STATISTICAL AND PREDICTIVE MODDELLING

Soumya kataria: 100897536

Soumya kataria: 100897536

## **Research requirements**

* A linear regression model needs to be created to understand the effect of smoking on expenses.
* A multivariate regression model needs to be created to understand the effect of all input variables on expenses.
* The linear regression model should be used to analyse the effect of the Smoker variable on the Expenses variable.
* The multivariate regression model should be used to analyse the effect of all independent variables (Age, Sex, BMI, Children, Smoker, and Region) on the Expenses variable.

##### Dataset used

* The dataset provided (MultiRegDataset.csv) needs to be used for both regression models.
* The dataset contains 1338 observations and 7 features (columns).
* The independent (input) variables in the dataset are Age, Sex, BMI, Children, Smoker, and Region.
* The dependent (output) variable in the dataset is Expenses.

## **Basic statistics of the dataset**

The dataset named "MultiRegDataset.csv" contains 1338 observations or rows, and 7 features or columns. The dataset has six independent or input variables, which include Age, Sex, BMI, Children, Smoker, and Region. The dependent or output variable is Expenses, which means that it is the variable being predicted or explained by the independent variables through the analysis which we are going to conduct.

The dataset has 6 independent variables out of which 3 variables are numerical including age, bmi and children whereas remaining 3 variables are categorical including sex, smoker and region. Below is the summary statistics for numerical variables and expenses which is our dependent variable and frequency table for the categorical variables.

The standard deviation for age, bmi, children and expenses has also been calculated for the analysis.

## **5 number summary for age, bmi, children and expenses**

age bmi children expenses

Min. :18.00 Min. :16.00 Min. :0.000 Min. : 1122

1st Qu.:27.00 1st Qu.:26.30 1st Qu.:0.000 1st Qu.: 4740

Median :39.00 Median :30.40 Median :1.000 Median : 9382

Mean :39.21 Mean :30.67 Mean :1.095 Mean :13270

3rd Qu.:51.00 3rd Qu.:34.70 3rd Qu.:2.000 3rd Qu.:16640

Max. :64.00 Max. :53.10 Max. :5.000 Max. :63770

Age: This variable measures the age of the individuals in the dataset. The minimum age is 18 and the maximum age is 64. The average age is 39.21 years old. The first quartile is 27, meaning that 25% of the individuals in the dataset are aged 18 to 27, and the third quartile is 51, meaning that 75% of the individuals in the dataset are aged 18 to 51.

BMI: This variable measures the body mass index of the individuals in the dataset. The minimum BMI is 16 and the maximum BMI is 53.1. The average BMI is 30.67. The first quartile is 26.3, meaning that 25% of the individuals in the dataset have a BMI of 16 to 26.3, and the third quartile is 34.7, meaning that 75% of the individuals in the dataset have a BMI of 16 to 34.7.

Children: This variable measures the number of children of the individuals in the dataset. The minimum number of children is 0 and the maximum number is 5. The average number of children is 1.095. The first quartile is 0, meaning that 25% of the individuals in the dataset do not have children, and the third quartile is 2, meaning that 75% of the individuals in the dataset have 2 or fewer children.

Expenses: This variable measures the medical expenses of the individuals in the dataset. The minimum expenses are 1,122 and the maximum expenses are 63,770. The average expenses are 13,270. The first quartile is 4,740, meaning that 25% of the individuals in the dataset have expenses of 1,122 to

4,740, and the third quartile is 16,640, meaning that 75% of the individuals in the dataset have expenses of 1,122 to 16,640.

## **Frequency tables**

Following are the frequency tables for the categorical variables in the

“MultiRegDataset” dataset. The first one shows the number of males and females in the dataset, the second one shows the number of smokers and non-smokers, and the third one shows the number of individuals in each of the

four regions

Frequency table for Sex

female male

662 676

Frequency table for smoker

no yes

1064 274

Frequency table for region

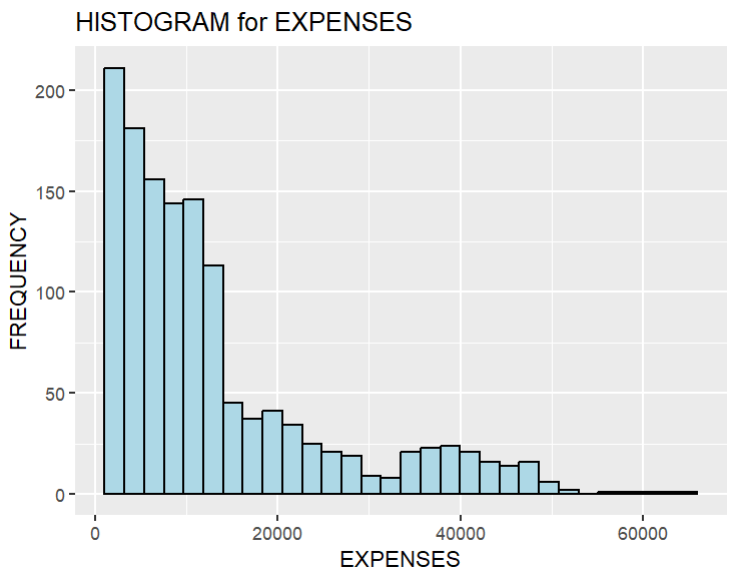
northeast northwest southeast southwest

324 325 364 325

## **Standard deviations for age, bmi, Children and expenses.**

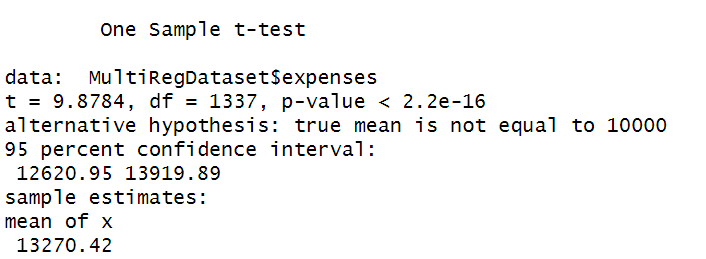
The SD is a measure of how spread out the data is from the mean. The SD of age is 14.05, which means that the age of individuals in the dataset varies from the mean age by approximately 14 years. The SD of BMI is 6.10, which means that the BMI of individuals in the dataset varies from the mean BMI by approximately 6.1. The SD of children is 1.21, which means that the number of children of individuals in the dataset varies from the mean number of children by approximately 1.21. The SD of expenses is 12,110.01, which means that the medical expenses of individuals in the dataset vary from the mean expenses by approximately $12,110.01.

## **Histogram for expenses**



The data is right/positively skewed implying the mean is greater than the median which is true (given in summary statistics. The peak of the data points are at the left side of the centre meaning that most of the data points are clustered on the left side of the distribution and the right tail is longer than the left tail.

## **Two-tailed T test**



#### ***Reason for choosing one sample T-test***

We conducted a two-tailed, one sample t-test because the population standard deviation was unknown. We opted for a two-tailed test because the alternative hypothesis suggested that the mean is not equivalent to 10,000; we did not need to determine whether the mean was greater or lesser than 10,000. Our objective was to determine if the population mean is equal to 10,000 or not, hence we chose a two-tailed, one sample t-test. In addition, we only had to consider expenses and it’s mean, and we were not working with two datasets or paired data.[[1]](#footnote-1)

#### ***Assumptions***

1. Data is independent.
2. Test variable i.e. expenses is continuous
3. Data is collected randomly.
4. The data is approximately normally distributed

Since Mr. john collected the data himself we are assuming that it is collected randomly and is independent. Expenses variable id continuous but the distribution isn’t close to normal but positively skewed. Hence, we are challenging assumption no.4 which can affect the authenticity of the results. [[2]](#footnote-2)

***Hypothesis statement:*** we want to test if, mean for expenses is equal to 10,000 or not.

* **Null hypothesis – Ho: µ = 10,000**
* **Alternative hypothesis - Ha: µ ≠ 10,000**

#### ***Analysis from the output of t-test***

* The t value is 9.8784 which is used to calculate the p value i.e.,2.2e-16. The df (degree of freedom) is 1337 because the number of observations is 1338 and df= n-1=1337, n represents the no. of observations.
* The next line states that alternative hypothesis is true mean is not equal to 10000(this is our alternative hypothesis). Here we reject the null hypothesis as significance level is 5% (0.05) and the P value (2.2e-16) which is way less than the significance level.
* In line 3 we have the confidence level as 95% (by default as it is the most commonly used value). Confidence internal + significance level = 1, which means significance level =5% or 0.05.
* The range of mean is 12620.95-13919.89. and confidence interval is 95% which means that we are 95% confident that the mean lies in this range.
* Mean of x (expenses) 13270.42 which is in the given range. Hence, we accept it as population mean.[[3]](#footnote-3)

***Conclusion***: since the P value (2.2e-16) is quite smaller than the significance level (0.05), we reject the null hypothesis which is Ho=10000 with 95%

confidence level. In other words, we have enough evidence that the mean of expenses is not 10,000.

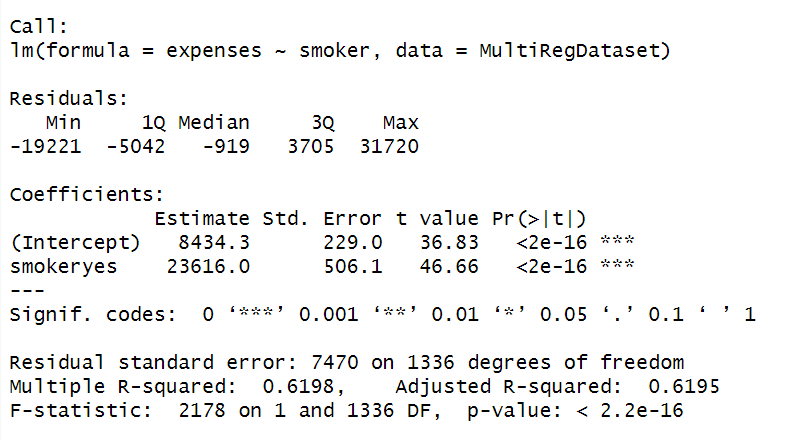
## **Simple linear regression model**

#### ***aim***

Linear regression is a statistical technique that aims to establish a mathematical relationship between one or more predictor variables (X) and a response variable (Y). This relationship is represented by a linear equation that can be used to predict the value of Y when the values of X are known.

The equation takes the form Y = β1 + β2X + ϵ, where β1 and β2 are coefficients that represent the intercept and slope of the line, and ϵ is the error term. The goal of linear regression is to find the values of β1 and β2 that minimize the difference between the predicted and actual values of Y.

We are using smoker as independent variable and expenses as dependent variable. We are going to use simple linear regression model to check if smoking affects expenses or not and if it does then how much do it affect the expenses for people who smoke or doesn’t smoke.



SECTION - 4

SECTION - 3

SECTION - 2

SECTION - 1

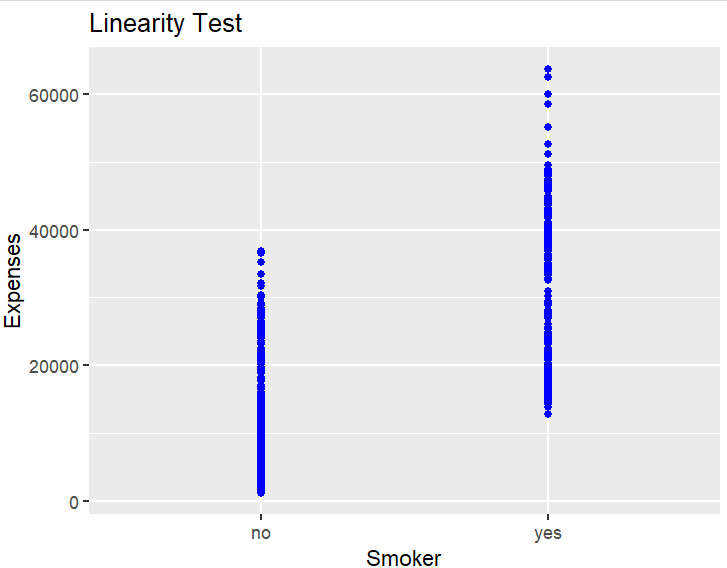
***Hypothesis statement:***

* **Ho: β=0, co-efficient β of the smoker is zero and not statistically significant**
* **Ha: β ≠0, co-efficient β of the smoker is not equal to zero and is statistically significant**

#### ***Assumptions***

1. Linearity: The relationship between the independent variable and the dependent variable is linear.
2. Independence: The observations in the dataset are independent of each other. There is no high correlation between the independent variables.
3. Homoscedasticity: The variance of the residuals (the differences between the predicted values and the actual values) is constant across all levels of the independent variable.
4. Normality: The residuals follow a near normal distribution.[[4]](#footnote-4)

#### ***Linearity test between expenses and smoker***



The scatterplot results indicate that there is no linear correlation between smoker and expenses. However, if we continue with this analysis without considering this fact, we may violate one of the assumptions and the results may be misleading.

**\*\*The independent variable smoker is categorical. This might affect the appearance of scatterplot that we’ll create to check the linearity.**

#### ***Analysis from the output of simple regression model.***

**Section-1** In the data frame titled "MultiRegDataset," the dependent variable is "expenses," and the independent variable is "smoker," and this line demonstrates that the linear regression model has been conducted using the "lm" function.

**Section-2** This section displays a statistical summary of the model residuals. The minimum residual is -19221, the first quartile (25th percentile), the median (50th percentile), the third quartile (75th percentile), and the highest residual (31720) are all values for the residual.

**Section-3** The intercept (8434.3) represents the predicted value of expenses for non-smokers (smoker = no/zero). The coefficient for the smoker variable (23616.0) represents the difference in the predicted value of expenses between smokers (smoker = yes/one) and non-smokers (smoker = no). Specifically, on average, smokers have predicted expenses that are 23616.0 higher than non-smokers.

**Section-4**

* Residual error-The estimated standard deviation of the errors or residuals is known as the residual standard error. It is an estimate of the dependent variable's variability that the independent variable is unaccounted for. The residual standard error in this instance is 7470. Smaller residuals are desirable since these are errors. Df is 1336 which is calculate by the formula (n-2) where n= sample size which is 1338.
* R-square and adjusted R-square- The R-squared value, which is 0.6198, shows how much of the variance in the dependent variable, "expenses," can be explained back to the independent variable, "smoker." The multiple R-squared value is adjusted based on the sample size and the number of independent variables using the adjusted R-squared value (0.6195).
* F-1 score- The F-statistic (2178) measures the overall significance of the regression model, and the associated p-value which is 2.2e-16 is less than the significance level (0.05), indicating that the model is statistically significant.

#### ***Conclusion:*** *Since the p-value (2.2e-16) is way smaller than the significance level 0.05, we reject the null hypothesis as the evidence is in favour of the alternate hypothesis. In other words, we can conclude our model is statistically significant and co-efficient β of the smoker is not equal to zero*

#### ***Equation***: (Y)expenses = 8434.3 + 23616.0 \* (X)smoker

#### Where smoker is a binary variable that takes the value 0 if the person is a non-smoker and 1 if the person is a smoker.

#### The intercept of 8434.3 represents the predicted value of expenses when the person is a non-smoker (smoker = 0). The coefficient of 23616.0 represents the increase in predicted expenses associated with being a smoker, compared to being a non-smoker*.*

## **Multiple linear regression model**

#### ***aim***

Multiple linear regression similar to simple linear regression but is used when there are two or more predictor variables that may influence the response variable. It is a statistical technique used to model the linear relationship between the response variable and multiple predictor variables We are using multi-linear regression model to see if any one of the independent variables affect expenses and if it does how much difference does it bring to expenses. This will be explained further in the estimated equation section with specific values as to the increase and it decrease the variable causes to expenses.

Equation: yi = B0 + B1xi1 + B2xi2 + ... + Bpxip + E where i = 1,2, ..., n

#### ***Hypothesis statement:***

* **Ho: βi = 0 (for i = 1, 2, 3,4,5,6)**
* **Ha: at least one βi ≠ 0 (for i = 1, 2, 3,4,5,6)**

**i=1-6 as we have 6 independent variables**

#### ***Assumptions***

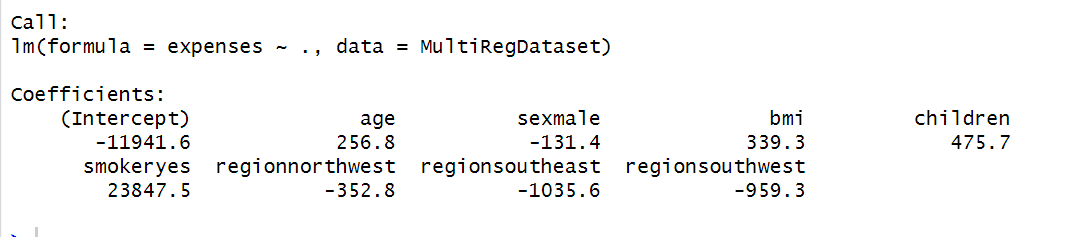
• The residuals of the model are nearly normal

• The variability of the residuals is nearly constant

• The residuals are independent and

• Each variable is linearly related to the outcome[[5]](#footnote-5)

#### ***Estimate equation***



Expenses = β0 + β1x1 + β2x2 + β3x3+ β4x4 + β5x5 + β6x6

(y)expenses = -11941.6 + 256.8 \* age - 131.4 \* sex(male) + 339.3 \* bmi + 475.7 \* children + 23847.5 \* smoker(yes) - 352.8 \* regionnorthwest - 1035.6 \* regionsoutheast - 959.3 \* regionsouthwest

where:

* Intercept (β0)= -11941.6
* expenses is the predicted value (y).
* age is the age of the individual(x1)
* sex(male) is a binary variable that takes the value 1 if the individual is male and 0 if the individual is female(x2)
* bmi is the body mass index of the individual(x3)
* children is the number of children the individual has(x4)
* smoker(yes) is a binary variable that takes the value 1 if the individual is a smoker and 0 if the individual is a non-smoker(x5)
* regionnorthwest, regionsoutheast, and regionsouthwest are binary variables that take the value 1 if the individual is from the corresponding region and 0 otherwise(x6) and if the individual is from regionnortheast then the value will remain 0.

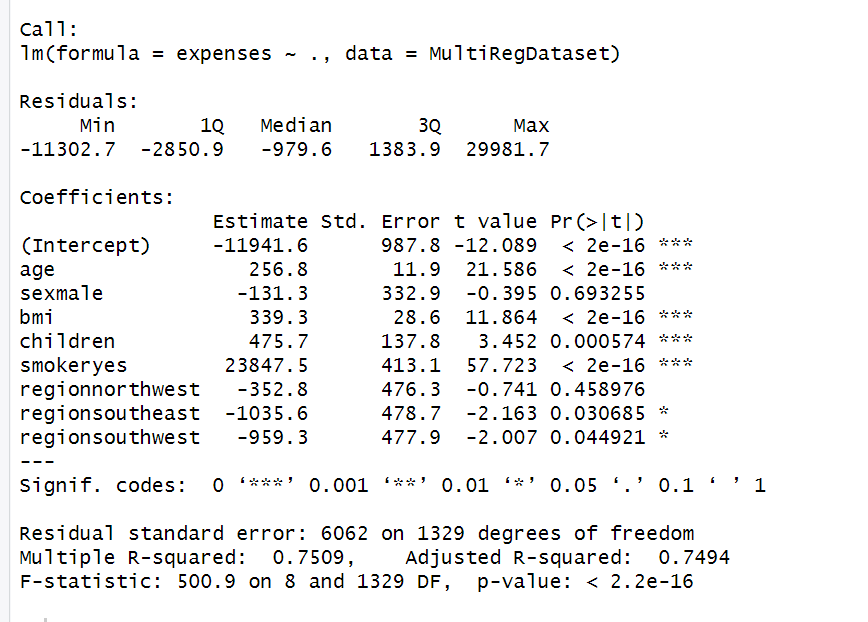
#### Co-eeficient values for the equation

* β1: Age = + 256.8
* β2: Sex = - 131.3 if Sex is ‘male’; 0 if Sex is ‘female’
* β3: BMI = + 339.3
* β4: Children = + 475.7
* β5: Smoker = 23847.5 if Smoker is ‘yes’; 0 if Smoker is ‘no
* β6: Region: - 352.8 if Region = ‘North West’ else 0

- 1035 if Region = ‘South East’ else 0

- 959.3 if Region = ‘South West’ else 0

0 if Region = ‘North East’



SECTION - 2

SECTION - 1

#### ***Analysis from the output of multiple regression model***

**Section-1**

* The model predicts expenses to be -11,941.6 when all other variables are zero, which is the value of the intercept (or constant; this value is not relevant in the context of expenses; it is only a statistical construct).
* For each one-year increase in age, expenses are estimated to increase by $256.80.
* Being male (as opposed to female) is estimated to be associated with a decrease in expenses of $131.30, but this effect is not statistically significant as p = 0.693
* For each one-unit increase in BMI, expenses are estimated to increase by $339.30.
* For each additional child, expenses are estimated to increase by $475.70
* Being a smoker (as opposed to a non-smoker) is estimated to be associated with a large increase in expenses of $23,847.50.
* Being from the northwest region (as opposed to the other reference region) is estimated to be associated with a decrease in expenses of $352.80, but this effect is not statistically significant (p = 0.459).
* Being from the southeast region (as opposed to the other reference region) is estimated to be associated with a decrease in expenses of $1,035.60, and this effect is statistically significant (p = 0.031).
* Being from the southwest region (as opposed to the reference region) is estimated to be associated with a decrease in expenses of $959.30, and this effect is statistically significant (p = 0.045).

**Section-2**

* The residuals are the differences between the actual expenses and the predicted expenses based on the model. The smallest residual is -11,302.7, the largest residual is 29,981.7, and the median residual is -979.6 (meaning the model predicted expenses that were $979.6 too low for half of the observations).
* The residual standard error is estimated to be 6,062, which is a measure of the average amount by which the model's predictions differ from the actual expenses (in this case, in dollars).
* The R-squared is 0.7509, which means that about 75% of the variation in expenses can be explained by the variables included in the model.
* The adjusted R-squared is 0.7494, which is a slightly lower value that takes into account the number of variables in the model and should be considered in this case instead of R-square.
* The F-statistic tests the overall significance of the model, and has a value of 500.9. The p-value is less than 2.2e-16, which is essentially zero, indicating that the model as a whole is statistically significant.
* Degree of freedom is not 1329 but 1331.

with n=1338 and k=6 (the number of independent variables), the degrees of freedom can be calculated as:

df = n - k - 1 = 1338 - 6 - 1 = 1331

where "n" is the total number of observations and "k" is the number of independent variables (excluding the intercept). Therefore, in this model, we have 1331 not 1329 degrees of freedom.

#### Conclusion

The p-value is 2.2e-16 which is less than 0.05 that is significance level. Therefore, we can conclude that we reject the null hypothesis stating that not one independent variable affects expenses. Instead, we fail to reject that at least one independent variable affects expenses.

# **OVERALL CONCLUSION**

First, we conducted a t-test to know if mean of expenses (we treated it as population mean) is equal to 10,000 or not. It turned out that expense mean is not equal to 10,000 but it actually is 13270.42. The p value was 2.2e-16 which was quite smaller than our significance level (0.05). Therefore, we rejected the null hypothesis with confidence interval of 95% as we had enough evidence to support that alternative hypothesis.

Then we conducted a simple linear regression where we concluded our model was statistically significant and the co-efficient of smoker is not equal to zero implying that independent variable smoker affect expenses. Here as well the significance level was 0.05 and P value was 2.2e-16 which is way smaller than the significance level. Also, we are 95% confident that the results are correct. Then we also calculated the amount of change which occurs in expenses when a person is smoker/non-smoker. It turned out that if a person smokes the expenses increase by 23616.0$ where as it a person doesn’t smoke the expenses doesn’t get affected.

Afterwards, we conducted a multi-linear regression model including all the independent variables given in the dataset to know if at least anyone of them affects expenses but it was already proved by the previous model so the main aim was to the level of change each independent variable brings in on expenses when they increase by 1 unit. The p-value was 2.2e-16 with significance level 0.05 and confidence interval 95%. Then we created an estimate equation for prediction. The equation is explained with the model above.

## Appendix (including the R code)[[6]](#footnote-6)

1. Summary statistics for numerical variables:

> summary(MultiRegDataset[c("age","bmi","children","expenses")])

age bmi children expenses

Min. :18.00 Min. :16.00 Min. :0.000 Min. : 1122

1st Qu.:27.00 1st Qu.:26.30 1st Qu.:0.000 1st Qu.: 4740

Median :39.00 Median :30.40 Median :1.000 Median : 9382

Mean :39.21 Mean :30.67 Mean :1.095 Mean :13270

3rd Qu.:51.00 3rd Qu.:34.70 3rd Qu.:2.000 3rd Qu.:16640

Max. :64.00 Max. :53.10 Max. :5.000 Max. :63770

2. Frequency table for categorical variables:

> table(MultiRegDataset$sex)

female male

662 676

> table(MultiRegDataset$smoker)

no yes

1064 274

> table(MultiRegDataset$region)

northeast northwest southeast southwest

324 325 364 325

3. Standard deviation of the numerical variables

> sd(MultiRegDataset$age)

[1] 14.04996

> sd(MultiRegDataset$bmi)

[1] 6.098382

> sd(MultiRegDataset$children)

[1] 1.205493

> sd(MultiRegDataset$expenses)

[1] 12110.01

4. Histogram for expenses

# Create histogram for Expenses column

ggplot(MultiRegDataset, aes(x = expenses)) +

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

ggtitle("HISTOGRAM for EXPENSES") +

xlab("EXPENSES") +

ylab("FREQUENCY")

5. two-tailed t test

t test code

T test two tailed

#One Sample t-test

t.test(MultiRegDataset$expenses, mu=10000)

6. Simple regression model

Scatterplot to check linearity between smoker and expenses

ggplot(MultiRegDataset, aes(x = smoker , y = expenses))+ geom\_point(colour = "blue") + geom\_smooth(method = "lm", fill = NA) + labs(title = "Linearity Test", x = "Smoker", y = "Expenses")

#Simple Regression Model

simple.fit<-lm(expenses ~ smoker, data = MultiRegDataset)

LinearModel<-simple.fit

summary(LinearModel)

7. multi-variable regression model

#Create the relationship model.

model <- lm(expenses~., data = MultiRegDataset)

# Show the model.

print(model)

#multiple reg model

multi.model <- lm(expenses ~., data = MultiRegDataset)

summary(multi.model)

1. Referred from assignment-3 submitted by me [↑](#footnote-ref-1)
2. Taken from course notes week-7 [↑](#footnote-ref-2)
3. Referred assignment-3 made by me [↑](#footnote-ref-3)
4. Referred course notes on linear regression model [↑](#footnote-ref-4)
5. Referred course notes on multi-linear regression model [↑](#footnote-ref-5)
6. The codes are taken from course notes and tutorials on t-test, simple regression model and multi-linear regression model [↑](#footnote-ref-6)