Health Insurance, is it worth it?

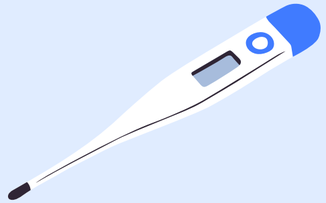
AN ANALYSIS OF MEDICAL INSURANCE CLAIMS

LT 7: Acot | Del Rosario | Jayme | Ruiz | Timajo | Velante

OUTLINE

Introduction

Exploratory Data Analysis



Hypothesis Testing

Multiple Linear Regression

Summary & Recommendation

THE HEALTH INSURANCE INDUSTRY

Introduction

Problem Statement

“What are the factors that affect the value of an insurance claim?”

Proponents

Helath Insurance Industry and its patients

Business Value & Objective

Optimize insurance premiums from the insurance companies based on factors correlated with high insurance claims

DATA DESCRIPTION

Data Source

Medial cost personal
dataset from kaggle.com
1338 entries
6 features

*HTTPS://WWW.KAGGLE.COM/MIRICHOI0218/INSURANCE

Dependent Variable

charges - individual medical costs billed by health insurance

Independent Variables

age - age
sex - male or female
bmi - body mass index
children - number of dependents
smoker - smoker or non-smoker
region - area of residence

HYPOTHESIS TESTING

Chi-Square Test of Independence

Mann Whitney U Test

Kruskal-Wallis Test

Friedman's Test

ASSUMPTIONS OF TESTS USED

Chi-Square Test of Independence

Mann Whitney U Test

Kruskal-Wallis Test

Friedman's Test

OBSERVATIONS: INDEPENDENT

DISTRBUTION: NOT NECESSARILLY NORMALLY DISTRIBUTED/ NON-PARAMETRIC

SAMPLING: RANDOM

α 0.05

CHI SQUARE TEST

Gender & Smoking

H0: There is no association between gender and smoking habits

P-VALUE IS 0.0065

REJECT NULL HYPOTHESIS

α 0.05

MANN-WHITNEY U TEST

	Charges of Smokers vs Charges of Non-smokers	Charges of lower BMI vs Charges of higher BMI	Charges of younger vs Charges of older
			
H0	There is no significant difference in the charges between smokers and non-smokers.	There is no significant difference in the charges between higher BMI and lower BMI	There is no significant difference in the charges between younger and older individuals.
P-VALUE	**1.77E-78**	**6.51E-53**	**4.92E-63**
	REJECT NULL HYPOTHESIS	REJECT NULL HYPOTHESIS	REJECT NULL HYPOTHESIS

*Note: The group used Mann Whitney and Kruskal-Wallis Test as the variances across groups were revealed to be non-homogenous based on Bartlett's test

α 0.05

KRUSKAL-WALLIS TEST

Region

There is no significant difference in the charges across regions

P-VALUE IS 0.00

REJECT NULL HYPOTHESIS

*Note: The group used Mann Whitney and Kruskal-Wallis Test as the variances across groups were revealed to be non-homogenous based on Bartlett's test

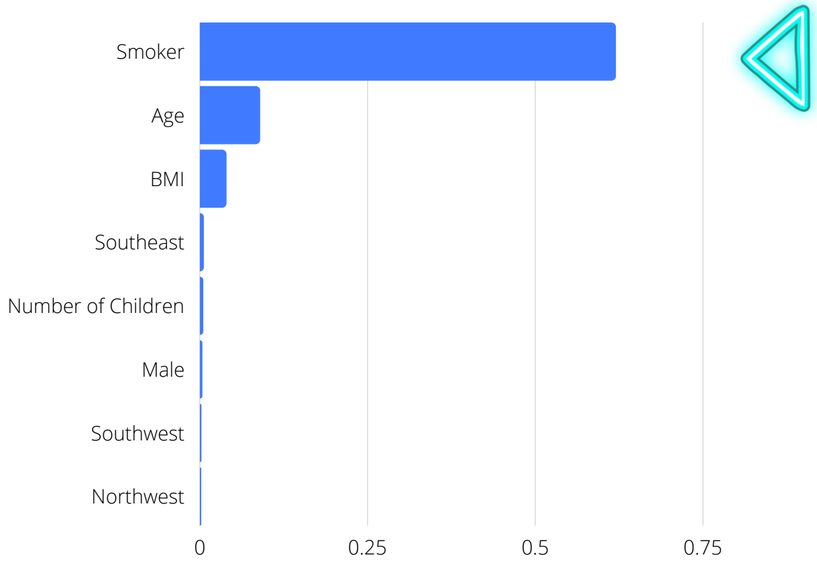
α 0.05

FRIEDMAN'S TEST

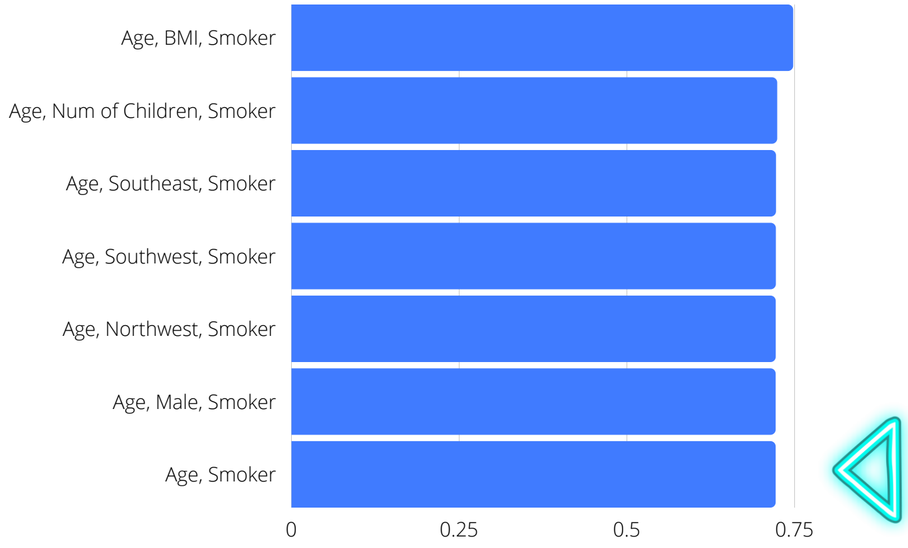
	Age as covariate	Children as covariate	BMI as coviariate
			
H0	There is no significant difference in the charges across regions with age as a subject	There is no significant difference in the charges across regions with number of children as a subject.	There is no significant difference in the charges across regions with BMI as a subject.
P-VALUE	**0.24**	**0.15**	**0.19**
	ACCEPT NULL HYPOTHESIS	ACCEPT NULL HYPOTHESIS	ACCEPT NULL HYPOTHESIS

Multiple Linear Regression Modelling

The linear regression model clearly shows that the smoker_yes variable has the largest impact on the charges dependent variable

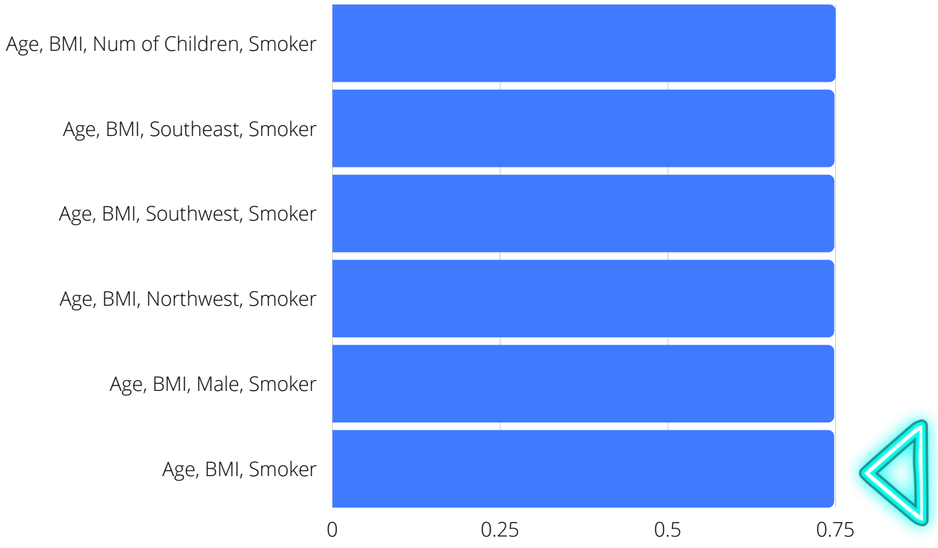


The age variable has the greatest impact on R-Squared, increasing from 62% to 72.1%.

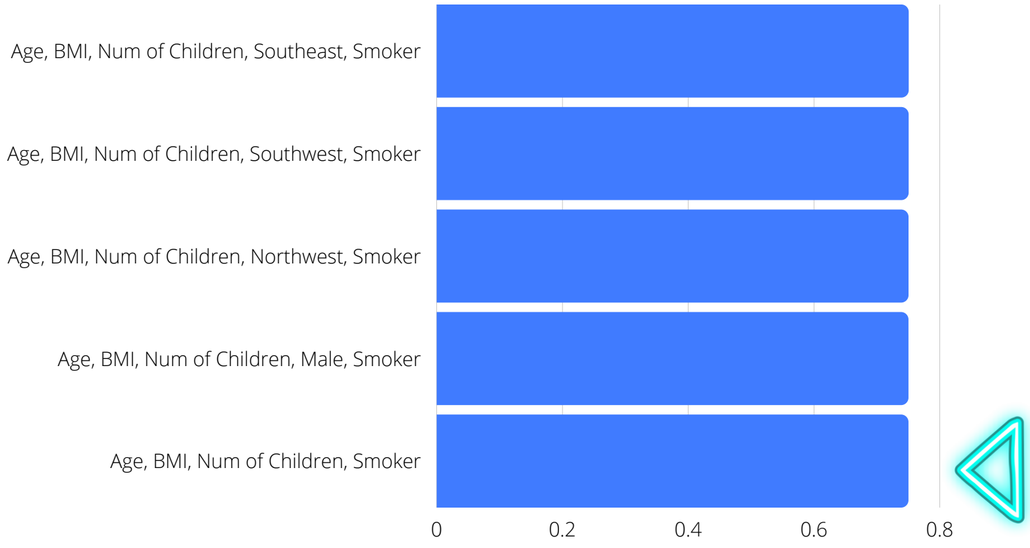


Multiple Linear Regression Modelling

Combining smoker_yes, age, and bmi increased R-Squared to 74.75%.



We therefore keep only the saved parameters smoker_yes, age, bmi and num_children



TEST OF NORMALITY

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	998.1
Date:	Wed, 15 Sep 2021	Prob (F-statistic):	0.00
Time:	21:17:32	Log-Likelihood:	-13551.
No. Observations:	1338	AIC:	2.711e+04
Df Residuals:	1333	BIC:	2.714e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.21e+04	941.984	-12.848	0.000	-1.4e+04	-1.03e+04
bmi	321.8514	27.378	11.756	0.000	268.143	375.559
smoker_yes	2.381e+04	411.220	57.904	0.000	2.3e+04	2.46e+04
num_children	473.5023	137.792	3.436	0.001	203.190	743.814
age	257.8495	11.896	21.675	0.000	234.512	281.187

Omnibus:	301.480	Durbin-Watson:	2.087
Prob(Omnibus):	0.000	Jarque-Bera (JB):	722.157
Skew:	1.215	Prob(JB):	1.53e-157
Kurtosis:	5.654	Cond. No.	292.

H0: The data is normally distributed.

Jarque-Bera test — SAMPLE SIZE BELOW MINIMUM

Shapiro-Wilk test — REJECT NULL HYPOTHESIS

Omnibus K-squared test — REJECT NULL HYPOTHESIS

RESULTS: MULTIPLE LINEAR REGRESSION MODELLING

OLS Regression Results

Dep. Variable:	charges	R-squared:	0.750
Model:	OLS	Adj. R-squared:	0.749
Method:	Least Squares	F-statistic:	998.1
Date:	Wed, 15 Sep 2021	Prob (F-statistic):	0.00
Time:	21:17:32	Log-Likelihood:	-13551.
No. Observations:	1338	AIC:	2.711e+04
Df Residuals:	1333	BIC:	2.714e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	-1.21e+04	941.984	-12.848	0.000	-1.4e+04	-1.03e+04
bmi	321.8514	27.378	11.756	0.000	268.143	375.559
smoker_yes	2.381e+04	411.220	57.904	0.000	2.3e+04	2.46e+04
num_children	473.5023	137.792	3.436	0.001	203.190	743.814
age	257.8495	11.896	21.675	0.000	234.512	281.187

Omnibus:	301.480	Durbin-Watson:	2.087
Prob(Omnibus):	0.000	Jarque-Bera (JB):	722.157
Skew:	1.215	Prob(JB):	1.53e-157
Kurtosis:	5.654	Cond. No.	292.

With an R-Squared of 0.75, the regression model explains a moderate amount of the annual charges, holding all other variables constant:

1. An increase in age by 1 year increases annual charges by $258
2. An increase of BMI by 1 unit increases annual charges of $322
3. An increase of 1 child increases annual charges by $474
4. Being a smoker increases annual charges by $23,810

SUMMARY

Based on multiple hypothesis testing methods used, the group has determined that the top factors that affect the value of an insurance claim are age, bmi, being a smoker, and number of children.

The main hypothesis testing methods explored in this study are Chi Square test, Mann Whitney U test, Kruskal-Wallis, and Friedman's Test. Bartlett's Test is also used to check for homogeniety

RECOMMENDATION

1. Insurance companies can include the factors affecting insurance claims in their computation for insurance premiums
2. Insurance companies can explore machine learning models as a way to predict the value of the claims
3. The study can be improved by including details such as comorbidities, income, family medical history, among other factors that could result to a more accurate forecast