UTM
UNIVERSITI TEKNOLOGI MALAYSIA

SCHOOL OF COMPUTING
Faculty of Engineering

FACULTY OF ENGINEERING

SCHOOL OF COMPUTING

SEMESTER 2/20212022

SECI2143-02 PROBABILITY & STATISTICAL DATA ANALYSIS

GROUP 6

PROJECT 2

LECTURER: DR. NOR AZIZAH ALI

NAME	MATRIC NO
LOO ZHI YUAN	A21EC0197
NG ZI XING	A21EC0213
YEW RUI XIANG	A21EC0149

Table of Contents

1.0 Introduction

In this project 2, we have chosen various data sets from websites such as Kaggle and Springboard. A few compulsory requirements for this project 2 are hypothesis testing 1-sample or 2-sample, correlation and regression. Therefore, we have chosen the most suitable dataset from Kaggle, “Medical Cost Personal Datasets”. The details are attached in the appendix.

In this project, we aim to investigate whether there is a relationship between the variables ages, sex(gender), BMI, smoker, region and charges. There are a few ways that we decided to help our group to analyse and interpret the data set. In the data analysis process, we will use the hypothesis 2 sample test, correlation test, regression test, chi-square test of independence and ANOVA test to test the dataset and provide the results to support or refute our expectations.

Therefore, we were expected to see that different gender will affect the insurance charges as males usually have better and stronger body conditions. Furthermore, we also expected that age would affect the insurance charges because the higher the age will bring more risk to body health. In our further expectation of this study, the value of Body Mass Index (BMI) will affect the insurance charges because usually, the individuals overweight or underweight have more body health issues. While we started this study, we expected that an individual who is a smoker would affect the insurance charges because smoking is bad for body health. As our last expectation, we believe that insurance charges in different regions will be affected as they have different living habits and environments.

2.0 Dataset

Our data set used collected from online sources (Kaggle). This secondary data or data set consists of 7 different variables, but we only choose 6 out of 7, which are ages, sex (gender), BMI, smoker, region and charges. The table below shows the detail of each variable regarding their level of measurement and type of data.

No	**Variable/Parameter**	**Level of Measurement**	**Type of Data**
1.	Ages	Ratio	Discrete
2.	Sex (Gender)	Nominal	Categorical
3.	BMI	Ratio	Continuous
4.	Smoker	Nominal	Categorical
5.	Region	Nominal	Categorical
6.	Charges	Ratio	Continuous

Table 2.1: Type of variable

We choose these variables because they are the most common factors that will influence the insurance charges for a person. Thus we want to evaluate whether these variables affect insurance charges.

Then, hypothesis testing, correlation, regression, chi-square test for independence and ANOVA are used to test the sample data. The ways of data analysis are by using RStudio to generate a graphical presentation of data and do some of the basic calculations. Then, the conclusions are drawn.

3.0 Data analysis

3.1 Hypothesis 2 Sample Test

Based on the data we obtained, we wish to determine if there is any difference in mean of insurance charges between the male and female. So that the hypothesis tests on mean for 2 independent samples are carried out and the population variances are unknown. The sample size of the insurance charge for male is $n_1 = 676$ while for females it is $n_2 = 662$. Assume the population variance for the both group are unequal, then the test statistic formula and the degree of freedom are shows at below:

Test statistic formula, $T_0^* = \frac{\overline{X_1} - \overline{X_2} - \Delta_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$; degree of freedom, $v = \frac{(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{S_2^2}{n_2})^2}{n_2 - 1}}$

where,

$\overline{X_1} = sample\ mean\ of\ first\ sample$

$\overline{X_2} = sample\ mean\ of\ second\ sample$

$S_1^2 = sample\ variance\ of\ first\ sample$

$S_2^2 = sample\ variance\ of\ second\ sample$

$n_1 = sample\ size\ of\ first\ sample$

$n_2 = sample\ size\ of\ second\ sample$

Hypothesis statement:

Null hypothesis, H_0: $\mu_1 = \mu_2$

Alternative hypothesis, H_1: $\mu_1 \neq \mu_2$

where μ_1 represent the population mean insurance charges for male while μ_2 represent population mean insurance charges for females.

Test Statistic:

The sample mean of insurance charges for male, $\overline{X_1} = 13956.750$, while sample mean of insurance charges for female is $\overline{X_2} = 12569.580$. The sample variance for the both sample are calculated by using the formula $s^2 = \frac{\Sigma_{i=1}^{N}(x_i - \bar{x})^2}{N-1}$. The results are $s_1^2 = 168247513$ and $s_2^2 = 123848048$. A significant level of 0.05 ($\alpha = 0.05$) is used to test the claim that there is a difference in mean of insurance charges between

the male and female. From the RStudio we obtain the test statistic value, $T_0^* = 2.101$ and the degree of freedom is $df = 1313.36$. The probability value that we obtain is $P - value = 0.018$.

Conclusion:
Since $P - value <$ α or $(0.018 < 0.050)$, thus reject null hypothesis, H_0. There is sufficient evidence to support the claims that there is a difference in mean of insurance charges between the male and female. So, we can be sure that different genders have different charges, because the body condition of different genders, like the probability of getting cancer, is different, thus the charges are also different to overcome the risk.

3.2 Correlation Test

In the correlation test, we measure the relationship between the ages and the insurance charges for each individual in a sample size of 1338. Since both of our data are ratio type, Pearson's technique was selected for calculating the sample correlation coefficient, r. The formula for calculating the sample correlation coefficient, r by using Pearson technique and the suitable test statistic formula were shown below.

sample correlation coefficient, $r = \frac{\Sigma xy - (\Sigma x \Sigma y)/n}{\sqrt{[(\Sigma x^2) - (\Sigma x)^2/n][(\Sigma y^2) - (\Sigma y)^2/n]}}$

Test statistic formula, $t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$; degree of freedom, $df = n - 2$

where,
$r = Sample\ correlation\ coefficient$
$n = Sample\ size$
$x = Value\ of\ the\ independent\ variable$
$y = Value\ of\ the\ dependent\ variable$

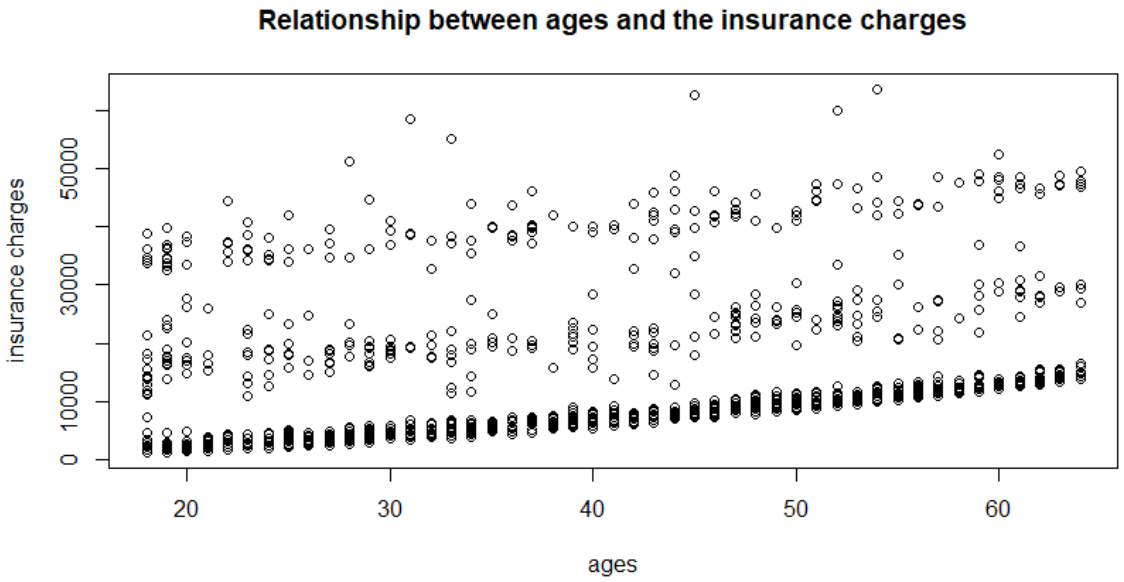

Figure 3.1: Scatter plot relationship between ages and insurance charges

From the scatter plot in Figure 3.1, we can see that there is a positive linear correlation between the ages and insurance charges.

Hypothesis statement:

Null hypothesis, H_0: $\rho = 0$ (no linear correlation)

Alternative hypothesis, H_1: $\rho \neq 0$ (linear correlation exists)

where ρ represent the population correlation coefficient.

Test statistic:

A significant level of 0.05 ($\alpha = 0.05$) is used to test the claim that there exists linear correlation between the ages and the insurance charges for each individual. From the RStudio we obtain the correlation coefficient, $r = 0.299$, test statistic value, $t = 11.453$ and the degree of freedom is $df = 1336$. The probability value that we obtain is $P - value = 2.443 \times 10^{-29}$.

Conclusion:

Since $P - value < \alpha$ or $(2.443 \times 10^{-29} < 0.050)$, thus reject null hypothesis, H_0.

There is sufficient evidence to support the claims that there exists linear correlation between the ages and the insurance charges for each individual. As the age is positively correlated with insurance charges, this makes common sense, because the older people, the weaker the body condition. So, the insurance company has to take more risk to pay for the people with high ages. But the correlation coefficient is low, so there is no very strong linear relationship.

3.3 Regression Test

In the regression test, we measure the relationship between the BMI and insurance charges for each individual with a sample size, $n = 1338$. The independent variable,x is the BMI and the dependent variable,y is insurance charges. The sample regression line provides an estimate of the population regression line.

Estimated Regression Model, $\widehat{y_i} = b_0 + b_1 x$

where,

$\widehat{y_i} = estimated\ dependent\ variable\ value$

$b_0 = estimated\ of\ the\ regression\ intercept$

$b_1 = estimated\ of\ the\ regression\ slope$

$x = independent\ variable$

Test statistic formula, $t = \frac{b_1 - \beta_1}{s_{b_1}}$; degree of freedom, $df = n - 2$

where,

$b_1 = Sample\ regression\ slope\ coefficient$

$\beta_1 = Hypothesized\ slope$

$s_{b1} = Estimator\ of\ the\ standard\ error\ of\ the\ slope$

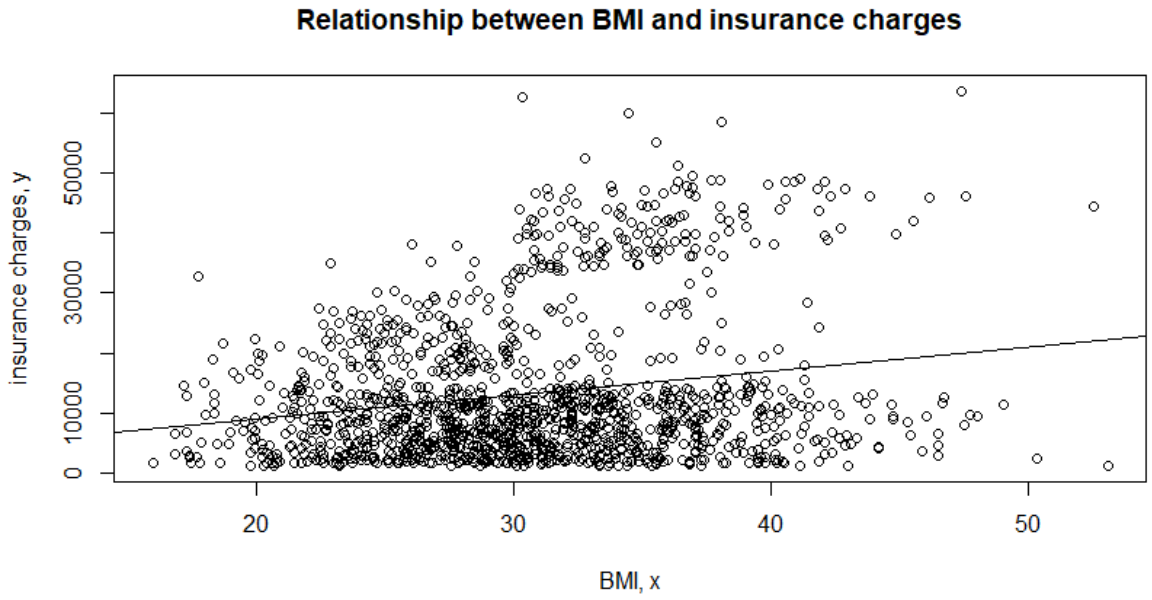

Figure 3.2: Scatter plot relationship between BMI and insurance charges

From the scatter plot in Figure 3.2, we can see that there is a positive linear relationship between the BMI and insurance charges. The best fit line shows an upward inclining.

From the summary, we can get the formula for estimated regression model which is

$$Y = 1192.94 + 393.87X$$

From the formula, the value of $b_0 = 1192.94$ is the estimated value of y when the value of x is equal to 0. While the value of $b_1 = 393.87$ measures the estimated change in the average value of insurance charges, y as a result of one-unit change in BMI, x.

The coefficient of determination is the portion of the total variation in the dependent variable that is explained by variation in the independent variable.

$$R^2 = \frac{SSR}{SST} = \frac{sum\ of\ squares\ explained\ by\ regression}{total\ sum\ of\ squares}$$

From the RStudio we obtain the value of coefficient of determination, $R^2 = 0.039$. Since $0 < R^2 < 1$, the result shows a weaker linear relationship between the BMI, x and insurance charges, y. Some but not all of the variation in insurance charges, y is explained by the variation in BMI, x.

Hypothesis statement:

Null hypothesis, H_0: $\beta_1 = 0$ (no linear relationship)

Alternative hypothesis, H_1: $\beta_1 \neq 0$ (linear relationship does exists)

where β_1 represent the population slope coefficient (hypothesized slope).

Test statistic:

A significant level of 0.05 ($\alpha = 0.05$) is used to test the claim that there exists a linear relationship between the BMI and the insurance charges for each individual. From the RStudio we obtain the sample regression slope coefficient, $b_1 = 393.87$, estimator of the standard error of the slope, $s_{b_1} = 53.25$, test statistic value, $t = 7.397$ and the degree of freedom is $df = 1336$. The probability value that we obtain is $P - value = 2.459 \times 10^{-13}$.

Conclusion:

Since $P - value < \alpha$ or ($2.459 \times 10^{-13} < 0.050$), thus reject null hypothesis, H_0. There is sufficient evidence to support the claims that there exists a linear relationship

between the BMI and the insurance charges for each individual. We can see the residual and residual standard error which is the difference between y and predicted y is larger, r square is low, so there is no strong linear relationship. But we can't make the conclusion that there is no relationship between charge and bmi, because we just prove that there is a linear relationship.

3.4 Chi-square Test of Independence

Chi-square test of independence is used to test for the relation between two nominal variables. In order to proceed with the Chi-square test, we summarize our data using a two way contingency table. In this test, we test whether there is a relationship between smoker and non-smoker in terms of insurance charges. The formula of test statistic is:

Test statistic formula, $\chi^2 = \Sigma_{all\ cells} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$; degree of freedom, $df = (r - 1)(c - 1)$

where,

$o_{ij} = Observed\ count\ at\ row\ i\ column\ j$

$e_{ij} = Expected\ count\ at\ row\ i\ column\ j$

$r = total\ number\ of\ rows$

$c = total\ number\ of\ columns$

Hypothesis statement:

Null hypothesis, H_0: insurance charges are independent on whether a person is a smoker or not

Alternative hypothesis, H_1: insurance charges are dependent on whether a person is a smoker or not

Test statistic:

A significant level of 0.05 ($\alpha = 0.05$) is used to test the claim that insurance charges are dependent on whether a person is a smoker or not. The below shows the two way contingency table created by using RStudio.

Smoker	(1121.8734, 19916.439783]	(1.99e+04, 2.62e+04]	(2.62e+04, 3.24e+04]	(32446.150705, 63770.42801]
no	1003	35	22	4
yes	60	51	13	150

Table: Observed frequency table

Smoker	(1121.8734, 19916.439783]	(1.99e+04, 2.62e+04]	(2.62e+04, 3.24e+04]	(32446.150705, 63770.42801]
no	845.3154	68.38864	27.832586	122.46338
yes	217.6846	17.61136	7.167414	31.53662

Table: Expected frequency table

From the RStudio we obtain the test statistic value, $\chi^2 = 788.79$ and the degree of freedom is $df = 3$. The probability value that we obtain is $P - value = 2.2 \times 10^{-16}$.

Conclusion:

Since $P - value < \alpha$ or $(2.2 \times 10^{-16} < 0.050)$, thus reject null hypothesis, H_0. There is sufficient evidence to support the claims that insurance charges are dependent on whether a person is a smoker or not. So, we can be sure that the charge is highly dependent on the smoker, this is naturally because smokers have more risk of getting some disease.

3.5 ANOVA Test

The purpose of the ANOVA test we conducted is to test for significance differences between means of insurance charges for four different regions. We use a 0.05 significance level to test the null hypothesis that the insurance charges for people in different regions will have the same mean. Due to the different sample size of each population being compared, we apply One-way ANOVA with Unequal Sample Sizes technique.

Test statistic formula, $F = \frac{SST/(k-1)}{SSE/(n-k)}$

numerator degrees of freedom $= k - 1$

denominator degrees of freedom $= n - k$

where,

$SST = \Sigma_{i=1}^{k} n_i (\overline{Y}_i - \overline{Y})^2$

$SSE = \Sigma_{i=1}^{k} \Sigma_{j=1}^{n_i} (Y_{ij} - \overline{Y}_i)^2$

$\overline{Y} = \frac{1}{k} \Sigma_{i=1}^{k} \overline{Y}_i$

$k = number\ of\ population\ or\ treatments\ being\ compared$

$n = \Sigma_{i=1}^{k} n_i$; $n_i = sample\ size\ for\ population\ i$

Hypothesis statement:

Null hypothesis, H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

Alternative hypothesis, H_1: $at\ least\ one\ mean\ is\ different$

where μ_1 represent the mean of insurance charges in the southwest, μ_2 represent the mean of insurance charges in the southeast, μ_3 represent the mean of insurance charges in the northwest, μ_4 represent the mean of insurance charges in the northeast.

Test statistic:

A significant level of 0.05 ($\alpha = 0.05$) is used to test the claim that the insurance charges for people in different regions will have the same mean. From the RStudio we obtain test statistic value, $F = 2.975$ and the degree of freedom is $df = 1336$. The probability value that we obtain is $P - value = 0.031$.

Conclusion:

Since $P - value < \alpha$ or $(0.031 < 0.050)$, thus reject null hypothesis, H_0. There is sufficient evidence to support the claims that the insurance charges for people in different regions will have at least one different mean. We can't assume that different region have same charge, different region have different culture and habit of diet, lead to the risk to getting some disease is different. So that the pay of the insurance is affected also by the regions.

4.0 Conclusion

In this project, we have learnt a lot, especially while choosing a dataset in which we must consider the requirements, such as suitable variables. For example, we must have at least two nominal variables in our dataset to perform a chi-square test of independence. Therefore, analyzing and discussing the dataset before choosing is very important to further analysis.

In the analysis process, we are using the RStudio to make the calculation needed to produce the desired output for further analysis in this project. Hence, understanding the R code and its output is essential to ensure we can produce the most accurate result and explanation. This analysis process helps us enhance our R programming skills to help us manage R programming in the future.

Using the hypothesis 2 sample test, we have found the most exciting result that there is a difference in the mean of insurance charges between the male and female, which means different insurance charges apply to the different gender. We always thought there should be no difference between healthy males and females, but the result extends our curiosity more comprehensively.

In conclusion, we determined that the ages positively correlate to the insurance charges but not in a very strong relationship. Then, the BMI and insurance charges are also related but without a strong relationship based on the regression test. From the chi-square test of independence, insurance charges are dependent on whether a person is a smoker or not. Lastly, the ANOVA test results show that the insurance charges will be affected by the regions. Hence, our expectations were proved based on the analysis results of the data set.

5.0 Appendix

Sample of Raw Dataset

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.9
18	male	33.77	1	no	southeast	1725.55
28	male	33	3	no	southeast	4449.46
33	male	22.705	0	no	northwest	21984.5
32	male	28.88	0	no	northwest	3866.86
31	female	25.74	0	no	southeast	3756.62
46	female	33.44	1	no	southeast	8240.59
37	female	27.74	3	no	northwest	7281.51
37	male	29.83	2	no	northeast	6406.41
60	female	25.84	0	no	northwest	28923.1
25	male	26.22	0	no	northeast	2721.32
62	female	26.29	0	yes	southeast	27808.7
23	male	34.4	0	no	southwest	1826.84
56	female	39.82	0	no	southeast	11090.7
27	male	42.13	0	yes	southeast	39611.8
19	male	24.6	1	no	southwest	1837.24
52	female	30.78	1	no	northeast	10797.3
23	male	23.845	0	no	northeast	2395.17
56	male	40.3	0	no	southwest	10602.4
30	male	35.3	0	yes	southwest	36837.5
60	female	36.005	0	no	northeast	13228.8
30	female	32.4	1	no	southwest	4149.74
18	male	34.1	0	no	southeast	1137.01
34	female	31.92	1	yes	northeast	37701.9
37	male	28.025	2	no	northwest	6203.9
59	female	27.72	3	no	southeast	14001.1
63	female	23.085	0	no	northeast	14451.8

Data Resource Link

https://www.kaggle.com/datasets/mirichoi0218/insurance

Video presentation link

Video presentation link: https://youtu.be/gxJOHwvZfU4

E-portfolio Link

NG ZI XING
LOO ZHI YUAN
YEW RUI XIANG