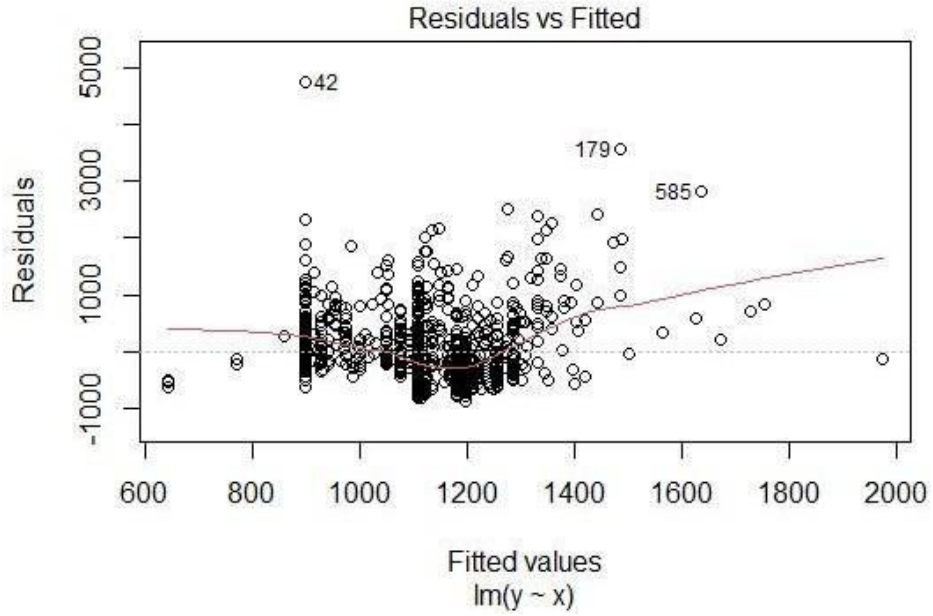
data: y

W = 0.85108, p-value < 2.2e-16

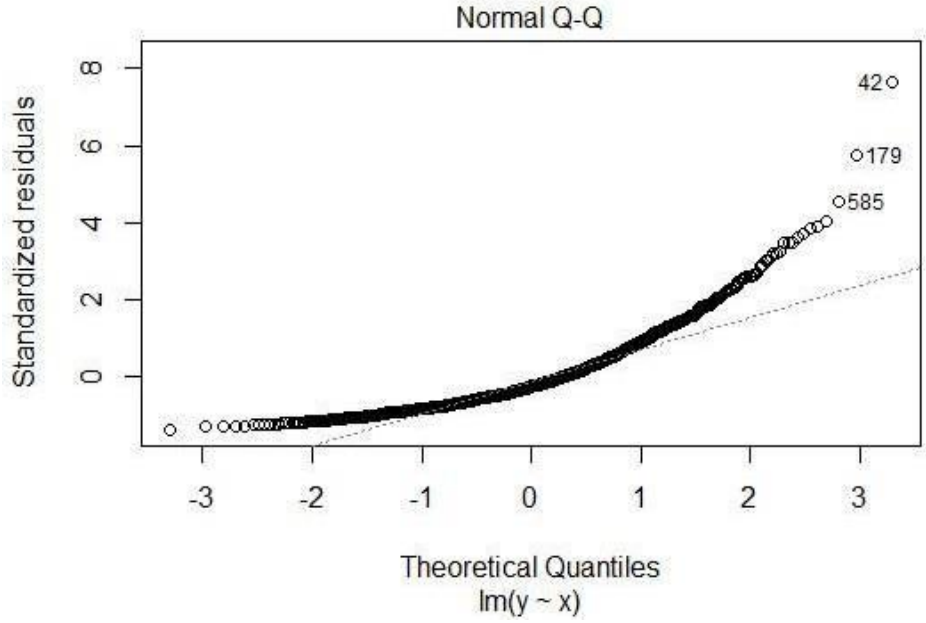
Predicted Value Link:

https://docs.google.com/spreadsheets/d/1S0AlZaOA5yenmwj4nuWG5Nr7HkQmoWVd/edit?usp=sharing&ouid=107511735992376290897&rtpof=true&sd=true

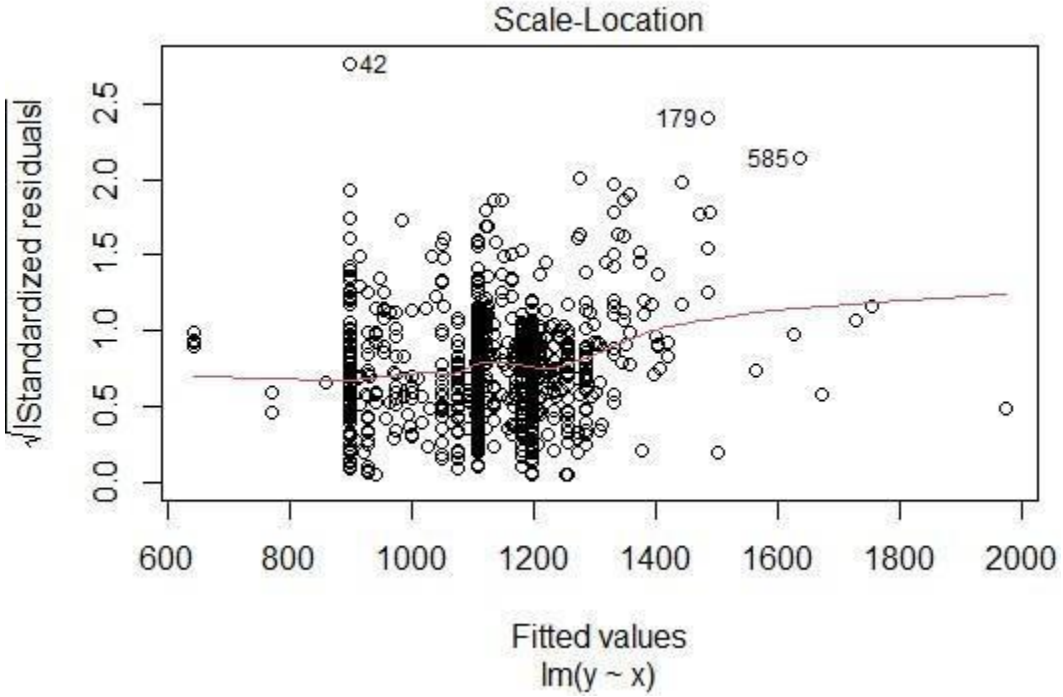
Plots:



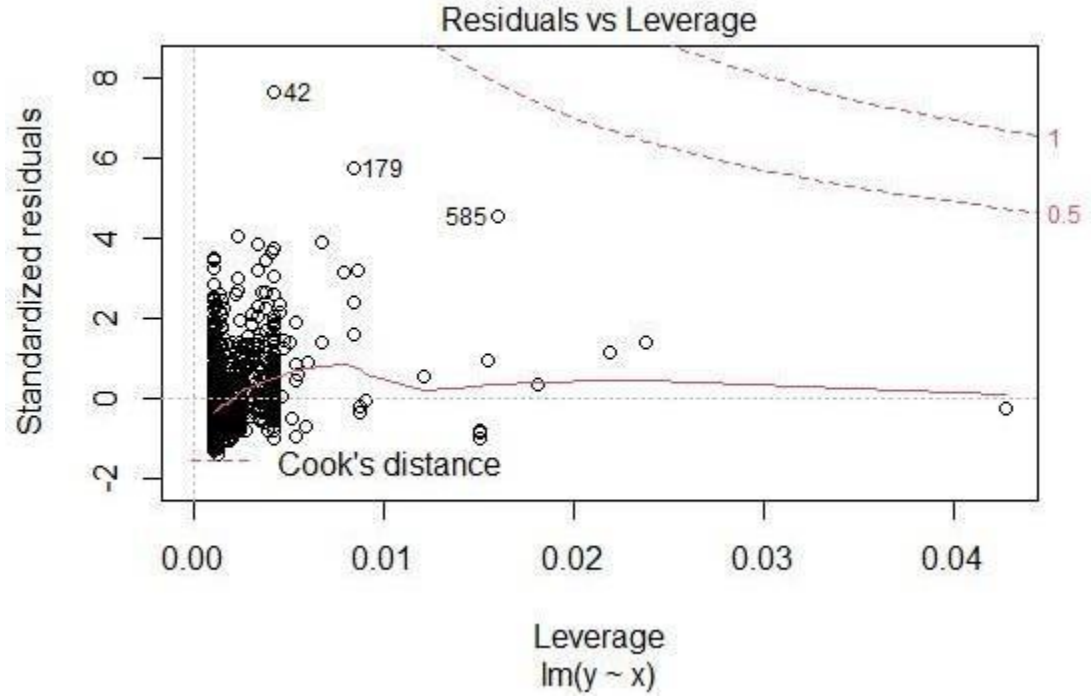
Interpretation: The model assumptions are accurate and the model is performing satisfactorily in terms of variability in data.



Interpretation: The distribution of the data closely follows normal distribution.



Interpretation: The spread of the residual is relatively constant and there is homoscedasticity in the data.



Interpretation: The model is less sensitive to outliers.

Mean Square Error:

```
> # Calculate Mean Squared Error
> # Example for regression
> mse <- mean(( final.data$current_premium - predictions)^2)
> mse
[1] 386464.5
```

Output:

The regression model suggests that there is a weak relationship between fixed expenses and current premiums having MSE value as 386464.5 which indicates that actual value and predicted value are far away from each other.

Also, R-square value is 0.04223 which indicates that 4.223% model is accurate.

In conclusion, the model is not a good fit and it is not dependable for estimating fixed expenses using current premium.