Final Project

Insurance Charges Modeling

Aline Rivera Mata
Pedro José Trillos Toro

Introduction

Insurance companies use price discrimination to protect themselves of the risk that implies giving insurance to someone. They charge higher rates to people who are more likely to claim the insurance and lower rates who are less likely to claim it.

Problem Statement: There is a lack of transparency in the insurance industry. Because of this, people are oftentimes unaware of what makes them pay more or less for insurance.

Study Significance: This study will be able to provide people an analysis of factors that might be making them pay more or less for their insurance. They can therefore identify if there is something they can change in their habits or if there are being treated unfairly by their insurance provider.

Limitations

- The gathered data belongs to insurance buyers solely in the United States. We cannot say that our findings are representative of the behavior of insurance charges in a global scale.

- There is a high likelihood that our model has omitted variables. Factors such as medical history, profession and hobbies may be some important ones.

Methodology

Variables

Dependent: charges

Independent:

1. age
2. sex (0=female)
3. bmi
4. children
5. smoker (0=non-smoker)
6. region (NE, NW, SE, SW)

Observations

1338

Type

Cross-Sectional

OLS linear regression using Microsoft Excel

Findings

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.8665524
R Square	0.750913
Adjusted R Square	0.7494136
Standard Error	6062.1023
Observations	1338

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	8	1.47235E+11	18404336091	500.81074	0
Residual	1329	48839532844	36749084.16		
Total	1337	1.96074E+11			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-11938.539	987.8191752	-12.08575302	5.579E-32	-13876.393	-10000.684	-13876.393	-10000.684
age	256.85635	11.89884907	21.58665523	7.783E-89	233.51378	280.19893	233.51378	280.19893
sex	-131.31436	332.9454391	-0.394402037	0.6933475	-784.47027	521.84155	-784.47027	521.84155
bmi	339.19345	28.59947048	11.86013055	6.498E-31	283.08843	395.29848	283.08843	395.29848
children	475.50055	137.8040925	3.450554599	0.000577	205.16329	745.8378	205.16329	745.8378
smoker	23848.535	413.1533548	57.72320196	0	23038.031	24659.038	23038.031	24659.038
southeast	-1035.022	478.6922095	-2.162186952	0.0307817	-1974.0968	-95.947326	-1974.0968	-95.947326
southwest	-960.05099	477.9330243	-2.008756337	0.0447649	-1897.6364	-22.4656	-1897.6364	-22.4656
northwest	-352.9639	476.2757859	-0.741091422	0.4587689	-1287.2982	581.3704	-1287.2982	581.3704

R-squared: 75%
Adjusted R-squared: 74.9%
F significance: 0

Model:

charges = -11938.54 + 256.86(age) + 339.19(bmi) + 475.50(children) + 23848.54(smoker) - 1035.02(southeast) - 960.05(southwest) + e

Non-significant variables:
sex and northwest

Findings

According to the independent variables' coefficients, ***ceteris paribus*:**

- With the increase of 1 year on the age of a client, their medical insurance costs rise by 256.86 USD.
- With the increase of 1 kg/m2 of a client, their medical insurance costs rise by 339.19 USD.
- With the increase of one more dependent of a client, their insurance costs rise by 475.50 USD.
- If a client smokes, their medical insurance costs rise 23848.54 USD.
- If a client lives in the USA's southeast region, their medical insurance costs decrease by 1035.02 USD.
- If a client lives in the USA's southwest region, their medical insurance costs decrease by 960.05 USD.

Conclusion

- The low p-values of the variables age, bmi, children, and smoker are consistent with the theory.
  - The older a person, the more likely it is that they get an illness.
  - People with higher weights are more propense to cardiovascular diseases.
  - The more dependents you have in your insurance plan, the higher the costs will be.
  - A person that smokes is more likely to have respiratory diseases.

- The higher p-values of the region variables and sex are also consistent with the theory.
  - Lifestyles and economic conditions across the USA's four regions are relatively homogeneous; therefore, it would not be logical for clients to be charged more or less for medical insurance according to where they reside.
  - Insurance costs should not vary much according to your gender.

Conclusion

- The study recommends the reader that in order to avoid higher medical insurance costs they should not smoke, try to be in a healthy weight (as there is nothing to be done in terms of height), subscribe less dependents to their insurance plan, and try to buy their insurance at a young age rather than when they’re older.

- It is recommended that future researchers explore what happens with the Adjusted R-squared if the region variables are dropped and omitted variables are included in the model.