Christopher Kuetsinya
EXST 7005
Dr. Gentimis April 19, 2021

Body Mass Index (BMI) of policyholders and Policy Charges of SIClife Insurance

INTRODUCTION

In insurance, effective policy pricing is vital. Charging the right amount insurance policies is very important for Insurance Companies to keep them in business (Wells, 2015) and as a prospective policyholder, knowing exactly the factors that affect the pricing of policies helps to know which Insurance company to patronize.

One of the factors that is considered when pricing insurance at SIClife is the body mass index of policyholders. Prospective policyholders must know this in other to make informed decisions about their weight and health. Also, this information helps prospective policyholders on whether or not to purchase health insurance from this company, since other insurance companies do not consider body mass index for insurance pricing.

The goal of this study is test the significance of the relationship between the body mass index of policyholders and the price of health insurance policies at SIClife.

DATASET DESCRIPTION

At the end of 2020, records of 200 health insurance policyholders were randomly selected from the "2020 New Clients" database of SIClife, an insurance company in Ghana. Records include responses given by the clients upon buying the policy (i.e. age, gender, BMI, number of children and smoking status). The underwriters use these responses to ascertain the charges for each individual.

For this study, interest is only in variables BMI and Charges.

bmi: the body mass index of each individual. It is derived by dividing the body mass by the square of the body height, and is universally expressed in units of kg/m^2.

charges: the policy amount each person must pay per annum to be covered. Charges are in Ghana cedis (GH₵).

Table 1: A part of the Data set of interest

Charges	bmi
16884.92	27.9
1725.55	33.77
4449.46	33
21984.47	22.71
3866.86	28.88
3756.62	25.74

STATISTICAL MODEL

Correlation Coefficient

In determining the relationship between the response variable (charges) and the explanatory variable (bmi), the correlation coefficient can be used. The correlation coefficient measures the strength of linear relationship between two quantitative variables (Freund, Mohr, & Wilson, 2010). Its value is between +1 and – 1 inclusive. Values of +1 and −1 signify an exact positive and negative relationship, respectively, between the variables (Freund, Mohr, & Wilson, 2010).
In this study, the Pearson's Correlation Coefficient denoted by r, will be used to measure the strength of linear relationship between the variables. It is estimated by

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}S_{yy}}}$$

Where

x_i is the values of the explanatory variable (bmi) of the ith observation from the data.
$\bar{x}$ is the mean of the values of the explanatory variables (bmi).
y_i is the values of the response variable (charges) of the ith observation from the data.
$\bar{y}$ is the mean of the values of response variable (charges).

Simple Linear Regression

Simple linear regression is a statistical method used for analyzing the relationship between two quantitative variables in such a manner that one variable can be predicted or explained using information from the other variable (Freund, Mohr, & Wilson, 2010). A simple linear regression model is given by

$$Y_i = \beta_o + \beta_1 X_i + \varepsilon_i$$

Where

- Y_i is the values of the response variable (charges) of the ith observation from the data
- β_0 is the intercept of the model. It is the value of the response variable (charges) when the value of the explanatory variable is zero. That is, when $X = 0, Y = \beta_o$
- β_1 is the slope of the model. It is the change in the response variable Y (charges), when the response variable X (bmi) is increased or decreased.
- X_i is the values of the explanatory variable (bmi) of the ith observation from the data.
- ε_i is the error term. It follows a normal distribution with mean 0 and variance σ^2.

To establish the relationship using the Simple Linear Regression Model, the parameters of the model i.e. β_0 and β_1 must be estimated. These parameters can be estimated using the Least Squares criterion. The goal of this criterion is to find the best estimates for the regression coefficients by making the sum of the squares of the errors a minimum. They are estimated by;

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}} \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

In simple linear regression some assumptions must be met in other for the test to be useful. These assumptions are;

- There is a linear relationship between the two variables.
- The error terms ae independent.
- The error terms are approximately normally distributed with mean 0 and variance σ^2.
- The error terms have a common variance σ^2.

The normality of residuals and homogeneity of variance will be checked.

Hypothesis Testing

After running the simple linear regression model, the test for regression effect can be turned into a hypothesis test with null and alternative hypothesis as $H_o: \beta_1 = 0, H_A: \beta_1 \neq 0$. This tests whether or not the coefficient of the explanatory variable is statistically significant to help know if the explanatory variable can predict the response variable (Freund, Mohr, & Wilson, 2010). This conclusion can be drawn from the p-value of the F value from the model in a table that resembles an ANOVA table.

Table 2: Statistical test for Regression table

Source	Degrees of freedom	Sum of squares	Mean square	F value	P value
Regression	p – 1	SSR = $\beta_1 \times S_{xy}$	MSR = SSR/df	MSR/MSE	p-value
Error	n – p	SSE = SST – SSR	MSE = SSE/df		
Total	n – 1	SST = S_{yy}			

The degrees of freedom for regression is calculated by the number of regression coefficients (p) minus 1. The error degree of freedom is calculated by total number of observations (n) minus the number of regression coefficients (p).

In using the least square criterion, the best parameters are computed to minimize the sum of square of the residuals which is also known as the Error of the residual sum of squares denoted by SSE. Now this SSE describes the variation in the response variable as an estimate of a linear model of the explanatory variable.

The first focus in this table is the Mean square error(MSE) which is the average variation for each of predictions and actual numbers. The goal here is to have a smaller MSE, since a smaller the MSE the better the linear regression fits the data.

The main focus of the table above, is to compare p-value to a 0.05 alpha level and draw conclusion.

LITERATURE REVIEW

Body Mass Index (BMI) is a person's weight in kilograms divided by the square of height in meters. A high BMI can be an indicator of high body fatness (CDC, 2021).

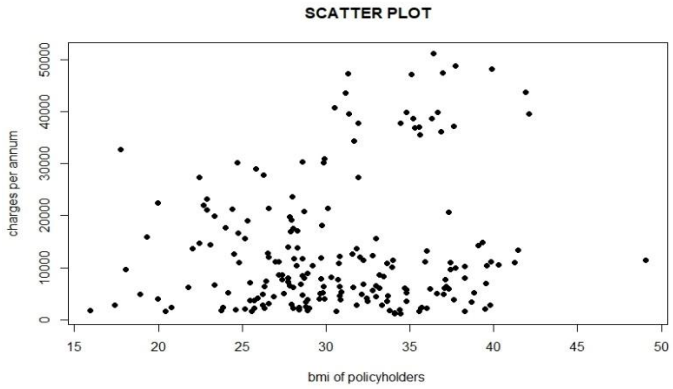
Excess or high BMI is associated with substantially shorter life expectancy and some chronic health conditions (Stenholm, et al., 2017). Health and life insurances are meant to meet the health

needs of the policyholder and also provide indemnity to any relative of the policyholder in case death depending on the terms of the policy. Therefore, insurance companies are very concerned with the elements that contributes to ill-health and shorter life expectancy to factor that in policy pricing so not run in debts or loss due to frequent claims payment (Wells, 2015).
With the prevalence obesity in Ghana and the world at large which constitutes a risk factor of many diseases like diabetes, hypertension and other heart diseases (Biritwum, Gyapong, & Mensah, 2005), I think it is appropriate to study the relationship between BMI and policy charges at SIClife insurance.

STATISTICAL ANALYSIS

The type of the relationship between the two variables was checked, the correlation coefficient and a correlation test was also computed. Also, some assumptions of the simple linear regression were checked. After, the model was fitted and a hypothesis testing was done to check for regression effect. All the computations and graphs were done using the R software.

Graph 1: Scatter plot of bmi of charges



The scatter plot indicates some sought of positive relationship between 'bmi' and 'charges'. However, a lot of dots can be found horizontally at the bottom of the plot. The correlation coefficient and test will provide a clear picture of the relationship.

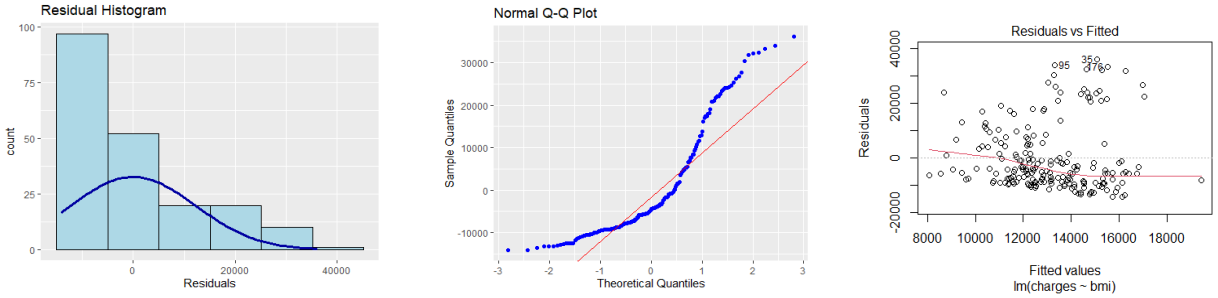
Table 3: Pearson Correlation Coefficient and Correlation test

Pearson's product-moment correlation

data: InsuranceData$bmi and InsuranceData$charges
t = 2.2387, df = 198, p-value = 0.02629
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
0.01879307 0.28955327
sample estimates:
cor
0.1571244

The value of the correlation coefficient is approximately 0.16 which indicates a positive linear relationship but a weak one since the correlation coefficient is close to zero. The hypothesis testing for the correlation coefficient produces a p-value less than 0.05, therefore the correlation coefficient significant different from zero. Here, we can safely say there is a correlation between bmi and charges.

Graph 2 **Graph 3** **Graph 4**



Graphs 2 and 3 can be used to check for normality of residuals. For graph 2, which is a histogram, we would like to see a bell-shaped curve indicating normality and in graph 3 (q-q plot), for the normality assumption to hold the dots in the plot must be on the red line. Both graphs do not provide what we look out for which means the residuals do not have a normal shape.

Table 4.

```
##  Shapiro-Wilk normality test
## H₀: normal distribution
## data:  SLR$residuals
## W = 0.83778, p-value = 1.138e-13
```

The Shapiro-Wilk test can also be used to test for normality of residuals. Here, the p-value is less than 0.05 so we reject the null hypothesis and conclude that the residuals are not normally distributed. This conclusion in agreement with the graphs 2 and 3 hence the normality assumption is violated.

Graph 4 tests the homogeneity of variance assumption. The residuals in the graph are not evenly distributed around zero and most of the observations are closely packed hence the homogeneity of variance assumption is violated.
The simple linear regression will still be fitted although the assumptions tested were violated.

Table 5. Parameter Estimates

Parameters	Estimate	Std. Error	t-value	Pr(>| t |)
Intercept	2580.4	4777.0	0.540	0.5897
bmi	34.3	153.4	2.239	0.0263*

The model estimated is; $\widehat{\boldsymbol{charges}} = \mathbf{2580.4} + \mathbf{34.3}(\boldsymbol{bmi})$
Charges increase by 34.3 for every 1-unit increase in bmi.

Table 6. Hypothesis testing of Regression effect

Source	Degrees of freedom	Sum of squares	Mean square	F value	P value
Regression	1	7.5007e+08	750072217	5.012	0.02629 *
Error	198	2.9632e+10	149656002		
Total	199	3.0382e+10			

H_0: $\beta_1 = 0$. The regression coefficient for the variable bmi is zero.
H_A: $\beta_1 \neq 0$. The regression coefficient for the variable bmi is significantly different from zero.
The p-value is less than 0.05 so we reject the null hypothesis and conclude that the regression coefficient of bmi is significantly different from zero.

CONCLUSION

The correlation coefficient of the variables of the in this study which is approximately 0.16 indicates a weak linear relationship between 'bmi' and 'charges'.
Secondly, the normality of residuals and homogeneity of variance assumptions were not met for this study but the simple linear regression was still fitted.
Also, the hypothesis test was significant which means the variable 'bmi' is a good predictor of the variable 'charges'.
In conclusion, all the factors that contribute in determining the price of a policy should be studied using a complex statistical to bring out good predictions.

BIBLOGRAPHY

1. Biritwum, R., Gyapong, J., & Mensah, G. (2005). The Epidemiology of Obesity in Ghana. *Ghana Medical Journal*, 82–85.
2. CDC. (2021). Defining Adult Overweight & Obesity. https://www.cdc.gov/obesity/adult/defining.html
3. Freund, R. J., Mohr, D., & Wilson, W. J. (2010). *Statistical Methods.* Burlington: Academic Press, 325–357.
4. Stenholm, S., Head, J., Aalto, V., Kivimäki, M., Kawachi, I., Zins, M., . . . Vahtera, J. (2017). Body mass index as a predictor of healthy and disease-free life expectancy between ages 50 and 75: a multicohort study. *International Journal of Obesity*.
5. Wells, A. (2015). The Price of Price Optimization In Insurance. *Insurance*. https://www.insurancejournal.com/news/national/2015/11/17/389153.htm

APPENDIX

```r
#'packages
library(olsrr)

#' Opening data in R
InsuranceData <- read.csv("Project data.csv", header = TRUE)

#' Scatter plot to show relationship between charges and bmi
with(InsuranceData, plot(bmi,charges, main = "Scatterplot"))

#' Pearson's Correlation Coefficient
cor(InsuranceData$bmi, InsuranceData$charges, method = c("pearson"))

#' Simple Linear Regression Model
SLR <- lm(charges ~ bmi, data = InsuranceData)
summary(SLR)

#' Test for Normality of Residuals
ols_plot_resid_hist(SLR)

ols_plot_resid_qq(SLR)

shapiro.test(SLR$residuals)

#' Test for Homogeneity of variance
plot(SLR, which=c(1,2))

#' Hypothesis testing for regression effect
anova(SLR)
```