|  |  |
| --- | --- |
|  |  |

|  |  |
| --- | --- |
| health insurance claims in the United States | Logo  Description automatically generated  **Statistics for Data Science**  *April 12, 2021*  **Group 13**  Ifenna Okeke  Daniel Park  Wendy Porter  Jessica Qu  David Ravenscroft |
|  |  |

**Background**

Large insurance companies in the United States have, until recently, been insulated from the forces of rapid technological change thanks to the nature of the insurance industry. Any new competitors face heavy regulation, large in-force books, and require an exceptionally large amount of capital to carry the risk. Without these competitive pressures, the industry has been slow to take advantage of technological advancements and the boom in big data that could provide better accuracy in the underwriting and pricing processes, as well as in the fraudulent claims detection process.

New, smaller firms that have embraced big data and technology are starting to break through the large insurance company strongholds. Upstarts like Lemonade use AI and chatbots to provide quotes and process claims in minutes.[[1]](#footnote-1) Health and Insurance consulting firms, such as X by 2, are even proposing that insurance companies can extend their predictive capabilities by leveraging nationally available data, such as weather patterns to mitigate the effects of storm damage.[[2]](#footnote-2)

Therefore, the time has come for insurance companies to consider their own technological journeys: “Companies that neglect to adapt to new business practices will be at a disadvantage by those that use digital technology to reduce costs and get better returns on their investments.”[[3]](#footnote-3)

**Objective**

To better understand how insurance companies can use data analysis to expedite processes such as underwriting, an analysis will be conducted on health insurance claims in the United States, as found on the Kaggle website.[[4]](#footnote-4)

The main objective of the study is to determine which health risk factors more reliably correlate with individual medical charges billed to the insurance company (‘charge’), thereby understanding which health risk factors may result in higher health insurance premiums.

The main hypothesis statement for the project is as follows:

*“The most influential health risk factors on charges can be identified through appropriate hypothesis testing and linear regression analysis to produce a predictive health insurance premium model.”*

The analysis will be restricted to the parameters as provided in the dataset(Appendix A). While there may be external factors that may influence the charges such as proximity to environmental toxins, global pandemics, genetic predisposition to certain illnesses, etc. these factors are beyond the scope of this analysis and will not be considered.

**Data Overview**

1. There are 1338 health insurance claims records for the United States (time frame is not indicated). The entries contain 7 attributes for each claim with no null values.
2. The column names were examined to ensure that they did not contain any unusual characters or ambiguous names. Column names appeared clear, and no unusual characters were found.
3. Examining the number of unique values for each attribute did not reveal any extraordinary values and all the discrete and continuous variables appeared to have a good range of values, thereby allowing for a robust analysis.

**Data Preparation**

***Data Errors***

As mentioned above, no values appeared to need correction.

***Feature Engineering***

With the intention of conducting linear regression further in the analysis, some of the categorical values (‘smoker’, ‘sex’, ‘region’, ‘bmi’) were converted to dummy variables to become binary yes/no variables. As is recommended for building a regression analysis with dummy variables, the n-1 of the columns for each category was added to the existing dataset. For example, converting the smoker category to dummy variables produced two columns: ‘smoker\_yes’ and ‘smoker\_no’. However, only the ‘smoker\_yes’ column was concatenated to the existing dataset.[[5]](#footnote-5)

No further feature engineering was deemed necessary during this phase of the analysis.

***Outliers***

At this juncture, the removal of outliers was deemed unnecessary.

**Exploratory Data Analysis**

A visual examination of the distribution of charges reveals that there is a downward trend of frequency as the total dollar amount of charges increases. The charges range in value from $1,121 to $63,770, with a mean of $13,270.

***Quantitative Variables***

Scatterplots were used to examine the relationship between charges and the continuous and discrete variables in the data set:

* Age
* BMI
* Number of Children

1. **Age** – Three (3) distinct bands of charges were revealed that spanned all the ages: approximately $0 - $10k, $15k - $30k, and $35k - $45k. When overlayed with the categorical variable ‘sex’, there appeared to be an even distribution of gender across the scatterplot, as well as across all 3 bands of charges. However, when overlayed with the categorical variables ‘smoker’ and ‘bmi\_group’:
   1. The highest band of charges consisted almost exclusively of smokers
   2. The middle band of charges contained an even mixture of smokers and non-smokers
   3. The lowest band of charges consisted almost exclusively of non-smokers

Lastly, when overlaid with ‘bmi\_group’:

* 1. The highest band of charges consisted almost exclusively of those with a high BMI
  2. The middle band of charges contained an even mixture of those with a low and high BMI
  3. The lowest band of charges contained an even mixture of those with a low and high BMI

1. **BMI** - This feature has a somewhat upward trend in charges as BMI increases. However, it must be noted that an inspection of BMI revealed that in the overwhelming majority of claims (1093/1338 or 81.7%), the individuals had BMIs that fell in the Overweight and Obese categories (Appendix B). When overlayed with the categorical variable ‘sex’, there appeared to be an even distribution of gender across the scatterplot. However, when overlayed with the categorical variable, ‘smoker’, the highest charges involved the smokers. Moreover, the charges for smokers increased more rapidly as their BMI increased, than did the charges for non-smokers.
2. **Children** – This is a discrete variable. The dollar values of the charges remain consistent for 0-3 children but decreases sharply from 4 children onward. However, a previous histogram revealed that the frequency of claims declined rapidly as the number of children decreased, with very few claims for individuals with 4 or more children. When overlayed with the categorical variable ‘sex’, there appeared to be an even distribution of gender across the number of children. However, when overlayed with the categorical variable ‘smoker’, the highest charges for each number of children were for smokers. Similarly, when overlayed with the categorical variable, ‘bmi\_group’, the highest charges for each number of children were for those with a high\_bmi.

Of the quantitative variables, only ‘**age’** and ‘**bmi’** demonstrate a relationship with charges.

***Categorical Variables***

Boxplots were used to examine the relationship between the sale price and the categorical variables in the data set:

* Smoker
* Gender
* BMI Grouping
* Region

1. ***Smoker*** - The median charges of smokers is much higher than for non-smokers. The distribution of charges for smokers is also much larger than for non-smokers. However, non-smokers appear to have more outliers than smokers.
2. ***Gender*** – The median charges of males and females appear to be equivalent. Females have a larger number of outliers, but a smaller distribution than men.
3. ***BMI Grouping*** - The median charges of those with a low BMI and a high BMI appear to be equivalent. Those with a high BMI have a larger number of outliers and a larger distribution than those with a low BMI.
4. ***Region*** – The median charges across all 4 regions appear to be fairly equivalent. However, the mean and the distribution of the data for the Southeast region are higher.

Of the categorical variables, smoker (y/n) shows the highest likelihood of correlation with charge.

***Correlation Matrix***

A correlation matrix was plotted to check correlation between different variables and charge, as well as the different variables to each other.

Table

Description automatically generated

The matrix confirmed the observations made earlier in our exploratory analysis, i.e., that some of the variables that show a relationship to charge are:

* Smoker
* Age
* BMI

**Hypothesis Testing**

In addition to the correlation matrix from the EDA, a variable’s influence on a claimant’s charges can be determined by comparing the mean charges of two populations. For instance, the mean charges of smokers versus non-smokers. A two-sided t-test provides a **t-statistic** and **p-value**.

The larger the t-statistic, the greater the difference between the two groups. The p-value, usually expressed as percentage, indicates how likely the results are to have occurred by chance. A p-value of 0.05, for instance, means that there is only a 5% chance that the results from the experiment occurred by chance. The critical value, which defines whether a p-value is **statistically significant** and acceptable or not is typically pre-defined at 5%.

For the purposes of our hypothesis testing, the critical value was set to 5%. In other words:

p\_value > 0.05: Fail to reject the null hypothesis of the statistical test

p\_value ≤ 0.05 Reject the null hypothesis of the statistical test

***Smokers vs Non-Smokers***

Specifically, the hypothesis test was used to determine whether the mean value of charges differed between claimants who smoked and claimants that did not. The hypotheses being tested were:

*H0: There is no significant difference in the charges between smokers and non-smokers.*

*H1: There is a significant difference in the charges between the smokers and non-smokers.*

The dataset contained 274 claimants who are smokers. Since a significantly larger proportion of claimants are non-smokers, 274 non-smoking claimants were chosen at random for testing purposes.

A simple visual comparison using boxplots and tables indicated that not only did the smokers and non-smokers have disparate means, but that the group of smokers had a larger range and distribution variability of charges. Comparing these two groups using a t-test confirmed that these differences are statistically significant:

|  |  |
| --- | --- |
| **t-statistic** | 30.77 |
| **p-value** | 2.50e-121 |

Given that the p-value is far below our critical value of 5%, we reject the null hypothesis that there is no significant difference in the charges between smokers and non-smokers. In other words, the average charge for claimants who smoked is statistically higher than the average charge for claimants who did not smoke.

***Age***

As mentioned during the EDA, the age variable displayed an unusual trend of clustering around 3 distinct bands. This trend is also present when the dataset is limited to smoking claimants only. In order to determine whether the age variable influences the charges of smokers, these bands were used to create two separate datasets (a lower charge group/band and a higher charge group/band) and the following hypotheses were tested:

*H0: There is no significant difference in the mean ages of smokers between the lower charge group and the higher charge group*

*H1: There is a significant difference in the mean ages of smokers between the lower charge group and the higher charge group*

A simple visual comparison using boxplots and tables indicated that the low charge group and the high charge group had similar mean ages, as well as a similar distribution of ages. Comparing these two groups using a t-test confirmed that these differences are not statistically significant:

|  |  |
| --- | --- |
| **t-statistic** | -1.09 |
| **p-value** | 0.29 |

Given that the p-value is above our critical value of 5%, we fail to reject the null hypothesis that there is no significant difference in the ages of smokers between the low charge group and the high charge group. In other words, the average age for claimants in the low charge group and the high charge group are very similar and is not an influence on whether a claimant’s charge is in one band or another.

**Ordinary Least Squares (OLS)**

In the simplest of terms, a linear regression shows how much one thing affects another. For example, starting from when the outdoor temperature is at 10˚C, as the temperature increases by one degree, an ice cream shop sells 10 more ice cream cones. One of the most common methods used to define this relationship is called **Ordinary Least Square (OLS)**. This method minimizes the sum of squared residuals - the difference between the observed and actual dependent variable values squared. One of the key indicators of how well the resultant model explains the relationship between the variables is the R-Squared value. This value describes the percent variation in the dependent variable that is explained by the independent variable(s); the higher the R-Squared value the better.

The OLS method was used to further refine our list of variables that appear to have a correlation to charges. Each variable was examined independently first, then in combinations. The linear regression model clearly shows that 'smoker\_yes' has the largest impact on charges with an R-Squared value of (0.62), which is larger than for any of the other independent variables. We therefore select the 'smoker\_yes' variable to take forward to the next stage of the model analysis – determining whether a combination of variables provides a more accurate prediction of charges.

This next step of the analysis shows that the 'age' variable has the greatest impact on R-Squared, increasing from 0.62 to 0.721. We therefore take this forward to the next step, along with the 'smoker\_yes' variable. The next variable we add to the model is 'bmi', which increases the R-Squared value from 0.721 to 0.747. At this step, the increases in R-Squared are very minor. The maximum increase in R-Squared is for the 'children' variable, which only increases R-Squared from 0.747 to 0.750. At this stage of the analysis, there is no further increase in R-Squared. We therefore keep only the saved parameters ('smoker\_yes', 'age', 'bmi' & 'children').

**R-Squared Values for the Independent Variables**

|  |  |
| --- | --- |
| **Independent Variable(s)** | **R-Squared** |
| smoker\_yes + age + bmi + children | 0.75 |
| smoker\_yes + age + bmi + sex\_male | 0.747 |
| smoker\_yes + age + bmi + region\_northwest | 0.748 |
| smoker\_yes + age + bmi + region\_southeast | 0.748 |
| smoker\_yes + age + bmi + region\_southwest | 0.748 |

Although additional combinations of variables were tested, the R-Squared value did not increase. Additionally, influence plots were run on each variable to determine whether certain outliers positively or negatively influenced the regression analysis. After all identified influencers were removed, the regression analysis was executed. The R-Squared only marginally improved, thereby confirming the original conclusion. In summary, the variables that have an impact on the insurance charges are, in descending order of impact:

* Smoker
* Age
* BMI
* Number of Children

**Conclusion**

An analysis of US health insurance claims data was conducted to determine whether certain health factors could be identified and statistically proven to influence the total dollar amount that an insured claimed. Specifically, the main hypothesis is:

*The most influential health risk factors on charges can be identified through appropriate hypothesis testing and linear regression analysis to produce a predictive health insurance premium model.*

After reviewing the selected data using simple exploratory data analysis, followed by hypothesis testing and linear regression analysis, ***the conclusion is if a claimant smoked, their charges tended to increase as their age, BMI, and the number of children increased***.

***Next Steps***

Although the provided health insurance claims data is robust, there were some limitations identified through our data preparation phase. The model would benefit from having a more evenly distributed population of smokers vs non-smokers. Also, the data would benefit from a more even distribution of claimants who have a BMI within the normal range vs overweight/obese.

Lastly, with the benefit of more evenly distributed data, the model produced by linear regression can then be applied to machine learning models, thereby automating the underwriting process, and saving insurance companies time/money by expediting the underwriting process.

**Appendices**

**Appendix A**

***Full Description of Variables Included in the Dataset***

|  |  |
| --- | --- |
| **Variable** | **Description** |
| **age** | Age of primary beneficiary |
| **sex** | Insurance contractor gender (female / male) |
| **bmi** | Body mass index, providing an understanding of body, weights that are relatively high or low relative to |
| **children** | Number of children covered by health insurance / Number of dependents |
| **smoker** | Smoker / Non - smoker |
| **region** | The beneficiary's residential area in the US, northeast, southeast, southwest, northwest |
| **charges** | Individual medical costs billed by health insurance |

**Appendix B**

***BMI Categories***

|  |  |
| --- | --- |
| **Category** | **Description** |
| **Underweight** | <18.5 |
| **Normal weight** | 18.5–24.9 |
| **Overweight** | 25–29.9 |
| **Obese** | 30 or greater |

1. Meckbach, Greg. (July 3, 2020) *How insurtech Lemonade’s IPO is going so far*. Canadian Underwriter. https://www.canadianunderwriter.ca/insurance/how-insurtech-lemonades-ipo-is-going-so-far-1004193926/ [↑](#footnote-ref-1)
2. Burhani, Yunus. (March 1, 2019) *How Insurers Can Really Embrace Big Data..*X By 2. https://xby2.com/how-insurers-can-really-embrace-big-data/ [↑](#footnote-ref-2)
3. Digital McKinsey.(March 2017) *Digital disruption in insurance: Cutting through the noise.* https://www.mckinsey.com/~/media/mckinsey/industries/financial%20services/our%20insights/time%20for%20insurance%20companies%20to%20face%20digital%20reality/digital-disruption-in-insurance.ashx [↑](#footnote-ref-3)
4. *US Health Insurance Dataset: Insurance Premium Charges in US with important details for risk underwriting.* Kaggle. https://www.kaggle.com/teertha/ushealthinsurancedataset?select=insurance.csv [↑](#footnote-ref-4)
5. When categorical variables are converted to dummy variables, the regression test looks at the dummy variable in comparison to something else. If the charges are higher for a smoker, it is higher relative to something. Therefore, the dropped variable becomes the anchor, the thing with which the variable is compared to. [↑](#footnote-ref-5)