PREDICTING MEDICAL CHARGES USING LINEAR REGRESSION

By: Aldwin Dave Conahap and Abegail Romas

I. Introduction

Affordable health insurance is being offered by ACME Insurance, Inc. to thousands of customers all over the United States. With this, they aim to estimate the annual medical expenditure for new customers, using information such as their age, sex, BMI, children, smoking habits, and region of residence. The estimates will play a role in determining the annual insurance premium (amount paid every month) offered to every customer.

As Crow et al. (2014) wrote, increased BMI is a significant predictor in the amount of health resources utilized for medical procedures and hospitalization. Similarly, as found in another study, expenditures for injury, respiratory, and circulatory diseases all increased with a person's BMI (Pan et al., 2012).

Hence, the goal in this paper is to demonstrate univariate linear regression to predict the medical charges for smokers using body mass index (BMI).

II. Methodology

a. Description of data

Since an overview of the previous literature suggests a link between medical charges and body mass index (BMI), the data set selected for this paper is a CSV file from Kaggle containing verified historical data, consisting of information and actual medical charges incurred by over 1,300 customers (Singh, 2020). The file has a total of seven columns, namely, age, sex, BMI, children, smoker, region, and charges. However, in this paper, we are only interested in using the BMI of smokers to predict the charges. Thus, the only columns used are BMI and charges, which are the independent variable and the target variable, respectively.

b. Techniques

The dataset was separated into two categories: smoker and nonsmoker. Then, Pearson's correlation analysis was performed to examine the relationship between the body mass index (BMI) and medical charges of smokers from the data. The univariate regression analysis was performed using machine learning algorithms. This involved the scikit-learn library for splitting the data into the train set and test set to fit and predict the regression model. Additionally, various machine learning regression techniques were tested to determine the best line of fit. Consequently, the ordinary least squares method was chosen.

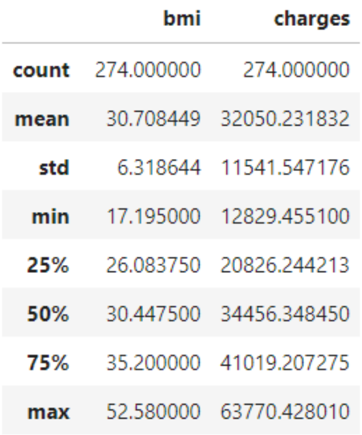
To evaluate the accuracy of the model, the coefficient of determination (r^2) and the mean absolute percentage error (MAPE) were determined. Then, we verified the assumptions for linear regression to assess the reliability of the model.

Other analyses were performed using the Python programming language, which involved pandas and NumPy libraries for data manipulation, computations, and analysis; scikit-learn, SciPy, and statsmodel for the statistical analyses; and matplotlib and seaborn for data visualizations.

III. Results and Discussion

Correlation Analysis

Descriptive statistics were used to summarize the statistical features to be used in the analyses (see *Figure 1*). The mean BMI of the smoker patients is approximately 30.7 (Obese), and the mean medical charge is nearly $\$\ 32{,}050.23$. As shown in *Figure 2*, an outlier has also been observed in the BMI of the smoker patients.



	bmi	charges
count	274.000000	274.000000
mean	30.708449	32050.231832
std	6.318644	11541.547176
min	17.195000	12829.455100
25%	26.083750	20826.244213
50%	30.447500	34456.348450
75%	35.200000	41019.207275
max	52.580000	63770.428010

Figure 1. Descriptive statistics

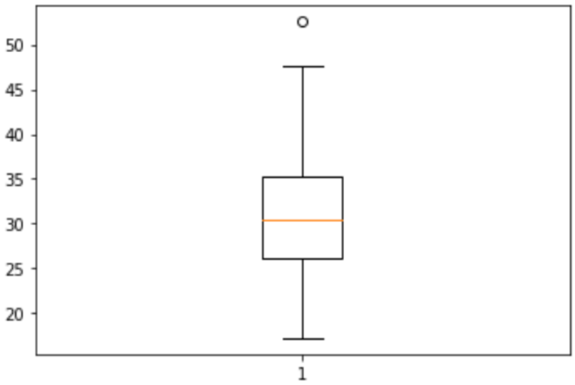


Figure 2. Boxplot of BMI

Using SciPy, it has been found that the BMI of the smoker patients and their medical charges have a very strong correlation with a correlation coefficient $r = 0.806481$, and a p-value $5.019668631794482e - 64 < 0.05$ indicating a significant positive linear relationship. This implies that as the BMI of a smoker patient increases, the medical charges increase as well.

Model

Using scikit-learn, the estimated regression model from the ordinary least squares method is given by

$$y = 1483.46313735x - 13474.70112447$$

where x - the BMI of the smoker patient
y - the medical charges

The regression model above indicates that the average value of the medical charges increases by approximately 1483.46313735 per increase in the smoker patient's BMI. In addition, the portion of the medical charges not explained by the BMI of the smoker patient is -13474.70112447. After splitting the data into test set and train set using the scikit-learn library, approximately 52.27% of the data fit the regression model.

Model Evaluation

The accuracy metrics considered to evaluate the model's performance are the coefficient of determination (r^2) and the mean absolute percentage error (MAPE). Since $r^2 = 0.6100751870181192$, then approximately 61% of the total variation of the medical charges is explained by the variation of the smoker patients' BMI. The mean absolute percentage error 0.19112567256721544 also implies that the average of the absolute percentage errors between the forecasted value and the actual value is approximately 19%.

Assumptions

In this section, we verify the assumptions of univariate linear regression, namely linearity, normality of error terms/residuals, homoscedasticity, and autocorrelation of residuals.

Linearity

Since linear regression requires a relationship between the variables, we will use visualization to determine the relationship between BMI and medical charges of smokers. From the scatter plot in *Figure 3* below, we can observe a linear relationship between the two variables. Thus, a linear model may be appropriate for the data.

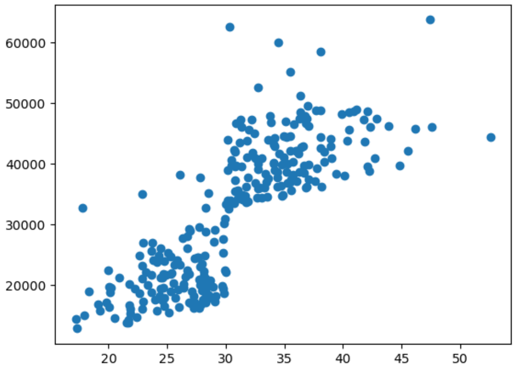


Figure 3. Scatter Plot (BMI vs. Medical Charges for Smokers)

Normality of error terms/residuals

Null Hypothesis (H_O)**:** The error terms are normally distributed.
Alternative Hypothesis (H_A)**:** The error terms are not normally distributed.

Using the Shapiro-Wilk test, we found the p-value $0.11940812319517136 > 0.05$. This is supported by the approximately bell-shaped Gaussian distribution plot and the quantile-quantile plot with values close to the line (see *Figure 4*). Thus, we fail to reject the null hypothesis and we conclude that the error terms are approximately normally distributed.

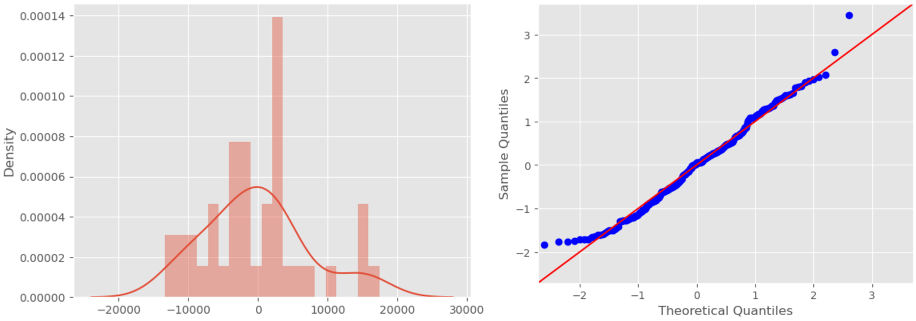


Figure 4. Plots for testing normality of error terms

Homoscedasticity

Null Hypothesis (H_O)**:** The error terms are homoscedastic.
Alternative Hypothesis (H_A)**:** The error terms are heteroscedastic.

Using the Breusch-Pagan test, the p-value is 0.2603027734470592, which is greater than 0.05. Therefore, we fail to reject the null hypothesis and we conclude that the error terms are homoscedastic.

Autocorrelation of residuals

Null Hypothesis (H_O)**:** Autocorrelation is not present in the error terms.
Alternative Hypothesis (H_A)**:** Autocorrelation is present in the error terms.

Using the Durbin-Watson test, we obtained the p-value $1.915 \approx 2$. Hence, we fail to reject the null hypothesis and we can conclude that there is no serial autocorrelation in the error terms at 5% level of significance.

Prediction of Medical Charges

After the assumptions for linear regression have been verified, the obtained model can now be used to predict the medical charges. In the table in *Figure 5* below, we have shown various values for body mass index and used the model to predict their corresponding medical charges.

Smoker's Body Mass Index	Predicted Medical Charges
18.5	$13969.37
24.9	$23463.53
29.9	$30880.85
34.9	$38298.16

Figure 5. Values of BMI and Predicted Medical Charges

Lastly, *Figure 6* below shows the predicted values using the BMI from the data set are plotted against the actual values.

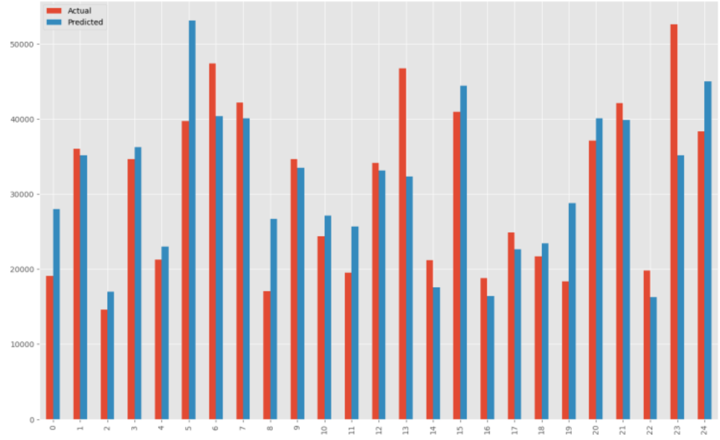


Figure 6. Actual Values vs. Predicted Values using the model

IV. Summary and conclusion

In general, a data set consisting of information on insurance customers' body mass index (BMI) and medical charges has been separated into two categories: smoker and nonsmoker. The BMI in the smoker category has then been used as a predictor for the target variable, the medical charges. It has been found by correlation analysis that the two variables have a strong positive linear relationship. Thus, we proceeded with determining the best line of fit from the machine learning regression techniques and chose the ordinary least squares method. From there, the estimated regression model was given by $y = 1483.46313735x - 13474.70112447$, where y is the medical charges and x is the BMI of the smoker. The model was evaluated using accuracy metrics (i.e., coefficient of determination and mean absolute percentage error) and the assumptions for linear regression were verified to determine the validity of the model. Finally, the model was used to predict the medical charges of the smoker patients. This can be utilized by ACME Insurance, Inc. to determine the annual insurance premium (amount paid every month) offered to every customer.

A multivariate regression analysis considering other possible predictors, instead of only the body mass index, could provide a better prediction for the medical charges. This paper is also limited only to the smoker category from the data set. Thus, future analyses could also be performed with the nonsmoker category. Lastly, the model can also be improved further to minimize the errors and improve the accuracy of the predictions of medical charges.

References

Crow, L., Tan, Y., Cavanaugh, T. M., Heaton, P., Diwan, T., Succop, P., ... & Boone, J. (2014). Impact Of Bmi On Charges And Reimbursement In Kidney Transplant Hospitalization Of Deceased And Living Donor Recipients. Value in Health, 17(3), A233.

Pan, W. H., Yeh, W. T., Chen, H. J., Chuang, S. Y., Chang, H. Y., Chen, L., & Wahlqvist, M. L. (2012). The U-shaped relationship between BMI and all-cause mortality contrasts with a progressive increase in medical expenditure: a prospective cohort study. Asia Pacific journal of clinical nutrition, 21(4), 577-587.

Singh, J. (2020). *Medical Charges Prediction*. Retrieved from https://www.kaggle.com/jagjeet555/medical-charges-prediction