***1.Research Scenario:***

The health insurance industry is critical in providing financial safety to individuals' medical expenses. Health insurance companies cover many healthcare service fees, including doctor visits, clinic stays, prescription drugs, and medical procedures. In recent years, the health insurance industry has faced several challenges, including rising healthcare costs, regulatory changes, and the ongoing impact of the COVID-19 pandemic. The global health insurance market was valued at $1.5 trillion in 2020. As the industry continues to evolve, it will be necessary for health insurance companies to remain focused on improving healthcare outcomes, reducing costs, and providing innovative products and services that meet the changing needs of consumers.

***1.1Research questions:***

What are the main factors contributing to health insurance costs in the United States, and how can we estimate the health insurance cost based on personal condition?

***2.About the Dataset***

A medical insurance company released this dataset containing information on nearly 1000 customers. The dataset includes details on people's medical history, such as whether they have certain medical conditions or have undergone surgeries, and their age, height, and weight. The data also includes the number of medical insurance premiums each individual pays. This dataset could help explore the relationship between medical history, demographics, and insurance premiums and provide insights into how insurance companies price their policies.

**Links to Dataset:**

<https://www.kaggle.com/datasets/tejashvi14/medical-insurance-premium-prediction>

**Content:**

1. Age: Age of customer.
2. Diabetes: Whether the person has abnormal blood sugar levels. [Yes, NO]
3. Blood Pressure Problems: Whether the person has abnormal blood pressure levels. [Yes, NO]
4. Any Transplants: Any major organ transplants. [Yes, NO]
5. Any Chronic Diseases: Whether customer suffers from chronic ailments like Asthma, Etc. [Yes, NO]
6. Height: Height of customer.
7. Weight: Weight of customer.
8. Known Allergies: Whether the customer has any known allergies. [Yes, NO]
9. History Of Cancer in Family: Whether any blood relative of the customer has had any form of cancer. [Yes, NO]
10. Number Of Major Surgeries: The number of major surgeries that the person has had. [0, 1, 2, 3]
11. Premium Price: Yearly premium price

***3.Statistical methods***

I will use correlation tests, multi-linear regression, and one-way ANOVA to investigate my research question.

***4.Result***

There is no missing data in this data set. This dataset is generated by 986 customers with 11 elements related to their health background.

***4.1 Check for collinearity***

图表, 瀑布图

描述已自动生成

First, use the correlation plot among quantitative data to check for collinearity. Since none of the independent variables is highly correlated (correlation > 0.8) to others, there is no evidence of collinearity. Therefore, all the quantitative independent variables are open to using.

***4.2 Multi-linear regression***

**To understand what determines the health insurance cost, we first examine the demographic variable (age, height, and weight). We will utilize a multi-linear regression model to determine how these variables contribute to health insurance costs.**

The scatterplot matrix below shows the associations between the continuous factors. As displayed in the scatterplot matrix, the premium price is strongly associated with age (correlation >= 0.60, very strong), weakly associated with weight (correlation = 0.10 to 0.19, weak), and negligible association with height (correlation < .10, negligible). Since high has a negligible correlation with the premium price, we will not consider that as a factor in our study.图片包含 图示

描述已自动生成

1. **Hypothesis:**

H0: βage =βweight=0 (Age, and weight are not significant predictors of premium price)

H1: βage≠0 and/or βweight≠0 (Age and/or weight are significant predictors of prestige score)

α=0.05

1. **Test statistic selected:**

F= df = 2, 983 (n−k−1)

1. **Decision rule:**

Reject H0 if F≥ 3.00488. Otherwise, do not reject H0.

1. **Global f-test:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals: |  |  |  |  |
| Min | 1Q | Median | 3Q | Max |
| -11988.1 | -2929.2 | -692.1 | 1446.6 | 22465.5 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | Estimate | Std. Error | t-value | Pr(>|t|) |
| (Intercept) | 6044.812 | 878.004 | 6.885 | 1.03e-11 \*\*\* |
| demo$Age | 313.413 | 9.988 | 31.380 | < 2e-16 \*\*\* |
| demo$Weight | 67.684 | 9.777 | 6.923 | 7.96e-12 \*\*\* |
| Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 | | | | |
| Residual standard error: 4376 on 983 degrees of freedom | | | | |
| Multiple R-squared: 0.5104, Adjusted R-squared: 0.5094 | | | | |
| F-statistic: 512.4 on 2 and 983 DF, p-value: < 2.2e-16 | | | | |

F= ≈512.4 df = 3, 982

1. **Conclusion**

Reject H0 since 512.4>3.00488. We have significant evidence at the α=0.05 level that age, weight, and height, when taken together, are significant predictors of premium price (here, p < 2.2e-16).

The least-squares regression equation for the data looking at the association between premium price and these factors was calculated to be:

ŷ =6044.812+313.413\*age+68.684\*weight

Take a closer look at how each variable influences the premium price. We can say this is a significant predictor when |t|≥1.96238.

We found that we have significant evidence at the α=0.05 level that **age** is a significant predictor of premium price after adjusting for height and weight. (31.38>1.962, p< 2e-16). **Weight** is also a significant predictor of premium price after height and age adjustment. (6.923>1.962, p=7.15e-12).

**Model assumption**

**图表, 直方图

描述已自动生成**

Based on the Residual plot, we can tell our model does **not** **meet the assumption** of linearity and constant variance. Since we only collect data from each participant once, we can assume that our data are independent. Based on the histogram, the distribution of the residuals is slightly right skewed.

By using R, we could find outliers in the premium price, and their ID is: 205, 296, 927, 929, 977, 985, and outliers in weight, and their ID is: 155, 159, 184, 187, 193, 196, 197, 204, 205, 208, 218, 219, 222, 226, 228, 229.

***4.3 ANOVA***

**After looking at how demographic variables (age, height, and weight) predict health insurance costs, we want to explore how personal history will influence health insurance costs. We will utilize the one-way ANOVA model to determine how each variable contributes to health insurance costs.**

The correlation plot below shows all correlations among different variables. Following the rule of thumb, we only look at those variables at least have a weak/small association with a premium price. The discrete independent variable we want to test on are blood pressure problem, transplants, chronic disease, and number of major surgeries.

**Rule of thumb for interpreting strength in our field:**

< .10 = negligible

.10 - .19 = weak/small

.20 - .39 = moderately strong

.40 - .59 = strong

>= .60 = very strong

***图表, 散点图

描述已自动生成***

***4.3.1 ANOVA - Blood pressure problem***

***图表, 箱线图

描述已自动生成***

There are 524 customers does not have a blood pressure problem and 462 customers who have a blood pressure problem. Customers with blood pressure problems (25448.05 on average) tend to pay higher than those without blood pressure problems (23356.87 on average).

1. **Hypothesis:**

H0:μYes=μNo (All underlying population means are equal)

H1:μi≠μj for some i and  j. (Not all the underlying population means are equal)

α=0.05

1. **Test statistic selected:**

F= df = 1, 984 (n−k)

1. **Decision rule:**

Reject H0 if F≥ 3.850926. Otherwise, do not reject H0.

1. **Test result:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals: |  |  |  |  |
| Min | **1Q** | **Median** | **3Q** | **Max** |
| -10448.1 | **-2448.1** | **-356.9** | **3551.9** | **15643.1** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| (Intercept) | **23356.9** | **269.3** | **86.747** | **< 2e-16 \*\*\*** |
| data$g1 | **2091.2** | **393.3** | **5.316** | **1.31e-07 \*\*\*** |
| Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 | | | | |
| Residual standard error: 6163 on 984 degrees of freedom | | | | |
| Multiple R-squared: 0.02792, Adjusted R-squared: 0.02693 | | | | |
| F-statistic: 28.26 on 1 and 984 DF, p-value: 1.31e-07 | | | | |

F= ≈28.26

1. **Conclusion**

Reject H0 since 28.26>3.85. We have significant evidence at the α=0.05 level that these groups are different in their premium price whether the costumer has abnormal blood pressure (here, p<0.001). The difference in premium price based on whether you have abnormal blood pressure is significant (here, p<0.001).

The regression model in this case is: ŷ =23356.9+ 2091.2groupyes.

***4.3.2 ANOVA – Transplants***

***图表, 箱线图

描述已自动生成***

There are 931 customers does not have a transplant and 55 customers who have a transplant. Customers who have a transplant (31763.64 on average) tends to pay higher than those customer does not have a transplant (23897.96 on average).

1. **Hypothesis:**

H0:μYes=μNo (All underlying population means are equal)

H1:μi≠μj for some i and  j. (Not all the underlying population means are equal)

α=0.05

1. **Test statistic selected:**

F= df = 1, 984 (n−k)

1. **Decision rule:**

Reject H0 if F≥ 3. 850926. Otherwise, do not reject H0.

1. **Test result:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals: |  |  |  |  |
| Min | **1Q** | **Median** | **3Q** | **Max** |
| -16764 | **-2898** | **-898** | **4102** | **16102** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| (Intercept) | **23898.0** | **196.1** | **121.845** | **< 2e-16 \*\*\*** |
| data$g3 | **7865.7** | **830.4** | **9.472** | **<2e-16 \*\*\*** |
| Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 | | | | |
| Residual standard error: 5985 on 984 degrees of freedom | | | | |
| Multiple R-squared: 0.08355, Adjusted R-squared: 0.08262 | | | | |
| F-statistic: 89.71 on 1 and 984 DF, p-value: < 2.2e-16 | | | | |

F= ≈89.71

1. **Conclusion**

Reject H0 since 89.71 >3.85. We have significant evidence at the α=0.05 level that these groups are different in their premium price whether the customer has a transplant (here, p<0.001).The difference in premium price based on whether to have a transplant is significant (here, p<0.001).

The regression model in this case is: ŷ = 23898.0 + 7865.7groupyes.

***4.3.3 ANOVA - Chronic disease***

***图表, 箱线图

描述已自动生成***

There are 808 customers who do not have a chronic disease, and 178 have a chronic disease. Customers with chronic disease (27112.36 on average) tend to pay higher than those customers does not have a chronic disease (23725.25 on average).

1. **Hypothesis:**

H0:μYes=μNo (All underlying population means are equal)

H1:μi≠μj for some i and  j. (Not all the underlying population means are equal)

α=0.05

1. **Test statistic selected:**

F= df = 1, 984 (n−k)

1. **Decision rule:**

Reject H0 if F≥ 3. 850926. Otherwise, do not reject H0.

1. **Test result:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals: |  |  |  |  |
| Min | **1Q** | **Median** | **3Q** | **Max** |
| -9112.4 | **-4112.4** | **-725.2** | **4274.8** | **15274.8** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| (Intercept) | **23725.2** | **215.1** | **110.307** | **< 2e-16 \*\*\*** |
| data$g5 | **3387.1** | **506.2** | **6.691** | **3.71e-11 \*\*\*** |
| Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 | | | | |
| Residual standard error: 6114 on 984 degrees of freedom | | | | |
| Multiple R-squared: 0.04352, Adjusted R-squared: 0.04255 | | | | |
| F-statistic: 44.77 on 1 and 984 DF, p-value: 3.713e-11 | | | | |

F= ≈44.77

1. **Conclusion**

Reject H0 since 44.77 >3.85. We have significant evidence at the α=0.05 level that these groups are different in their premium price by whether the customer has the chronic disease (here, p<0.001). The difference in premium price is based on whether the costumer have or not have the chronic disease is (here, p<0.001).

The regression model in this case is: ŷ = 23725.2 + 3387.1groupyes.

***4.3.4 ANOVA - Number of major surgeries.***

***图表, 箱线图

描述已自动生成***

There are 479 customers does not have major surgeries before, 372 customers who have had one major surgery before, 119 customers who have had two major surgeries before, and 16 customers who have had three major surgeries before. Customers who have had two major surgeries before (28084.03 on average) tend to pay the highest, and customers who have had three major surgeries before (28000.00 on average) follow closely. Then customers who have had one major surgery before (24741.94 on average) place in third. Customers who have had 0 major surgeries before (22968.68 on average) tend to pay the least amount of premium price.

1. **Hypothesis:**

H0:μ0=μ1=μ2 =μ3 ( (All underlying population means are equal)

H1:μi≠μj for some i and  j. (Not all the underlying population means are equal)

α=0.05

1. **Test statistic selected:**

F= df = 3, 982 (n−k)

1. **Decision rule:**

Reject H0 if F≥ 2.613967. Otherwise, do not reject H0.

1. **Test result:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Residuals: |  |  |  |  |
| Min | **1Q** | **Median** | **3Q** | **Max** |
| -9741.9 | **-1968.7** | **31.3** | **2031.3** | **16031.3** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Coefficients: |  |  |  |  |
|  | **Estimate** | **Std. Error** | **t-value** | **Pr(>|t|)** |
| (Intercept) | **22968.7** | **275.1** | **83.477** | **< 2e-16 \*\*\*** |
| data$g7 | **1773.3** | **416.2** | **4.261** | **2.23e-05 \*\*\*** |
| data$g8 | **5115.3** | **616.8** | **8.293** | **3.60e-16 \*\*\*** |
| data$g9 | **5031.3** | **1530.4** | **3.288** | **0.00105 \*\*** |
| Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1 | | | | |
| Residual standard error: 6022 on 982 degrees of freedom | | | | |
| Multiple R-squared: 0.07394, Adjusted R-squared: 0.07111 | | | | |
| F-statistic: 26.14 on 3 and 982 DF, p-value: 2.871e-16 | | | | |

F= ≈26.14

1. **Conclusion**

Reject H0 since 26.14> 2.61. We have significant evidence at the α=0.05 level that these groups are different in their premium price by the number of major surgeries a customer has (here, p<0.001).The difference in premium price between a customer having 0 major surgery compares to having one major surgery (here, p<0.001), having two major surgeries (here, p<0.001), and having three major surgeries (here, p=0.00105) are all significant.

The regression model in this case is:

ŷ = 22968.7 + 1773.3group1 + 5115.3group2 + 5031.3group3.

***5 Conclusion and limitation***

***5.1 Conclusion***

This study aimed to investigate the factors contributing to the cost of health insurance in the United States and to estimate the health insurance cost based on personal conditions using a dataset containing information on nearly 1000 customers.

Through multi-linear regression analysis, it was found that age and weight were significant predictors of health insurance cost. Premium price was strongly associated with age and weakly associated with weight.

The ANOVA tests were used to examine the relationship between various medical conditions and premium prices in a health insurance company. The results showed significant differences in premium prices between customers with transplants, chronic diseases, blood pressure problems, and the number of major surgeries. These findings suggest that certain medical conditions can impact health insurance premiums.

***5.2 limitation***

One of the limitations of this study is that the dataset used is limited to one medical insurance company. Therefore, it may not be generalizable to the entire population. Moreover, the data does not contain information about customers' health habits, such as exercise, diet, and smoking, which could influence the cost of health insurance.

Other than that, this research may commit the type 1 error (rejecting a null hypothesis when it is true). In this study, I chose a significance level (alpha) of 0.05, meaning there is a 5% chance of rejecting the null hypothesis when it is true. Future research could decrease the significance level (alpha) to test for the result.

In summary, while this study provides valuable insights into the factors that contribute to the cost of health insurance in the United States, there are limitations to the dataset used and the analysis conducted. Further research is needed to explore the broader context of the cost of health insurance, including other factors such as medical procedures and prescription drug costs.