**Background**

ACME Insurance, a leading health insurance company in the United States, aims to develop a mechanism to estimate the annual insurance premium charge for new customers. To achieve this objective, data for 1,338 existing customers has been obtained, including their age, sex, BMI (Body Mass Index), number of children, smoking habits, and location. Secondary data publicly available on Kaggle has been used for the project. To analyze and predict annual medical charges for potential, Minitab has been extensively used, employing tools such as descriptive statistics, scatterplots, correlation analysis, various graphs and plots, and linear regression modeling.

**Problem Statement**

Can a linear regression model be developed to accurately predict the annual medical charges for a potential customer based on relevant details collected from them? This study aims to identify the key variables that affect medical charges or costs and determine the extent of their impact. If we are able to estimate the value in ‘charges’ column using the historical data, then we should be in a position to predict charges for potential customers too. Various predictor variables will be considered to establish the relationship between the dependent variable (charge) and the independent variables. The findings of this study will aid in improving the accuracy of medical cost predictions and help insurance companies to better assess and manage risk.

**Analysis**

Collected data has been analyzed to discover patterns using descriptive statistical tools Mean, Median, standard deviation, skewness, distribution analysis, different charts and plots, tally tables etc.

**Statistics**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Variable** | **N** | **N\*** | **Mean** | | **SE Mean** | **StDev** | **Minimum** | **Q1** | **Median** | **Q3** | **Maximum** |
| age | 1338 | 0 | 39.207 | | 0.384 | 14.050 | 18.000 | 26.750 | 39.000 | 51.000 | 64.000 |
| bmi | 1338 | 0 | 30.663 | | 0.167 | 6.098 | 15.960 | 26.273 | 30.400 | 34.700 | 53.130 |
| children | 1338 | 0 | 1.0949 | | 0.0330 | 1.2055 | 0.0000 | 0.0000 | 1.0000 | 2.0000 | 5.0000 |
| charges | 1338 | 0 | 13270 | | 331 | 12110 | 1122 | 4734 | 9382 | 16687 | 63770 |
| **Variable** | **Skewness** | | |
| age | 0.06 | | |
| bmi | 0.28 | | |
| children | 0.94 | | |
| charges | 1.52 | | |

Chart

Description automatically generatedChart, histogram

Description automatically generatedChart, histogram

Description automatically generatedChart

Description automatically generated

**Tally**

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **sex** | **Count** | | **Percent** | | **children** | | **Count** | **Percent** | **smoker** | **Count** | **Percent** |
| female | 662 | | 49.48 | | 0 | | 574 | 42.90 | no | 1064 | 79.52 |
| male | 676 | | 50.52 | | 1 | | 324 | 24.22 | yes | 274 | 20.48 |
| N= | 1338 | |  | | 2 | | 240 | 17.94 | N= | 1338 |  |
|  |  | |  | | 3 | | 157 | 11.73 |  |  |  |
|  |  | |  | | 4 | | 25 | 1.87 |  |  |  |
|  |  | |  | | 5 | | 18 | 1.35 |  |  |  |
|  |  | |  | | N= | | 1338 |  |  |  |  |
| **region** | | **Count** | | **Percent** | |
| northeast | | 324 | | 24.22 | |
| northwest | | 325 | | 24.29 | |
| southeast | | 364 | | 27.20 | |
| southwest | | 325 | | 24.29 | |
| N= | | 1338 | |  | |

Chart, pie chart

Description automatically generated

Based on the initial analysis, the data shows a balanced representation of both male and female customers. The location of the customers is also evenly distributed across all four regions - Northeast, Northwest, Southeast, and Southwest. The typical ACME customer is around 35 years old and has a Body Mass Index (BMI) of 30.67. The majority of customers, around 43%, have no children, and almost one-fifth of the sample customers are smokers.

From the collected data, we can also infer that the average annual premium paid by customers is around $9,400. To ensure accuracy, the median was used instead of the mean, as the data is right-skewed.

We conducted a thorough analysis of the correlation among different variables using scatterplots. However, we found no significant correlation between individual variables. The highest correlation we observed was between charges and age, which was 0.3, indicating a moderate positive correlation between these variables.

A picture containing diagram

Description automatically generated

To further understand the data and the different relationships existing among, Probability plots and Bar charts have been used. Chart, line chart

Description automatically generated

In the above probability plot, blue denotes non-smokers and red denotes smokers. This plot signifies that annual medical charge for non-smokers is considerably lesser than the smokers.

Chart, bar chart

Description automatically generated

Bar chart of mean has been plotted to further investigate if medical charges is affected by the number of children that customers have. Interestingly enough, it has come to light that customers having 5 children are incurring the lowest medical charges.

We further investigated this with the help of Probability Plot for multiple variables Children and Smokers. The plot tells us that people with 4 or 5 children are mostly non-smokers.

Chart, histogram

Description automatically generated

This led us to believe that smoking can be significantly impacting the medical cost.

Further, we tried building simple linear regression models with charges as our dependent variable and Smoker being a predictor variable with confidence level of 95%.

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| charges | = | 8434 + 0.0 smoker\_no + 23616 smoker\_yes |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | 8434 | 229 | 36.83 | 0.000 |  |
| smoker |  |  |  |  |  |
| yes | 23616 | 506 | 46.66 | 0.000 | 1.00 |

**Model Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 7470.22 | 61.98% | 61.95% | 61.80% |

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 1 | 1.21520E+11 | 1.21520E+11 | 2177.61 | 0.000 |
| smoker | 1 | 1.21520E+11 | 1.21520E+11 | 2177.61 | 0.000 |
| Error | 1336 | 74554317947 | 55804130 |  |  |
| Total | 1337 | 1.96074E+11 |  |  |  |

This model explains 61.98% of the variation in the dependent variable. The adjusted R-square value, which adjusts for the number of predictors in the model, is 61.95%.

The Analysis of Variance table shows that the regression is significant with a p-value of 0.000 since it is lesser than level of significance 0.05. This means that the model is statistically significant and the independent variable "smoker" is a significant predictor of the dependent variable.

However, to ensure the validity of our model, we explored additional variables and combinations of variables present in our dataset.

Below is the linear regression model with charges as dependent variable and BMI as predictor variable.

**Regression Equation**

|  |  |  |
| --- | --- | --- |
| charges | = | 1193 + 393.9 bmi |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | 1193 | 1665 | 0.72 | 0.474 |  |
| bmi | 393.9 | 53.3 | 7.40 | 0.000 | 1.00 |

**Model Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 11873.9 | 3.93% | 3.86% | 3.62% |

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 1 | 7713391237 | 7713391237 | 54.71 | 0.000 |
| bmi | 1 | 7713391237 | 7713391237 | 54.71 | 0.000 |
| Error | 1336 | 1.88361E+11 | 140988645 |  |  |
| Lack-of-Fit | 546 | 77666813697 | 142246912 | 1.02 | 0.422 |
| Pure Error | 790 | 1.10694E+11 | 140119008 |  |  |
| Total | 1337 | 1.96074E+11 |  |  |  |

Though p-value of BMI is lesser than level of significance of 0.05 indicating that it is a significant variable, the adjusted R-square of the above model is only 3.86% which is extremely low suggesting that other variables need to be considered as well.

Through multiple iterations and testing, we arrived at a linear regression model that exhibited superior performance compared to the others and effectively addressed our problem statement.

**Optimal Multiple Regression Model**

Using the dependent variable 'charge' and continuous predictors 'BMI', 'Age', and 'number of children', as well as the categorical variable 'Smoker', a multiple linear regression model has been developed to analyze the relationship between these variables and their impact on charges. Confidence level has been considered as 95%.

**Regression Equation**

|  |  |  |  |
| --- | --- | --- | --- |
| **smoker** |  |  |  |
| no | charges | = | -12103 + 321.9 bmi + 257.8 age + 474 children |
|  |  |  |  |
| yes | charges | = | 11709 + 321.9 bmi + 257.8 age + 474 children |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | -12103 | 942 | -12.85 | 0.000 |  |
| bmi | 321.9 | 27.4 | 11.76 | 0.000 | 1.01 |
| age | 257.8 | 11.9 | 21.67 | 0.000 | 1.01 |
| children | 474 | 138 | 3.44 | 0.001 | 1.00 |
| smoker |  |  |  |  |  |
| yes | 23811 | 411 | 57.90 | 0.000 | 1.00 |

**Model Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 6067.79 | 74.97% | 74.89% | 74.74% |

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 4 | 1.46996E+11 | 36748942863 | 998.12 | 0.000 |
| bmi | 1 | 5088382317 | 5088382317 | 138.20 | 0.000 |
| age | 1 | 17296697217 | 17296697217 | 469.79 | 0.000 |
| children | 1 | 434769398 | 434769398 | 11.81 | 0.001 |
| smoker | 1 | 1.23448E+11 | 1.23448E+11 | 3352.91 | 0.000 |
| Error | 1333 | 49078450117 | 36818042 |  |  |
| Lack-of-Fit | 1326 | 48816632894 | 36814957 | 0.98 | 0.582 |
| Pure Error | 7 | 261817223 | 37402460 |  |  |
| Total | 1337 | 1.96074E+11 |  |  |  |

Chart, scatter chart

Description automatically generated

**Interpretation of Regression model**

As per the model, holding all other variables constant, for every one unit increase in BMI, the charges increase by $322. Similarly, for every one year increase in age, the charges increase by $258. And for every additional child, the charges increase by $474.

It also needs to be noted that the intercept is different for smokers and non-smokers, reflecting the fact that the medical charges for smokers and non-smokers are different even when all other variables are held constant.

The coefficients table is showing the constant term as -12,103 for non-smokers and 23811 for smokers, indicating that the average medical charges for smokers are significantly higher than for non-smokers. The coefficients for BMI, age, and children are the same for both smokers and non-smokers, indicating that the effect of these variables on charges is similar for both groups. Also, VIF (or Variance Inflation Factor) values are less than 1.01, suggesting that there is no significant collinearity among the predictor variables. Typically, a VIF value greater than 5 or 10 is considered to indicate significant collinearity.

Further, the R-square is 0.7497, which shows that the predictor variables explain about 75% of the variance in medical bills. The adjusted R-square is 0.7489, which adjusts for the number of predictor variables in the model.

Coming to Analysis of Variance table, it can be seen that the regression model as a whole is highly significant, since the p-value is lesser than 0.000 which is lesser than 0.05, our level of significance. Also, the individual predictor variables (BMI, age, children, and smoker) are also highly significant, indicating that they are each important predictors of the charges.

Additionally, the lack-of-fit test has a p-value of 0.582, denoting that there is no significant lack of fit between the model and the data. This suggests that the model is a good fit to the data and that the relationship between the predictor variables and charges is not significantly influenced by any factors outside the dataset.

Residual versus Fits plot shows that residual plots are bouncing randomly across the zero line, implying that the model is capturing the variation in the data, and the residuals (the difference between the actual values and the predicted values) are randomly distributed around zero. This suggests that there are no patterns or trends in the residuals, and the assumption about the linearity of the relationship is valid.

**Trying to improve the model further with interaction variables**

We also checked if model can further be improved by adding any interaction variables, just in case.

Keeping other predictors as is, we added interaction variables as ‘age\*BMI’, ‘age\*smoker’ and ‘BMI\*smoker’ just to check what impact it has on the prediction model.

**Regression Equation**

|  |  |  |  |
| --- | --- | --- | --- |
| **smoker** |  |  |  |
| no | charges | = | 53 + 205.8 age - 69.7 bmi + 1.98 age\*bmi |
|  |  |  |  |
| yes | charges | = | -20074 + 204.7 age + 1364.1 bmi + 1.98 age\*bmi |

**Coefficients**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Term** | **Coef** | **SE Coef** | **T-Value** | **P-Value** | **VIF** |
| Constant | 53 | 2045 | 0.03 | 0.979 |  |
| age | 205.8 | 49.6 | 4.15 | 0.000 | 26.96 |
| bmi | -69.7 | 65.9 | -1.06 | 0.290 | 8.95 |
| smoker |  |  |  |  |  |
| yes | -20127 | 1852 | -10.87 | 0.000 | 31.04 |
| age\*bmi | 1.98 | 1.57 | 1.26 | 0.208 | 37.38 |
| age\*smoker |  |  |  |  |  |
| yes | -1.1 | 23.9 | -0.05 | 0.963 | 8.95 |
| bmi\*smoker |  |  |  |  |  |
| yes | 1433.7 | 53.4 | 26.84 | 0.000 | 25.62 |

**Model Summary**

|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **R-sq** | **R-sq(adj)** | **R-sq(pred)** |
| 4908.14 | 83.65% | 83.57% | 83.46% |

**Analysis of Variance**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Source** | **DF** | **Adj SS** | **Adj MS** | **F-Value** | **P-Value** |
| Regression | 6 | 1.64011E+11 | 27335097917 | 1134.71 | 0.000 |
| age | 1 | 414761216 | 414761216 | 17.22 | 0.000 |
| bmi | 1 | 26959050 | 26959050 | 1.12 | 0.290 |
| smoker | 1 | 2843936735 | 2843936735 | 118.06 | 0.000 |
| age\*bmi | 1 | 38270030 | 38270030 | 1.59 | 0.208 |
| age\*smoker | 1 | 53016 | 53016 | 0.00 | 0.963 |
| bmi\*smoker | 1 | 17358405617 | 17358405617 | 720.57 | 0.000 |
| Error | 1331 | 32063634064 | 24089883 |  |  |
| Lack-of-Fit | 1302 | 31093106424 | 23881034 | 0.71 | 0.923 |
| Pure Error | 29 | 970527640 | 33466470 |  |  |
| Total | 1337 | 1.96074E+11 |  |  |  |

Though the adjusted R-square has improved and reached 83.57%, we are getting very high VIF for all predictor variables, indicating that there is a problem of endogeneity and this model is not a very good model to predict the dependent variable ‘charges’.

We are in a position now to conclude that our previous model was indeed the optimal model to predict charges that a customer will be incurring annually based on predictors BMI, age and children.

**Conclusion**

The multiple linear regression model developed for ACME Insurance has produced promising results, demonstrating that BMI, age, number of children, and smoking habits are key variables that affect medical charges. This model can accurately predict the annual medical cost for new customers based on relevant details collected from them. The findings of this study have significant implications for the insurance industry, as accurately assessing and managing risk is crucial for insurance companies. The high R-square value of 0.7497 indicates that the predictor variables explain about 75% of the variance in medical charges, which is a very good outcome for the model. The lack-of-fit test has also indicated that the model is a good fit to the data and that the relationship between the predictor variables and charges is not significantly influenced by any unmodeled factors. Overall, this study has contributed significantly to the development of an accurate and reliable mechanism for estimating medical charges, which can have a positive impact on the insurance industry and its customers.

**Reference doe data purposes**:

Trivedi, S. (2021). Health Insurance Price Prediction - Linear Regression. Kaggle. Retrieved from <https://www.kaggle.com/code/shubhamptrivedi/health-insurance-price-predict-linear-regression>.

**Project study performed by**:

1. Amandeep Singh
2. Dependra Nath Yogi
3. Madhukant Surgi

**Github Repository Location:**

<https://github.com/aman924/Data-Analytics-Final-Project>