# **Introduction**

The prediction of insurance charges plays a crucial role in the healthcare industry, as it helps individuals and insurance providers estimate the potential costs associated with various medical conditions. In this project, this study aim to explore the relationship between insurance charges and several predictors, including age, sex, body mass index (BMI), number of children, smoker status, and region. By utilizing simple and multiple linear regression techniques, this study intend to develop a model that can accurately predict insurance charges based on these factors.

The dataset used for this analysis comprises 1338 rows, each representing an individual's insurance profile. By leveraging this comprehensive dataset, this study can gain insights into the influence of different variables on insurance charges. This project holds significance as it can contribute to improved pricing strategies for insurance providers, better financial planning for individuals, and a deeper understanding of the factors that drive healthcare costs.

The predictors chosen for this analysis cover a range of demographic and lifestyle factors. Age, as a predictor, may reflect the influence of healthcare needs and potential chronic conditions that are more prevalent in certain age groups. Gender, as another predictor, can help us explore potential differences in healthcare costs between males and females. Body mass index (BMI) serves as an indicator of overall health and can provide valuable insights into the relationship between weight and insurance charges. The number of children may influence insurance charges, as dependents often require additional coverage. Smoker status is a vital predictor due to the substantial impact of smoking on health and associated medical expenses. Lastly, the region in which individuals reside may reveal variations in healthcare costs based on geographic factors such as availability of medical facilities and cost of living.

This project report will apply simple and multiple linear regression techniques to build predictive models that can estimate insurance charges based on the aforementioned predictors. The accuracy of the models will be evaluated using appropriate metrics, and the significance and interpretability of the predictor variables will be assessed. The insights gained from this analysis can have implications for insurance companies, policymakers, and individuals seeking insurance coverage, ultimately contributing to a more informed and data-driven decision-making process in the healthcare domain.

# **Part 1 – Simple Linear Regression**

This section will focus on utilizing simple linear regression to predict insurance charges (*Y*) based solely on an individual's age (*X*). Age is a fundamental demographic factor that has long been recognized as a potential predictor of healthcare costs. By examining the relationship between age and insurance charges, this study aim to uncover insights that can enhance our understanding of how age influences medical expenses and aid in accurate predictions.

Age is a critical variable as it reflects the passage of time and the progression of an individual's lifespan. It is reasonable to assume that as people age, they are more likely to require medical attention and incur higher healthcare costs. This is due to various factors such as the increased prevalence of chronic conditions, age-related diseases, and the need for specialized healthcare services. By examining the relationship between age and insurance charges, we can assess the extent to which age serves as a reliable predictor of medical expenses.

## **Mean and standard deviation for both variables**

The dataset contains a fairly large number of samples. Firstly, the mean and standard deviation of both variables can be calculated using built in function in R as follows.

> mean(insurance$age)

[1] 39.20703

> sd(insurance$age)

[1] 14.04996

>

> mean(insurance$charges)

[1] 13270.42

> sd(insurance$charges)

[1] 12110.01

The mean for age is at 39.20703 years with standard deviation at 14.04996 years. While for the charges, the mean is at $13,270.42 with standard deviation at $12,110.01.

## **Sample correlation**

Next, the correlation between the two variables can be calculated. Sample correlation, also known as the sample correlation coefficient or Pearson's correlation coefficient, measures the strength and direction of the linear relationship between two variables in the dataset. In this case, it will measure the strength and direction of the linear relationship between age and insurance charges.

> cor(insurance$age, insurance$charges)

[1] 0.2990082

A correlation coefficient of 0.299 indicates a moderate positive linear relationship between the two variables. It suggests that as age increases, insurance charge tends to increase as well, but the relationship is not very strong.

## **The slope and intercept for the least squares line and the prediction equation**

The analysis will continue by generating a simple linear regression model using R. The linear regression model will provide the value of slope and intercepts for the least squares line. This is also the coefficients for the prediction equation. The R result is shown as follows.

> simple\_model <- lm(charges ~ age, data = insurance)

> simple\_model

Call:

lm(formula = charges ~ age, data = insurance)

Coefficients:

(Intercept) age

3165.9 257.7

> round(summary(simple\_model)$coefficients, 3)

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

(Intercept) 3165.885 937.149 3.378 0.001

age 257.723 22.502 11.453 0.000

The coefficients section shows the estimated values for the intercept and the slope (or coefficient) associated with the predictor variable age. The intercept in the output, is the estimated value of the response variable (in this case, charges) when the predictor variable (age) is zero. In the given output, **the intercept is 3165.9**. However, it is important to note that interpreting the intercept in the context of age might not be meaningful, as age zero does not have a practical interpretation in this scenario.

The coefficient for age, represented as age in the output, is the estimated change in the response variable for each one-unit increase in the predictor variable. In this case, for every one-year increase in age, the charges are estimated to increase by $257.7. In other word**, the slope is 257.7.** Based on these numbers, **the prediction equation can be formulated as follows**.

Using three decimals, the equation is as follows.

The equation suggests that the model estimates an initial charge of 3165.9 when age is zero, but this value may not have practical significance. Furthermore, for each additional year of age, the charges are estimated to increase by 257.7 units. These coefficients allow for predictions and insights into the relationship between age and insurance charges, although it is essential to consider the limitations and context of the data when interpreting the results.

## **Sample predictions for any two data cases and the residuals**

To test the model, the study will choose two random samples from the dataset and calculate the predicted insurance charges for the samples using the model. Then, the predicted value will be compared to the actual observed value, which is also called as the residual. The process is as follows.

> # First sample

> # Generate random sample index

> sample\_1 <- sample(1:nrow(insurance),1)

> sample\_1

[1] 576

> # Obtain age of the random sample

> sample\_1\_age <- data.frame(age = insurance$age[sample\_1])

> sample\_1\_age

age

1 58

> # Calculate the predicted charges using the random sample age

> predicted\_value\_1 <- predict(simple\_model, newdata = sample\_1\_age)

> predicted\_value\_1

1

18113.8

> # The actual value for the random sample

> actual\_value\_1 <- insurance$charges[sample\_1]

> actual\_value\_1

[1] 12222.9

> # Residual is the difference between the actual value and the predicted value

> residual\_1 <- actual\_value\_1 - predicted\_value\_1

> residual\_1

1

-5890.899

The firs step is to generate the random sample index using sample function that will generate random index between 1 and 1338, which is the number of rows (nrow) of the insurance dataset. The generated index is 576. Then the age of the 576th sample is obtained, which is 58 years. Using the age, the model can compute the predicted insurance charges at $18,113.8. This value can also be computed manually using the prediction equation above.

The actual insurance charges observed for the sample is at $12,222.9. Thus, there are fairly large difference between the actual value and the predicted value. The difference is calculated as residual\_1, which is at -$5,890.9. Based on this sample, the predicted value overestimate the insurance charge.

The same process can be performed for the second sample. The calculation for the second sample is as follows.

> # Second sample

> # Generate random sample index

> sample\_2 <- sample(1:nrow(insurance),1)

> sample\_2

[1] 715

> # Obtain age of the random sample

> sample\_2\_age <- data.frame(age = insurance$age[sample\_2])

> sample\_2\_age

age

1 24

> # Calculate the predicted charges using the random sample age

> predicted\_value\_2 <- predict(simple\_model, newdata = sample\_2\_age)

> predicted\_value\_2

1

9351.228

> # The actual value for the random sample

> actual\_value\_2 <- insurance$charges[sample\_2]

> actual\_value\_2

[1] 2457.502

> # Residual is the difference between the actual value and the predicted value

> residual\_2 <- actual\_value\_2 - predicted\_value\_2

> residual\_2

1

-6893.726

The 715th data is the second random sample chosen. The sample’s age is 24. Plugging in the value to the prediction equation will result in the predicted charge value of $9,351.228, while the actual value for the sample is $2,457.502. Hence, the residual, or the difference in the predicted and actual value is $6,893.726. Similar to the previous sample, the predicted value overestimates the insurance charge.

## **Scatterplot with least squares line**

To better understand how the regression line is fitted on the data, the study can visualize the model using scatterplot. The scatterplot can be generated using ggplot library in R as follows.

> # Scatterplot with regression line

> insurance %>%

+ ggplot(aes(x=age,y=charges)) +

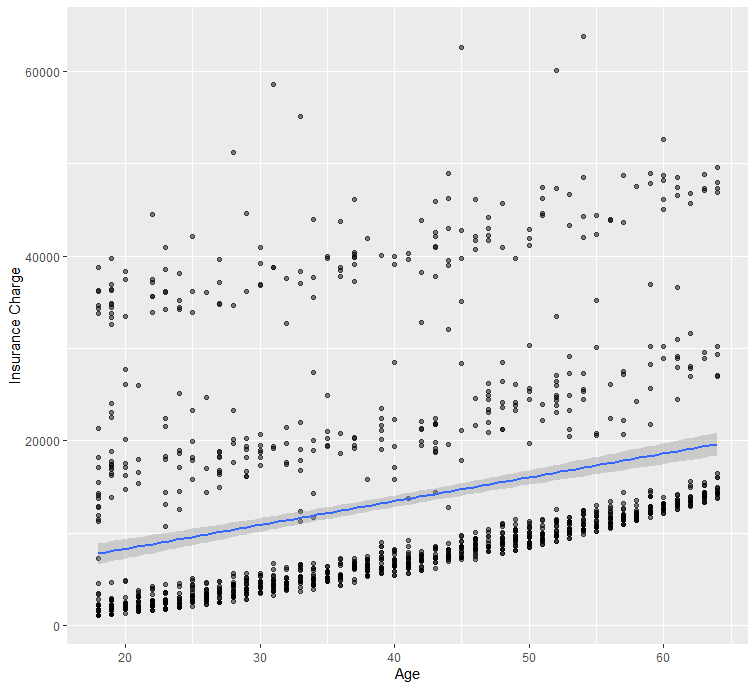
+ geom\_point(alpha=0.5) +

+ labs(x= "Age", y="Insurance Charge")+

+ geom\_smooth(method=lm)

`geom\_smooth()` using formula = 'y ~ x'

The code above results in the scatterplot below.



The scatterplot above shows that the difference between the line and the actual observed data can be quite large.

## **Test whether the slope is significantly different from zero**

In simple linear regression using R, the slope can be tested whether it is significantly different from zero by performing a hypothesis test. The most common test used is the t-test, which assesses the statistical significance of the slope coefficient. For this test, the null hypothesis, H0, is that the slope of this regression model does not significantly differ from zero, while the alternative hypothesis, HA, is that the slope does significantly differ from zero.

H0: β1 = 0

HA: β1 ≠ 0

The significance level used for this hypothesis test is α = 0.05 and the test p-value can be obtained using summary() function as follows.

> summary(simple\_model)

Call:

lm(formula = charges ~ age, data = insurance)

Residuals:

Min 1Q Median 3Q Max

-8059 -6671 -5939 5440 47829

Coefficients:

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

(Intercept) 3165.9 937.1 3.378 0.000751 \*\*\*

age 257.7 22.5 11.453 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11560 on 1336 degrees of freedom

Multiple R-squared: 0.08941, Adjusted R-squared: 0.08872

F-statistic: 131.2 on 1 and 1336 DF, p-value: < 2.2e-16

> round(summary(simple\_model)$coefficients, 3)

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

(Intercept) 3165.885 937.149 3.378 0.001

age 257.723 22.502 11.453 0.000

The "Estimate" column provides the estimated value of the slope as calculated earlier. The "Pr(>|t|)" column provides the p-value associated with the slope coefficient. To test whether the slope is significantly different from zero at a specific significance level (e.g., α = 0.05), the study can compare the p-value to the chosen significance level. If the p-value is less than the significance level (p-value < 0.05), it can be concluded that the slope is significantly different from zero.

In this case, the p-value for the slope coefficient is almost zero (lower than 2.10-16 or 0.000…) which is of course less than 0.0. It indicates that the slope is significantly different from zero. The \*\*\* symbols denote the level of significance (the more asterisks, the lower the p-value and higher the significance). It means that age variable is a significant variable.

## **R-squared**

The R-squared (or multiple R-squared) is a statistical measure that indicates the proportion of the variance in the response variable that can be explained by the predictor variables in a multiple linear regression model. It represents the goodness of fit of the regression model. The R-squared can be seen on the result of summary() function on the previous section.

Multiple R-squared: 0.08941

In this case, the R-squared is obtained at 0.08941. It suggests that approximately 8.941% of the variance in the response variable (insurance charges) can be explained by the predictor variable, age. It implies that there are many other factors that can explain the variation in insurance charges. It suggest that the study should probably explore more factors that can possibly affect the insurance charges.

## **Confidence interval for mean and prediction interval for individual values**

The confidence interval for mean and prediction interval will be calculated for the age sample of 58 years, which is calculated above (4). First, define the sample information, specifically the age and sample size for which the computation of the intervals will be calculated (age\_sample = 58 and sample\_size = number or rows of the insurance dataset = 1336).

> age\_sample <- 58

> sample\_size <- nrow(insurance)

Then, the standard error of the mean (SE\_mean) can be calculated. The residual standard error (RSE) is calculated by taking the square root of the sum of squared residuals divided by the degrees of freedom. The degrees of freedom are computed as the total number of observations minus the number of predictors (in this case, 2, including the intercept). Then, divide the RSE by the square root of the sample size to obtain the SE\_mean.

> # Compute the standard error of the mean

> RSE <- sqrt(sum(simple\_model$residuals^2) / (nrow(simple\_model$model) - 2))

> SE\_mean <- RSE / sqrt(sample\_size)

Next, set the desired confidence level for the confidence interval (e.g., 95%) by assigning it to the variable confidence\_level. Using the qt() function in R, the margin of error for the confidence interval and the prediction interval can be calculated. The qt() function provides the critical t-value corresponding to the desired confidence level and the degrees of freedom (sample\_size - 1).

> # Compute the confidence interval for the mean

> confidence\_level <- 0.95 # Set the desired confidence level (e.g., 95%)

> margin\_of\_error <- qt((1 - confidence\_level) / 2, df = sample\_size - 1) \* SE\_mean

For the confidence interval, the calculation will be to subtract and add the margin of error from the predicted\_value\_1 (which is calculated earlier) to obtain the lower and upper bounds of the interval, respectively. Then, store these values in the variable confidence\_interval.

> # Compute the confidence interval for the mean

> confidence\_level <- 0.95 # Set the desired confidence level (e.g., 95%)

> margin\_of\_error <- qt((1 - confidence\_level) / 2, df = sample\_size - 1) \* SE\_mean

> confidence\_interval <- c("fit : ", predicted\_value\_1, ", lower :", predicted\_value\_1 - margin\_of\_error, ", upper :", predicted\_value\_1 + margin\_of\_error)

For the prediction interval, calculate the standard error of prediction (SE\_prediction) by taking the square root of the sum of the squared RSE and the squared SE\_mean multiplied by (1 + 1/sample\_size). This accounts for the uncertainty in both the regression model and the estimation of the mean. Then, calculate the margin of error for the prediction interval using the same approach. Finally, compute the lower and upper bounds of the prediction interval and store them in the variable prediction\_interval.

> # Compute the prediction interval for an individual value

> SE\_prediction <- sqrt(RSE^2 + (SE\_mean^2 \* (1 + 1/sample\_size)))

> margin\_of\_error\_prediction <- qt((1 - confidence\_level) / 2, df = sample\_size - 1) \* SE\_prediction

> prediction\_interval <- c("fit : ", predicted\_value\_1, ", lower :", predicted\_value\_1 - margin\_of\_error\_prediction, ", upper :", predicted\_value\_1 + margin\_of\_error\_prediction)

The last two lines of code print the confidence interval for the mean and the prediction interval, respectively, using the cat() function.

> # Print the confidence interval for the mean and the prediction interval

> cat("Confidence Interval for the Mean:", "\n", confidence\_interval, "\n")

Confidence Interval for the Mean:

fit : 18113.796888743 ,

lower : 18733.7842023036 ,

upper : 17493.8095751824

> cat("Prediction Interval for Individual Values:", "\n", prediction\_interval)

Prediction Interval for Individual Values:

fit : 18113.796888743 ,

lower : 40800.5953284028 ,

upper : -4573.00155091676

The estimated mean insurance charge for individuals with an age of 58 years is approximately $18,113.80. The confidence interval for this mean is calculated as (lower bound: $17,493.81, upper bound: $18,733.78) at a 95% confidence level. This means that we are 95% confident that the true mean insurance charge for individuals with an age of 58 years falls within this range.

The predicted insurance charge for an individual with an age of 58 years is approximately $18,113.80. The prediction interval for this individual is calculated as (lower bound: $-4,573.00, upper bound: $40,800.60) at a 95% confidence level. This means that we are 95% confident that the true insurance charge for an individual with an age of 58 years lies within this prediction interval.

It's worth noting that the negative lower bound in the prediction interval is likely due to the limitations of the linear regression model in extrapolating beyond the observed range of the predictor variable. Therefore, it is advisable to interpret the prediction interval within the observed range of the data to ensure more reliable predictions.

## **Residual plot and test the assumptions of the regression model**

To create a residual plot in R, there are several steps to be performed. First create a data frame residual\_data containing the fitted values and residuals. The predict() function is used to obtain the predicted (fitted) values from the model.

Then, use ggplot() to create the plot, specifying Fitted as the x-axis variable and Residuals as the y-axis variable. The geom\_point() function adds points to the plot, and geom\_hline()at y = 0 to represent the zero residual line. The geom\_smooth() function is used to add the LOESS (Locally Estimated Scatterplot Smoothing) curve with a confidence band to the plot. The method = "loess" argument specifies that the LOESS smoothing method should be used. The se = TRUE argument requests the inclusion of the confidence band. Finally, set the x-axis label, y-axis label, and plot title using xlab(), ylab(), and ggtitle(), respectively. The code and the plot result is as follows.

> residuals <- residuals(simple\_model)

> residual\_data <- data.frame(Age = insurance$age, Residuals = residuals)

>

> # Create the residual plot using ggplot2

> ggplot(residual\_data, aes(x = Age, y = Residuals)) +

+ geom\_point() +

+ geom\_smooth(method = "loess", se = TRUE, color = "blue", fill = "lightblue") +

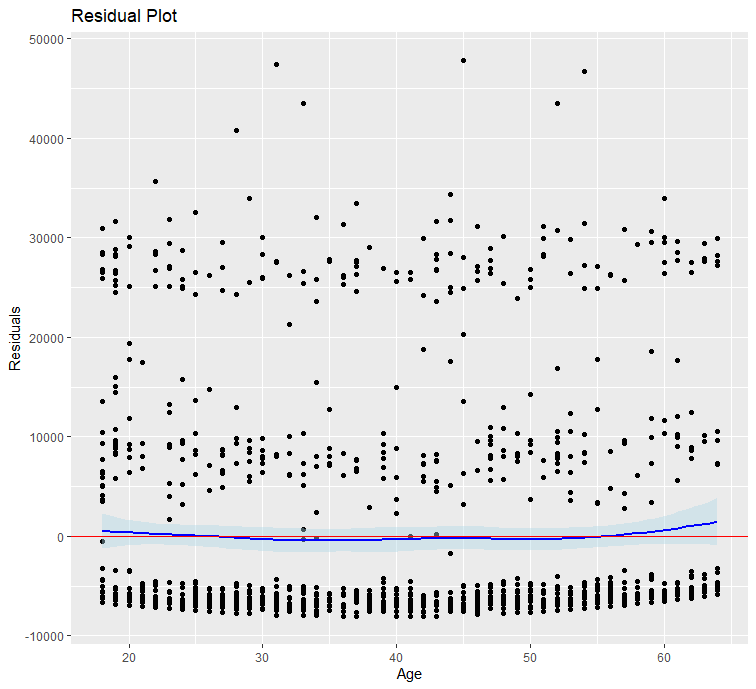
+ geom\_hline(yintercept = 0, color = "red") +

+ xlab("Age") +

+ ylab("Residuals") +

+ ggtitle("Residual Plot")

`geom\_smooth()` using formula = 'y ~ x'



This plot shows that the line of y=0 is contained within the confidence band of the LOESS curve. The spread of the points above zero is typically higher than the spread below zero, but the number of points is relatively equal, demonstrating that the linearity assumption is met. However, there may be non-random patterns that the residuals is concentrated at about -5,000, at between 5,000 and 10,000, as well as between 25,000 and 30,000. These patterns may indicate violations of the linearity assumption.

Aside from the linearity assumption, the regression model should be tested normality. To test for normality, the histogram of residuals can be generated using ggplot. The code and the histogram plot result is as follows.

> # Create a data frame with the residuals

> residual\_data\_2 <- data.frame(Residuals = residuals)

>

> # Create the histogram plot using ggplot2

> ggplot(residual\_data\_2, aes(x = Residuals)) +

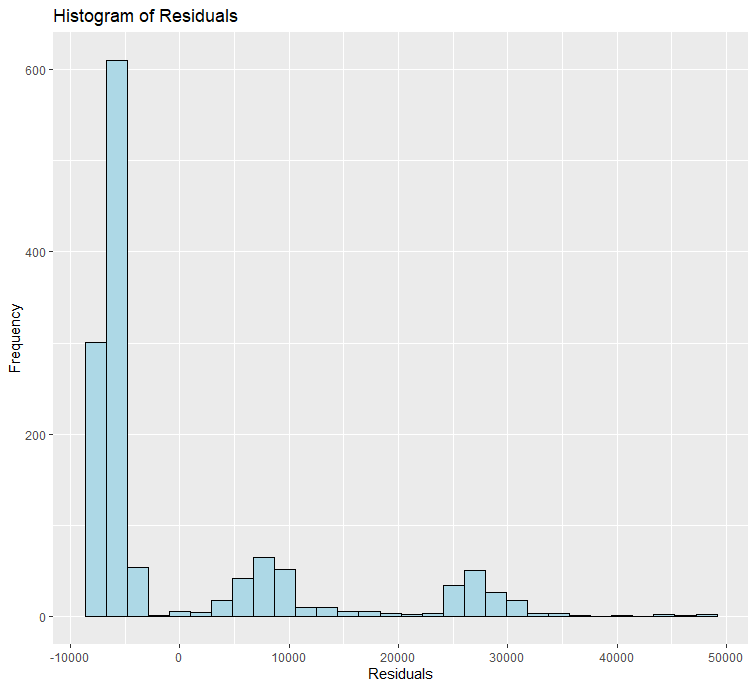
+ geom\_histogram(fill = "lightblue", color = "black") +

+ xlab("Residuals") +

+ ylab("Frequency") +

+ ggtitle("Histogram of Residuals")

`stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



> ggplot(residual\_data\_2, aes(sample = Residuals)) +

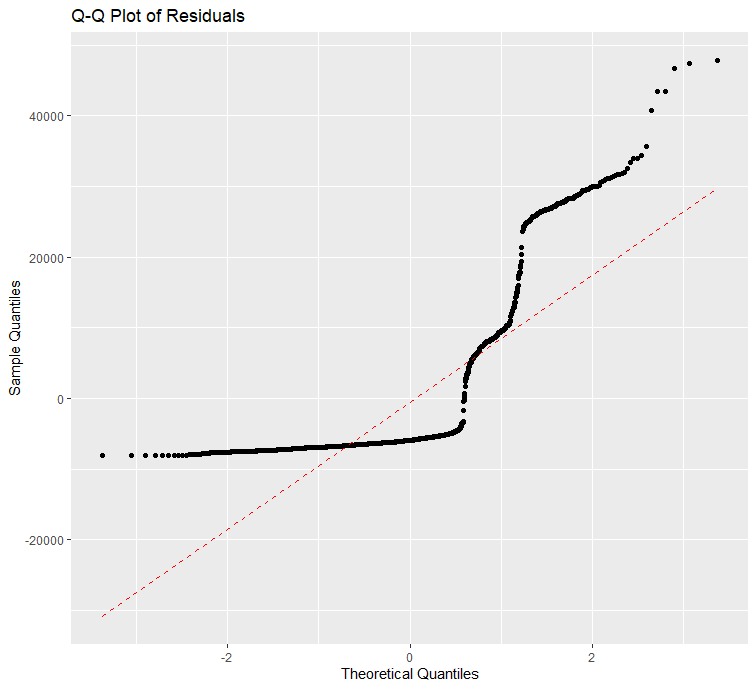
+ geom\_qq() +

+ stat\_qq\_line(distribution = qnorm, color = "red", linetype = "dashed") +

+ xlab("Theoretical Quantiles") +

+ ylab("Sample Quantiles") +

+ ggtitle("Q-Q Plot of Residuals")



The Q-Q plot visually compares the observed quantiles of a variable against the quantiles that would be expected if the variable followed a specific theoretical distribution, such as a normal distribution. If the points in the Q-Q plot approximately fall along a straight line, it indicates that the variable's distribution is close to the theoretical distribution being compared.

To better evaluate the plot, the Shapiro-Wilk test can be performed. It is a statistical test that provides a numerical measure of how well the data fit a normal distribution. It calculates a test statistic based on the correlation between the observed data and the expected values under the null hypothesis of normality. The null hypothesis of the test is that the residuals are normally distributed. The test result can help you assess the normality assumption of the residuals. The test result is a p-value that indicates the probability of obtaining the observed departure from normality by chance.

Shapiro-Wilk normality test

data: residuals

W = 0.65773, p-value < 2.2e-16

In this case, the test statistic (W) is 0.65773, and the p-value is near zero. Since the p-value is lower than the significance level (0.05), the null hypothesis rejected and conclude that there is enough evidence to suggest departure from normality.

These test shows that the regression models failed to satisfy its assumption requirements. It can be due to heteroscedasticity or missing variable. When the spread of the residuals varies across the range of fitted values, it indicates heteroscedasticity. Heteroscedasticity violates the assumption of constant variance in linear regression, and it can result in non-normal residuals. Robust regression techniques or transforming the response variable may help address this issue.

Missing variables can also be the cause of the failure in the test. If important predictor variables are missing from the model, the residuals may not capture all the relevant information, leading to non-normality. Including additional variables that have an impact on the response variable can help improve the model's goodness-of-fit and address the non-normality issue. Thus, the study should try to use multiple linear regression model that use other variables that may possibly affect insurance charges.

# **Part 2 – Multiple Linear Regression**

## **Descriptive Statistics**

age sex bmi children smoker

Min. :18.00 female:662 Min. :15.96 Min. :0.000 no :1064

1st Qu.:27.00 male :676 1st Qu.:26.30 1st Qu.:0.000 yes: 274

Median :39.00 Median :30.40 Median :1.000

Mean :39.21 Mean :30.66 Mean :1.095

3rd Qu.:51.00 3rd Qu.:34.69 3rd Qu.:2.000

Max. :64.00 Max. :53.13 Max. :5.000

region charges

northeast:324 Min. : 1122

northwest:325 1st Qu.: 4740

southeast:364 Median : 9382

southwest:325 Mean :13270

3rd Qu.:16640

Max. :63770

## **Correlation among all of the variables**

round(with(insurance, cor(cbind(age, sex, bmi, children, smoker, region, charges))),4)

age sex bmi children smoker region charges

age 1.0000 -0.0209 0.1093 0.0425 -0.0250 0.0021 0.2990

sex -0.0209 1.0000 0.0464 0.0172 0.0762 0.0046 0.0573

bmi 0.1093 0.0464 1.0000 0.0128 0.0038 0.1576 0.1983

children 0.0425 0.0172 0.0128 1.0000 0.0077 0.0166 0.0680

smoker -0.0250 0.0762 0.0038 0.0077 1.0000 -0.0022 0.7873

region 0.0021 0.0046 0.1576 0.0166 -0.0022 1.0000 -0.0062

charges 0.2990 0.0573 0.1983 0.0680 0.7873 -0.0062 1.0000

The table provided above shows the correlation coefficients between the variables in the insurance dataset. The correlation coefficient measures the strength and direction of the linear relationship between two variables. The brief interpretation of the correlation coefficients is as follows.

* Age and charges have a positive correlation coefficient of 0.2990, indicating a weak positive linear relationship. As age increases, the insurance charges tend to increase.
* Sex and charges have a correlation coefficient of 0.0573, suggesting a weak positive linear relationship. However, the correlation is quite small, indicating that gender has a limited impact on insurance charges.
* BMI and charges have a correlation coefficient of 0.1983, indicating a moderate positive linear relationship. As BMI increases, the insurance charges tend to increase.
* Children and charges have a correlation coefficient of 0.0680, suggesting a weak positive linear relationship. The number of children has a limited impact on insurance charges.
* Smoker and charges have a strong positive correlation coefficient of 0.7873, indicating a strong positive linear relationship. Smokers tend to have significantly higher insurance charges compared to non-smokers.
* Region and charges have a weak negative correlation coefficient of -0.0062, indicating a very weak negative linear relationship. The region of residence has a minimal impact on insurance charges.

It's important to note that correlation does not imply causation. While these correlation coefficients indicate the strength and direction of the linear relationship between variables, they do not provide information about causality or the underlying mechanisms at play.

## **Multiple Linear Regression Using All Variables**

> multiple\_model <- lm(charges ~ ., data=insurance)

> summary(multiple\_model)

Call:

lm(formula = charges ~ ., data = insurance)

Residuals:

Min 1Q Median 3Q Max

-11304.9 -2848.1 -982.1 1393.9 29992.8

Coefficients:

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

(Intercept) -11938.5 987.8 -12.086 < 2e-16 \*\*\*

age 256.9 11.9 21.587 < 2e-16 \*\*\*

sexmale -131.3 332.9 -0.394 0.693348

bmi 339.2 28.6 11.860 < 2e-16 \*\*\*

children 475.5 137.8 3.451 0.000577 \*\*\*

smokeryes 23848.5 413.1 57.723 < 2e-16 \*\*\*

regionnorthwest -353.0 476.3 -0.741 0.458769

regionsoutheast -1035.0 478.7 -2.162 0.030782 \*

regionsouthwest -960.0 477.9 -2.009 0.044765 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6062 on 1329 degrees of freedom

Multiple R-squared: 0.7509, Adjusted R-squared: 0.7494

F-statistic: 500.8 on 8 and 1329 DF, p-value: < 2.2e-16

The output provided above is the summary of a multiple linear regression model. The "Estimate" column shows the estimated coefficients for each predictor variable in the model. For instance, the estimated coefficient for age is 256.9, which suggests that for each year increase in age, the insurance charges are expected to increase by $256.9, holding other variables constant. Similarly, the other variables such as sex, BMI, children, smoker, and region have estimated coefficients indicating their association with the insurance charges.

The "Pr(>|t|)" column provides the p-values associated with each estimated coefficient. It indicates the statistical significance of each variable in relation to the insurance charges. Variables with p-values less than the chosen significance level (e.g., 0.05) are considered statistically significant. In this model, age, BMI, children, smoker, and regions (southeast and southwest) are statistically significant.

The multiple R-squared value (0.7509) represents the proportion of variance in the insurance charges that can be explained by the predictor variables in the model. In this case, approximately 75.09% of the variation in insurance charges is accounted for by the variables in the model. The adjusted R-squared value (0.7494) adjusts the multiple R-squared for the number of predictors in the model, providing a more accurate measure of the model's goodness of fit.

F-statistic: The F-statistic (500.8) tests the overall significance of the model by comparing the variability explained by the model to the variability not explained. In this case, the F-statistic is highly significant (p-value < 2.2e-16), indicating that the model as a whole is statistically significant.

Residuals: The "Residuals" section provides information about the distribution of the residuals, including the minimum, first quartile, median, third quartile, and maximum values. The residual standard error (6062) estimates the standard deviation of the residuals, indicating the average distance between the observed values and the predicted values by the model.

Overall, the multiple linear regression model shows a good fit to the data, with a high multiple R-squared value and statistically significant predictors. The model suggests that age, BMI, number of children, smoking status, and region have significant associations with insurance charges, while sex and certain regions (northwest) are not statistically significant predictors in this model.

## **Multiple Linear Regression Using only Significant Variables**

> #model experimentation 3: drop sex variable

> multiple\_model\_2 <- lm(charges ~ . -sex, data=insurance)

> summary(multiple\_model\_2)

Call:

lm(formula = charges ~ . - sex, data = insurance)

Residuals:

Min 1Q Median 3Q Max

-11367.2 -2835.4 -979.7 1361.9 29935.5

Coefficients:

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

(Intercept) -11990.27 978.76 -12.250 < 2e-16 \*\*\*

age 256.97 11.89 21.610 < 2e-16 \*\*\*

bmi 338.66 28.56 11.858 < 2e-16 \*\*\*

children 474.57 137.74 3.445 0.000588 \*\*\*

smokeryes 23836.30 411.86 57.875 < 2e-16 \*\*\*

regionnorthwest -352.18 476.12 -0.740 0.459618

regionsoutheast -1034.36 478.54 -2.162 0.030834 \*

regionsouthwest -959.37 477.78 -2.008 0.044846 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6060 on 1330 degrees of freedom

Multiple R-squared: 0.7509, Adjusted R-squared: 0.7496

F-statistic: 572.7 on 7 and 1330 DF, p-value: < 2.2e-16

> #model experimentation 3: drop sex and region variables

> multiple\_model\_3 <- lm(charges ~ age + bmi + children + smoker, data=insurance)

> summary(multiple\_model\_3)

Call:

lm(formula = charges ~ age + bmi + children + smoker, data = insurance)

Residuals:

Min 1Q Median 3Q Max

-11897.9 -2920.8 -986.6 1392.2 29509.6

Coefficients:

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

(Intercept) -12102.77 941.98 -12.848 < 2e-16 \*\*\*

age 257.85 11.90 21.675 < 2e-16 \*\*\*

bmi 321.85 27.38 11.756 < 2e-16 \*\*\*

children 473.50 137.79 3.436 0.000608 \*\*\*

smokeryes 23811.40 411.22 57.904 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6068 on 1333 degrees of freedom

Multiple R-squared: 0.7497, Adjusted R-squared: 0.7489

F-statistic: 998.1 on 4 and 1333 DF, p-value: < 2.2e-16

Comparing this model to the previous model, dropping the "sex" and "region" variables did not have a substantial impact on the model's performance. The adjusted R-squared value is slightly lower (0.7489) compared to the previous model (0.7494), suggesting a very slight decrease in the model's goodness of fit. However, the model with the reduced set of variables still provides a strong fit to the data and all remaining predictors (age, BMI, children, and smoker) are statistically significant.

When comparing models, it is generally preferable to choose a model with a simpler structure that achieves similar predictive performance. Removing non-significant variables helps to simplify the model and avoid overfitting. However, it's important to consider other factors as well, such as the interpretation and theoretical significance of the variables, the practical implications of the model, and potential confounding factors.

Therefore, based solely on the adjusted R-squared values, the revised model with fewer significant variables (adjusted R-squared of 0.749) is likely the better choice.

5

> # Generate random sample index

> sample\_3 <- sample(1:nrow(insurance),1)

> sample\_3

[1] 1066

> # Obtain age of the random sample

> sample\_3\_df <- data.frame(age=insurance$age[sample\_3], bmi=insurance$bmi[sample\_3], children=insurance$children[sample\_3], smoker=insurance$smoker[sample\_3])

> sample\_3\_df

age bmi children smoker

1 42 25.3 1 no

> # Calculate the predicted charges using the random sample age

> predicted\_value\_3 <- predict(multiple\_model\_3, newdata = sample\_3\_df)

> predicted\_value\_3

1

7343.253

> # The actual value for the random sample

> actual\_value\_3 <- insurance$charges[sample\_3]

> actual\_value\_3

[1] 7045.499

> # Residual is the difference between the actual value and the predicted value

> residual\_3 <- actual\_value\_3 - predicted\_value\_3

> residual\_3

1

-297.7537

> #Fourth sample

> # Generate random sample index

> sample\_4 <- sample(1:nrow(insurance),1)

> sample\_4

[1] 1090

> # Obtain age of the random sample

> sample\_4\_df <- data.frame(age=insurance$age[sample\_4], bmi=insurance$bmi[sample\_4], children=insurance$children[sample\_4], smoker=insurance$smoker[sample\_4])

>

> # Calculate the predicted charges using the random sample age

> predicted\_value\_4 <- predict(multiple\_model\_3, newdata = sample\_4\_df)

> predicted\_value\_4

1

9449.719

> # The actual value for the random sample

> actual\_value\_4 <- insurance$charges[sample\_4]

> actual\_value\_4

[1] 10577.09

> # Residual is the difference between the actual value and the predicted value

> residual\_4 <- actual\_value\_4 - predicted\_value\_4

> residual\_4

1

1127.368

The output provided above is the summary of another multiple linear regression model where the variables "sex" and "region" have been dropped. Coefficients: The "Estimate" column shows the estimated coefficients for each predictor variable in the model. In this case, the model includes age, BMI, children, and smoker. The estimated coefficients indicate the association between each predictor variable and the insurance charges. For example, the estimated coefficient for age is 257.85, suggesting that for each year increase in age, the insurance charges are expected to increase by $257.85, holding other variables constant. Similarly, the other variables (BMI, children, and smoker) have estimated coefficients indicating their association with the insurance charges.

The "Pr(>|t|)" column provides the p-values associated with each estimated coefficient. Variables with p-values less than the chosen significance level (e.g., 0.05) are considered statistically significant. In this model, all the predictor variables (age, BMI, children, and smoker) are statistically significant, as their p-values are much smaller than 0.05.

The multiple R-squared value (0.7497) represents the proportion of variance in the insurance charges that can be explained by the predictor variables in the model. In this case, approximately 74.97% of the variation in insurance charges is accounted for by the variables in the model. The adjusted R-squared value (0.7489) adjusts the multiple R-squared for the number of predictors in the model, providing a more accurate measure of the model's goodness of fit.

F-statistic: The F-statistic (998.1) tests the overall significance of the model by comparing the variability explained by the model to the variability not explained. In this case, the F-statistic is highly significant (p-value < 2.2e-16), indicating that the model as a whole is statistically significant.

Residuals: The "Residuals" section provides information about the distribution of the residuals, including the minimum, first quartile, median, third quartile, and maximum values. The residual standard error (6068) estimates the standard deviation of the residuals, indicating the average distance between the observed values and the predicted values by the model.