HEALTH INSURANCE PREMIUM PREDICTION

Authors:

Bhagyashri Raghunath Gadkari (BRG210002)

Huda Abdul Rawoof (HXR190026)

Jawadul Fathima (JXF200023)

Prathamesh Ajay Pawar (PAP210000)

Priyanka Pimpalekar (PXP210021)

Rohan Satish Patil (RSP210001)

Sangram Tushar Mohite Patil (SXM210110)

Saurabh Dhananjay Jadhav (SDJ200000)

Professor: Quanquan Liu

The University of Texas at Dallas

1. Introduction

The right health insurance plan is essential in today's rapidly growing economy, rapidly deteriorating health, and exponentially increasing medical costs. Health Insurance is a type of insurance that covers the medical expenses of the benefactor. To benefit from a health insurance policy, the benefactor must pay a particular premium annually. As an insurance provider and benefactor, it helps to be able to predict the insurance costs(premium) to be paid and what it maybe depends on.

Our project aims to estimate the effects of the benefactor's lifestyle behavior factors like smoking habits (smoker and non-smoker), age, BMI, etc. on individual medical costs billed by health insurance companies annually. The amount of the premium (insurance charges) for a health insurance policy depends on person to person, as many factors affect the insurance charges. For example, age, where a young person is less likely to have major health problems compared to an older person. Thus, treating an older person will be more expensive compared to a young one which is why an older person is required to pay a higher premium compared to a young person. Just like age, many other factors affect the premium for a health insurance policy. Hence, our project helps determine the statistically significant factors for premium prediction.

2. Literature Review

Some of the recent literature that describes the various mechanisms of estimating the costs of physical healthcare is summarized below.

[1] reviewed past studies that used statistical methods for insurance costs analysis and claims prediction. Methods such as logit analysis, Tobit, adjusted Tobit, GLM(gamma/log gamma), Cox proportional hazard model (semi-parametric), weighted regression, mixed model, nonparametric additive regression, OLS log, threshold logit models, Weibull (parametric) and recursive partitioning algorithms have been implemented and compared.

In [2], it is observed that unplanned 30-day readmissions are a common occurrence among congestive heart failure (CHF) patients, posing major health concerns and increasing healthcare costs. There have been various studies conducted to recognize high-hazard individuals.

Because of the aging populations and enhanced therapy of fundamental conditions, cardiac arrest is among the most complicated chronic disorders with a higher incidence. The incidence is projected to climb gradually, reaching 3% of the population in Western countries [3]. It is the leading reason for hospitalizations in people aged 65 and above, leading to substantial expenses and a significant societal effect.

Powers et al. [4] evaluated several regression statistical modeling approaches for predicting prospective total annual health costs (medical plus pharmacy) of health plan participants using Pharmacy Health Dimensions (PHD), a pharmacy claims-based risk index. Their models included ordinary least squares (OLS) regression, log-transformed OLS regression with a smearing estimator, and 3 two-part models using OLS regression, log-OLS regression with a smearing estimator, and generalized linear modeling (GLM), respectively. The results showed that most Ph.D. drug categories could solely predict future total health costs.

3.1 Data Description:

Our data is sourced from Kaggle. It has one table – insurance.csv which consists of 1338 observations and 7 columns.

Variable Name	Description
age	age of primary beneficiary
sex	sex of primary beneficiary
BMI	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
children	Number of children covered by health insurance / Number of dependents
smoker	Whether or not the person is a smoker
region	the beneficiary's residential area in the US: northeast, southeast, southwest, northwest
charges	Individual medical costs billed by health insurance

3.2 Data Preprocessing:

3.2.1 Elimination of missing records, duplicate records, and outliers

Checking the availability of missing records, we found that there was none in our data. We found one duplicate record along with a few outliers which were eliminated from our data.

3.2.2 Conversion of categorical variables to factors

We have 3 nominal categorical variables in our data – sex, smoker, and region. These categorical variables are converted into indicators.

Sex takes up the value ‘male’ and ‘female’ which upon conversion into factors indicates 0 when it’s a female or 1 when it’s a male.

Smoker takes up the value ‘no’ and ‘yes’ that indicates whether an individual is a non-smoker or smoker. On conversion, it indicates 0 for a non-smoker and 1 for a smoker.

Region comprises of northeast, northwest, southeast, southwest which upon conversion assume values of 1 to 4 to indicate each region.

```
. tabulate smoker , generate(dsmoker)
```

smoker	Freq.	Percent	Cum.
no	1,064	79.52	79.52
yes	274	20.48	100.00
Total	1,338	100.00	

```
. tabulate sex, generate(dsex)
```

sex	Freq.	Percent	Cum.
female	662	49.48	49.48
male	676	50.52	100.00
Total	1,338	100.00	

```
. tabulate region, generate(dregion)
```

region	Freq.	Percent	Cum.
northeast	324	24.22	24.22
northwest	325	24.29	48.51
southeast	364	27.20	75.71
southwest	325	24.29	100.00
Total	1,338	100.00	

```
. tabulate children , generate(dchildren)
```

children	Freq.	Percent	Cum.
0	574	42.90	42.90
1	324	24.22	67.12
2	240	17.94	85.05
3	157	11.73	96.79
4	25	1.87	98.65
5	18	1.35	100.00
Total	1,338	100.00	

```
tabulate dsex
```

dsex	Freq.	Percent	Cum.
0	662	49.48	49.48
1	676	50.52	100.00
Total	1,338	100.00	

```
tabulate dsmoker
```

dsmoker	Freq.	Percent	Cum.
0	1,064	79.52	79.52
1	274	20.48	100.00
Total	1,338	100.00	

```
tabulate dregion
```

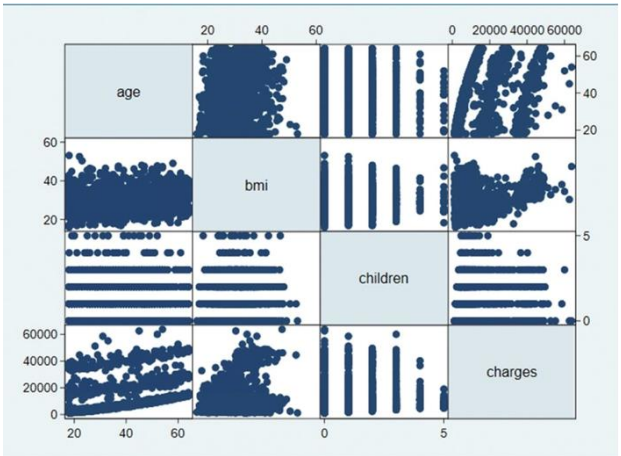
dregion	Freq.	Percent	Cum.
1	324	24.22	24.22
2	325	24.29	48.51
3	364	27.20	75.71
4	325	24.29	100.00
Total	1,338	100.00	

On observing our data, we see that:

- There are a higher number of males and non-smokers.
- Most of the population belongs to the southeast region.
- The average age of the beneficiaries is about 39 years.
- Most of the people have no children.

3.3 Exploratory Data Analysis:

3.3.1 Graph Matrix of numerical variables



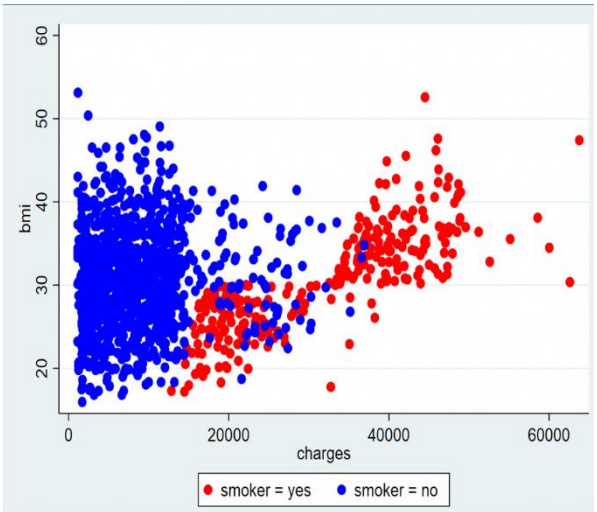
From the above graph, we can see a relationship between ages and charges along with bmi and charges. So, we will be focusing on these specific variables. Now, we need to decipher whether any of the categorical variables have a relationship with charges.

3.3.2 Box Plot of Smoker-Charges

From the above plot, we conclude that smokers are charged with higher premium rates of insurance compared to nonsmokers. This makes sense as smokers are more likely to face major health problems in the future than nonsmokers.

Let us observe the relationship between bmi, smoker and charges.

3.3.3 BMI vs Charges w.r.t Smoker



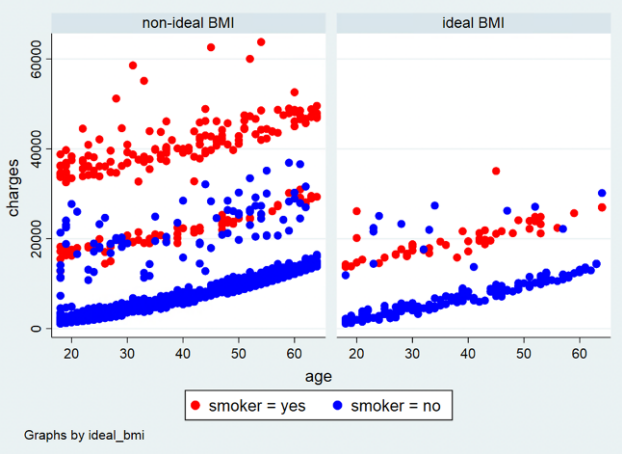
tabulate ideal_bmi

ideal_bmi	Freq.	Percent	Cum.
0	1,117	83.48	83.48
1	221	16.52	100.00
Total	1,338	100.00	

From the above plot, we again see how smokers are charged the highest premium rates, but we also notice that people non ideal BMI are also charged higher premium rates.

To look further into this, we generate a variable 'ideal_bmi' which indicates BMI from 18.5-24.9. We note that only about 16.5% of the population has an ideal bmi.

3.3.4 Distribution of ages over ideal and non-ideal BMI w.r.t smoker



The above graph shows the distribution of ages along with their charges over ideal and non-ideal BMIs w.r.t smoker. We can conclude that most smokers with non-ideal BMIs are charged the highest rates.

3.4 Correlation between the variables:

```
. corr charges bmi age dsex children dsmoker dregion
(obs=1,338)
```

	charges	bmi	age	dsex	children	dsmoker	dregion
charges	1.0000						
bmi	0.1983	1.0000					
age	0.2990	0.1093	1.0000				
dsex	-0.0573	-0.0464	0.0209	1.0000			
children	0.0680	0.0128	0.0425	-0.0172	1.0000		
dsmoker	0.7873	0.0038	-0.0250	-0.0762	0.0077	1.0000	
dregion	-0.0062	0.1576	0.0021	-0.0046	0.0166	-0.0022	1.0000

Smoker, age, and BMI are highly correlated with charges compared to the other variables.

4. Empirical Method

Empirical analysis is the statistical approach used to find the answer to real-world complex problems. One of the key factors that the Empirical method considers is ‘research question’.

Data Modeling

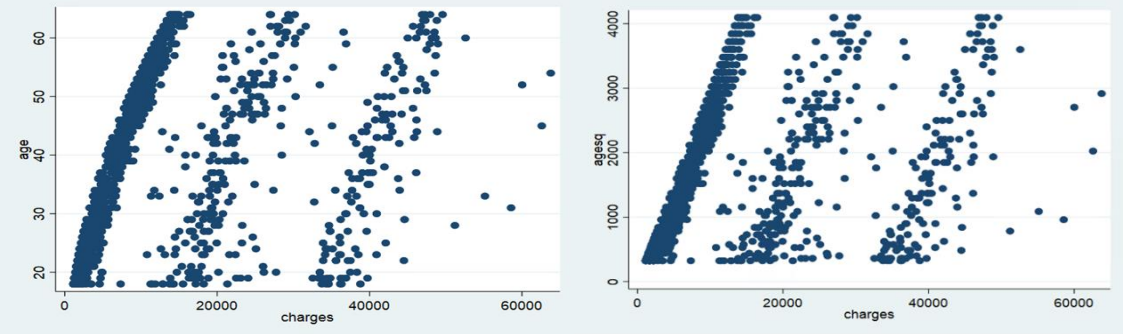
To predict the healthcare charges, the regression technique is considered, that can be defined as the model, which is a linear function that defines relationship between a dependent variable and an independent variable with an error term.

Models used for predicting:

1. Polynomial Regression with an interaction term
2. Multiple Linear Regression
3. Simple Linear Regression

1. Polynomial Regression with an interaction term

It is a type of linear regression wherein the relationship between independent and dependent variables is an n-th degree polynomial.



(a) (b)

From the above chart (a) we can see that there is a slight curve which indicates that age and charges have a non-linear relationship. To make the relationship linear we transform the age variable to age square which can be seen in chart (b).

Since age and charges have a non-linear relationship, we selected polynomial regression as the third model. From EDA we have seen that if a person is a smoker, then with an increase in BMI the charges also increase. To account for this interaction, we used an interaction term bmi_smoker.

$Charges=\beta0+\beta1age+\beta2age2+\beta3sex+\beta4bmi+\beta5smoker+\beta6smoker*bmi+u$

```
. reg charges age agesq dsex bmi dsmoker dsmokerbmi
```

Source	SS	df	MS	Number of obs	=	1,338
				F(6, 1331)	=	1145.41
Model	1.6426e+11	6	2.7377e+10	Prob > F	=	0.0000
Residual	3.1813e+10	1,331	23901384.9	R-squared	=	0.8378
				Adj R-squared	=	0.8370
Total	1.9607e+11	1,337	146652372	Root MSE	=	4888.9

charges	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	80.75709	62.62343	1.29	0.197	-42.0943	203.6085
agesq	2.345004	.7818462	3.00	0.003	.8112188	3.878789
dsex	473.1501	268.8065	1.76	0.079	-54.18047	1000.481
bmi	6.2722	24.97508	0.25	0.802	-42.72262	55.26702
dsmoker	23836.81	332.2756	71.74	0.000	23184.97	24488.65
dsmokerbmi	1433.737	53.08696	27.01	0.000	1329.593	1537.88
_cons	715.8651	1374.629	0.52	0.603	-1980.81	3412.54

It is evident from the results that the age^2, sex, smoker, smoker*bmi variables are statistically significant. At the average bmi, smoker has a statistically significant positive effect on charges.

R-squared= 0.8378, which means that the model explains around 84% of the variability in the data.

2. **Multiple Linear Regression:**

Multiple linear regression is a statistical technique that uses several explanatory variables to predict the outcome of a response variable.

$Charges=\beta0+\beta1age+\beta2sex+\beta3bmi+\beta4children+\beta5smoker+\beta6region+u$

```
. reg charges age dsex bmi children dsmoker dregion
```

Source	SS	df	MS	Number of obs	=	1,338
				F(6, 1331)	=	668.12
Model	1.4720e+11	6	2.4533e+10	Prob > F	=	0.0000
Residual	4.8874e+10	1,331	36719766.5	R-squared	=	0.7507
				Adj R-squared	=	0.7496
Total	1.9607e+11	1,337	146652372	Root MSE	=	6059.7

charges	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
age	257.2881	11.88574	21.65	0.000	233.9712	280.6049
dsex	131.1106	332.8107	0.39	0.694	-521.7801	784.0013
bmi	332.5701	27.72217	12.00	0.000	278.1862	386.954
children	479.3694	137.6442	3.48	0.001	209.3461	749.3926
dsmoker	23820.43	411.8429	57.84	0.000	23012.5	24628.37
dregion	-353.64	151.9266	-2.33	0.020	-651.6817	-55.59836
_cons	-11592.92	994.9299	-11.65	0.000	-13544.72	-9641.121

It is evident from the results that the age, bmi, children, smoker, region variables are statistically significant.

The R-squared value of the model is 0.7507, which tells that model explains around 75% of the total variability in the observed data.

3. **Simple Linear Regression:**

Simple Linear Regression is a model that describes the relationship between one dependent variable and one independent variable.

From the Exploratory Data Analysis, it was visible that the ‘smoker’ variable had a significant impact on the dependent variable ‘charges.

The first model considered ‘charges’ as the dependent variable and ‘smoker’ as the independent variable.

charges = $\beta 0$ + $\beta 1$*smoker+ u

```
. reg charges dsmoker
```

Source	SS	df	MS	Number of obs	=	1,338
				F(1, 1336)	=	2177.61
Model	1.2152e+11	1	1.2152e+11	Prob > F	=	0.0000
Residual	7.4554e+10	1,336	55804130.6	R-squared	=	0.6198
				Adj R-squared	=	0.6195
Total	1.9607e+11	1,337	146652372	Root MSE	=	7470.2

charges	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
dsmoker	23615.96	506.0753	46.66	0.000	22623.17	24608.75
_cons	8434.268	229.0142	36.83	0.000	7985.002	8883.535

It is evident from the results that the ‘smoker’ variable is statistically significant.

The R-squared value of the model is 0.6198, which tells that model explains around 62% of the total variability in the observed data.

5. Results

Models	R-Squared	RMSE
Simple Linear Regression	62%	7470.2
Multiple Linear Regression	75%	6059.7
Polynomial Regression with interaction term	84%	4888.9

In our research, from the factors chosen, age, BMI, and smoker were found to be most impactful. Further, it can be inferred that smoking plays a major role in predicting insurance premium costs.

Even if the 'sex' variable is not statistically significant in any of the models, as a belief, it plays a crucial role in calculating insurance costs. Hence, we have taken the variable into consideration.

6. Conclusion

As per our findings, Polynomial Regression with an interaction term explains 84% of the data and a strong correlation between actual and predicted values. Hence, it is the best model among the chosen ones.

In the future, we would like to consider factors like medical history and various disease conditions to get more accurate predictions and improve the distribution of residuals.

7. References

1. https://doi.org/10.1016/j.jbi.2019.103256
2. https://doi.org/10.1007/s13167-019-00188-9
3. https://pubmed.ncbi.nlm.nih.gov/16224298
4. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8906954/
5. Health Insurance Premium Prediction with Machine Learning. Available online: https://thecleverprogrammer.com/2021/10/26 /health-insurance-premium-prediction-with-machine-learning/
6. https://www.hindawi.com/journals/mpe/2021/1162553/
7. ul Hassan, C.A.; Iqbal, J.; Hussain, S.; AlSalman, H.; Mosleh, M.A.A.; Sajid Ullah, S. A Computational Intelligence Approach for Predicting Medical Insurance Cost. Math. Probl. Eng. 2021, 2021, 1162553. [Reference paper]
8. Cevolini, A.; Esposito, E. From Pool to Profile: Social Consequences of Algorithmic Prediction in Insurance. Big Data Soc. 2020, 7. [Reference]