Analysis of factors surrounding medical insurance premiums

Haoying Shen

June 18, 2021

Abstract

This study uses the insurance data that includes factors such as age, BMI, number of children, region, and smoking. This data was analyzed for the magnitude of effects on the medical insurance cost and potential discontinuities in numerical data. The data is obtained through then Machine Learning with R book written by Brett Lantz. The data was analyzed using a linear regression model to determine the general effects of each of the predictors. BMI and age were investigated further using graphs to see if any particular ages or BMI showed unique changes, which would be represented as a discontinuity on the plot. From the study we have found a clear discontinuity when an individual's BMI is 30, leading to a jump in insurance cost. Whereas an increase in age showed a steady increase in cost. Overall, this study presented the magnitudes of the factors and a specific value of BMI that caused a large increase in cost. Future studies should include medical history.

Introduction

Medical insurance premiums have always been calculated differently between individuals, and what causes the prices to fluctuate is understood generally, however the specifics behind what causes price differentials is not clear [1]. This study will investigate the magnitude of effect on insurance premiums based on BMI, age, number of children, smoking, and region of residence. The data that will be used is a dataset that includes the cost of insurance for a vast set of individuals for an unnamed insurance company. Understanding the magnitude of effect for the factors and what cut offs are in place that will vastly change the price will allow individuals to focus on what to change or plan for to get the best deal for insurance. One example of potential cut off lies in BMI, where 30+ indicates obesity [2]. This study will aim to determine how much of a change in premium these cut offs will induce. This leads us to our research question of whether BMI, age, number of children, smoking, and region of residence effect the price of medical insurance.

Data

Data Collection Process

The dataset used is obtained through the Machine Learning with R book written by Brett Lantz, where the dataset is freely available through GitHub at: https://github.com/stedy/Machine-Learning-with-R-datasets [3]. The dataset column names were cleaned to have their first letter be capitalized, with the exeption of the charges and bmi column which was renamed to "Cost" and capitalized to "BMI" respectively. The numerical data was also cleaned by rounding the data to the nearest integer.

Data Summary

The dataset includes some of the factors that may determine the medical insurance premiums that individuals must pay. This includes factors such as BMI, number of children, age, region, sex, and whether they smoke or not.

The rounding of the numerical variables was done using the `mutate_at(vars(Cost, BMI), funs(round(., 0)))` function. The renaming was done directly inside the csv.

Variables

Age - A continuous numerical variable that indicates the age of the individual for that observation.

Sex - A binary categorical variable that indicates the sex of the individual for that observation.

BMI - A continous numerical variable that indicates the body mass index of the individual for that observation.

Children - A discrete numerical variable that indicates the number of children an individual has.

Smoker - A binary categorical variable that indicates if the individual in the observation smokes or not.

Region - A categorical variable that determines the area of residence of the individual.

Cost - A continous numerical variable that indicates the individual's insurance premium.

Table 1: Frequency and Percentage of Smokers

	no	yes
Frequency	1064	274
Percentage	80	20

Table 2: Frequency and Percentage of Sexes

	female	male
Frequency	662	676
Percentage	49	51

Table 3: Frequency and Percentage of Region

	northeast	northwest	southeast	southwest
Frequency	324	325	364	325
Percentage	24	24	27	24

Table 4: Summary and spread of numerical data

	Age	BMI	Children	Cost
Min.	18	16	0	1122
1st Qu.	27	26	0	4740
Median	39	31	1	13983
Mean	40	32	2	19292
3rd Qu.	51	35	2	18154
Max.	64	53	5	63770

All tables for this report was programmed using `kableExtra` [4].

Table 1 shows the frequency and percentage of smokers to non-smokers for this dataset, we can see that only 20% of the sample smokes. Table 2 presents the percentage and frequency of the sexes within the dataset, where a relatively even distribution between males and females can be seen. Table 3 illustrates the frequency

and percentage of region, where most of the respondents come from the southeast region. Table 4 shows the summary and spread of the numerical variables. Age shows a distribution between 18 and 64, where the mean age is 40. BMI shows a distribution between 16 and 53, where the mean BMI is 32. The number of children ranges from 0 to 5, where the mean number of children is 2. The cost of the insurance ranges from 1122 to 63770 dollars, with the average being $19292.

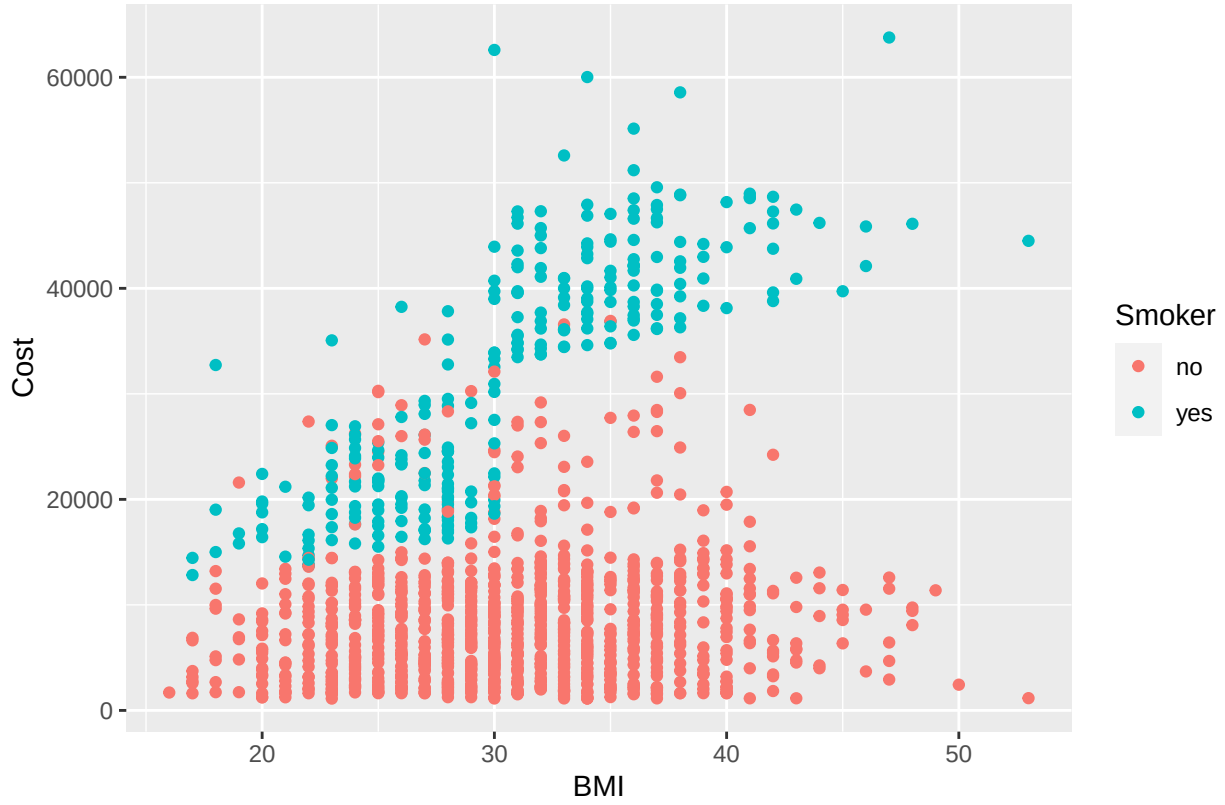

Figure 1. Insurance cost by BMI seperated by smokers and non−smokers

Figure 1 presents the cost of insurance in comparison with BMI, and is separated by smokers and non-smokers. From the graph we can see that in general as BMI increases the cost of insurance also seems to increase. Additionally, smokers generally pay more for the same BMI as non-smokers. More importantly, for the smoker category, what seems to be a regression discontinuity can be seen at a BMI of 30.

All analysis and figures for this report was programmed using `R version 4.0.4`[5].

Methods

Model Specifics

For this study, the assumption that none of the variables share any relationship will be present. We will create a linear regression model to determine the effects of BMI, age, sex, region, number of children and smoking on the cost of medical insurance cost. We will then investigate the graphs of each individual predictor to discover any regression discontinuities, which are jumps in the graph that may indicate a cut off point where the response variable, cost in this case greatly changes. The model used will can be represented as:

$$Y = \beta_0 + \beta_1 x_{bmi} + \beta_2 x_{age} + \beta_3 x_{sex} + \beta_4 x_{region} + \beta_5 x_{child} + \beta_6 x_{smoke}$$

Where Y represents the cost of the medical insurance. β_0 represents the baseline or the intercept which occurs when all other predictors are 0. β_1 is the effect on cost that BMI contributes. β_2 represents the effect

on cost based on age. β_3 represents the contribution to determining insurance cost that sex provides. β_4 is the effect on cost based on region. Finally, β_5 and β_6 represent the effect on cost that number of children and smoking contributes, respectively.

Regression discontinuities

To determine any regression discontinuities, we will split the data based on smokers and non-smokers, and then graph the relationship between each of the numerical predictors Age and BMI separately. This will allow us to determine which values will have unique effects on the response.

Results

Table 5: Summary of estimates for the linear regression model on insurance costs

term	estimate	std.error	statistic	p.value
(Intercept)	-11926.859	987.486	-12.078	0.000
BMI	338.981	28.601	11.852	0.000
Age	256.848	11.900	21.584	0.000
Sex: Male	-125.226	332.944	-0.376	0.707
Region: Northwest	-362.523	476.309	-0.761	0.447
Region: Southeast	-1033.336	478.707	-2.159	0.031
Region: Southwest	-973.574	478.064	-2.036	0.042
Children	473.708	137.815	3.437	0.001
Smoker	23844.161	413.175	57.710	0.000

Summary statistics calcuated using r's `summary()` and cleaned using the `broom` package [5,6].

From the summary table in table 5, we can see that the intercept shows a negative estimate, this is due to assuming all other effects are 0, which indicates an age of zero as well which makes sense as children should not need their own medical insurance. BMI indicates that for each point of increase the medical cost will increase by 338.98 dollars. The estimate for age indicates that for each increase in age 256.85 dollars will be added to cost. The table also shows that males tend to have a lower insurance cost in comparison to females by 125 dollars. In comparison to the baseline region of northeast, the northwest is cheaper by $362, the southeast region is cheaper by $1033, and the southwest region is cheaper by $973. Furthermore, the table shows that each children increases the insurance cost by $473.71. Finally, smoking increases the cost of insurance by $23844 on average. All of the predictor's part from sex and being in the northwest region show significance under the 95% confidence interval as the p-values are less than 0.05.

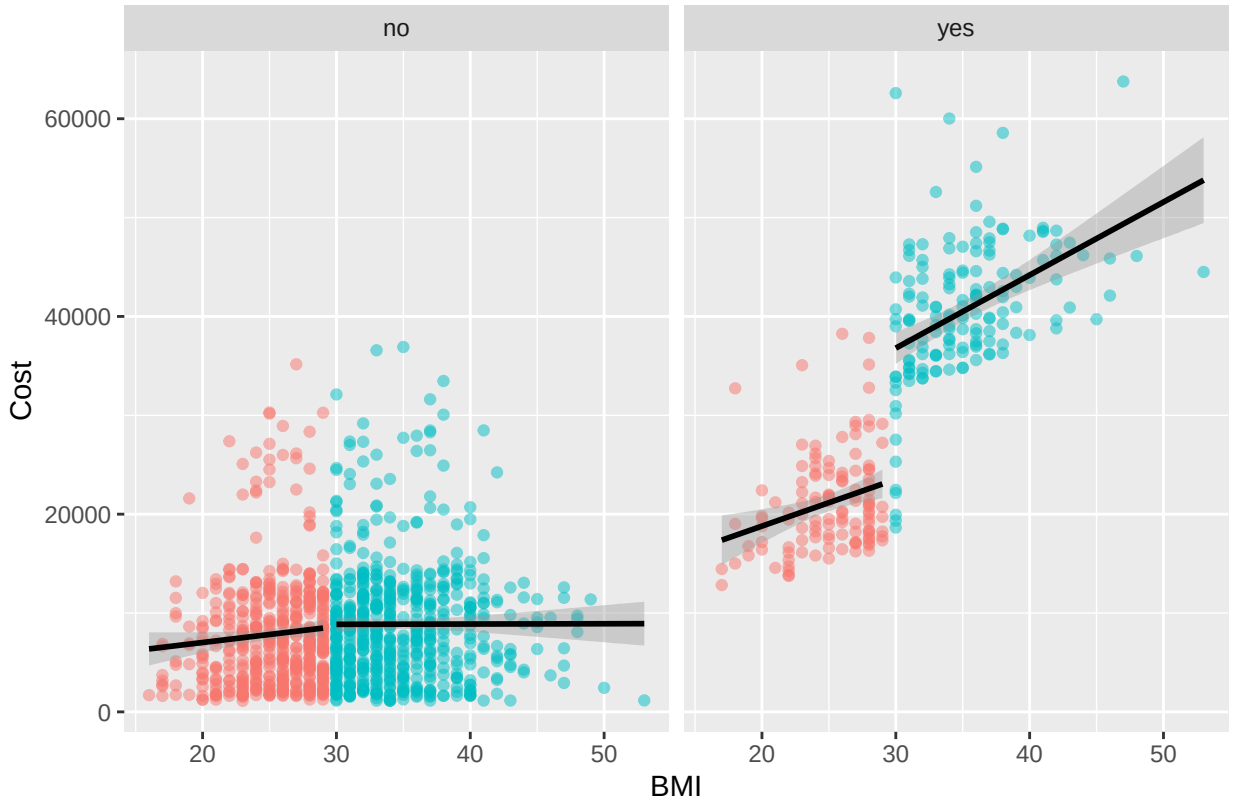

Figure 2. Insurance cost by BMI faceted by smokers and non−smokers

Figure 2 shows the relationship between BMI and Cost, which is separated between smokers and non-smokers. We can see that for both groups as BMI increases so does the cost of insurance. As for the smokers, we see a clear regression discontinuity at 30, which is the point at which people are determined to be obese on the BMI scale [2]. This indicates a large jump caused by the classification of obesity.

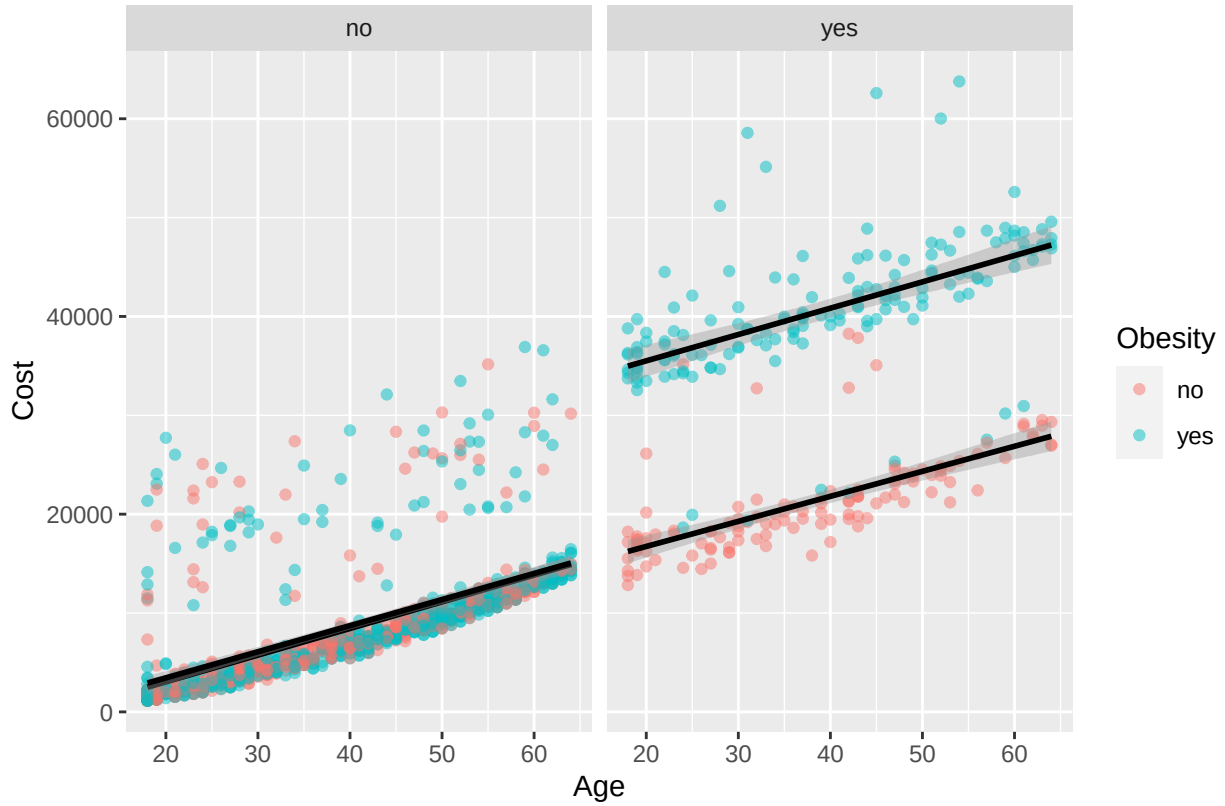

Figure 3. Cost by Age seperated by smoking and BMI below and above 30

Figure 3 shows the relationship between Age and Cost, which is separated between smokers and non-smokers. Additionally, we have coloured the points by those above and below the BMI of 30, which indicates obesity. From the relationship, we can clearly see that as age increases so does the insurance premium. On the smoker side we can see a clear separation between those who are obese and those who are not. Reinforcing the regression discontinuity for obese smokers shown in figure 2.

Overall, the results obtained seem reasonable, as those who are obese tend to be at higher risk, which would increase the insurance premium, the regression discontinuity when BMI is greater than 30 illustrate this precisely, as when BMI increases above 30, it is deemed to be obese [2]. Additionally, as one ages, the risk will also increase making it reasonable to have a higher premium. Smoking also exposes one to many diseases, which also aligns with the result of higher premiums.

Conclusions

To conclude, a reasonable model and analysis of insurance data has been completed. Through the data we determined which predictors from BMI, age, sex, region, number of children, and smoking lead to what effects on the cost of insurance. We have also explored the trend of increase for BMI and age, which we have found that BMI illustrates a clear discontinuity where once those who have surpassed a BMI of 30 have a jump in insurance cost. This is reasonable under the fact that a BMI of 30 and over is classified as obese [2]. These results can help educate individuals on the magnitude of how their lifestyle may affect their insurance premiums. This study is certainly not without its limitations, where previous medical history, which is a large determining factor for most insurance companies is not included in the dataset [7]. Next steps for the future would be to include a more extensive medical history, as that may have more of an effect than any of the predictors in this study.

Bibliography

1. Kaur, T. (2018). "Factors affecting health insurance premiums: Explorative and predictive analysis" Creative Components. 72. https://lib.dr.iastate.edu/creativecomponents/72
2. U.S. Department of Health and Human Services. (n.d.). Calculate Your BMI - Standard BMI Calculator. National Heart Lung and Blood Institute. https://www.nhlbi.nih.gov/health/educational/lose_wt/BMI/bmicalc.htm.
3. Lantz, B. (2015). Machine learning with R: discover how to build machine learning algorithms, prepare data, and dig deep into data prediction techniques with R. Packt Publishing.
4. Hao Zhu. (2020). kableExtra: Construct Complex Table with 'kable' and Pipe Syntax. http://haozhu233.github.io/kableExtra/, https://github.com/haozhu233/kableExtra.
5. R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
6. David Robinson, Alex Hayes and Simon Couch. (2021). broom: Convert Statistical Objects into Tidy Tibbles. https://broom.tidymodels.org/, https://github.com/tidymodels/broom.
7. Huddleston, C. (2021, May 20). How Life Insurance Companies Get Intel On You. Forbes. https://www.forbes.com/advisor/life-insurance/personal-data/.

Appendix

Preview of the data

Table 6: Preview of data structure

Age	Sex	BMI	Children	Smoker	Region	Cost
19	female	28	0	yes	southwest	16885
18	male	34	1	no	southeast	1726
28	male	33	3	no	southeast	4449
33	male	23	0	no	northwest	21984
32	male	29	0	no	northwest	3867
31	female	26	0	no	southeast	3757