Info 511 Final Project

By: James Christensen

Is there a practical way for individuals to reduce how much their insurance company spends on them, and thereby reduce how much the individual spends on insurance?

Medical Insurance Cost Prediction Dataset

- Originates from Kaggle. Author is Mohan Krishna Thalla
- Dataset has information on 100,000 individuals such as demographics, health history, medical insurance expenditures, and lifestyle
- Data covers the globe and the year range of 2014-2024
- “The dataset was compiled from publicly available health surveys, insurance research studies, and anonymized online healthcare data” (Thalla)

Exploratory Data Analysis

Missing Value Analysis

Variable Name	P-value
Total_claims_paid (Y)	0.569 (t-test)
BMI	0.893 (t-test)
Income	0.734 (t-test)
Alcohol_freq	0.470 (Chi-sq test)
Employment_status	0.749 (Chi-sq test)
Urban_rural	0.079 (Chi-sq test)

Model, Variables, and Description

$$Y_{totalClaimsPaid} = \beta_0 + \beta_1 X_{income} + \beta_2 X_{bmi} + \beta_3 X_{smokerFormer} + \beta_4 X_{smokerCurrent} + \beta_5 X_{alcoholWeekly} + \beta_6 X_{alcoholDaily} + \beta_7 X_{suburban} + \beta_8 X_{rural} + \beta_9 X_{unemployed} + \beta_{10} X_{retired} + \epsilon_i$$

Variable Name	Description
Income	Annual Income (USD)
BMI	Body Mass Index
Smoker_Former	Used to smoke
Smoker_Current	Currently smokes
Alcohol_Weekly	Drinks weekly
Alcohol_Daily	Drinks daily
suburban	Lives in a suburban area
Rural	Lives in a rural area
Unemployed	Doesn’t have a job currently
Retired	No longer working

Models and methods

1. Linear model - No transformation
    a. Violated assumptions of normality and linearity
2. Linear model - Log transformation on total_claims_paid (Y)
    a. Violated assumption of normality, linearity, and homoscedasticity
    b. *Strange* behavior obvious from qq-plot and fitted vs. residual plot
        i. Two clear groups formed! One for no claims and one for at least one claim paid
3. Logistic regression - Model whether insurance company spent at least $1
    a. Pseudo R^2 of 0.00009
    b. *Not* one predictor statistically significant
4. Final linear model - Log transformation on (Y) and only on individuals where at least $1 was spent
    a. All linear regression assumptions met
    b. Low predictive power

Visual

Conclusion

- At an R^2 value of 0.029, our model doesn't explain enough variation to responsibly predict medical insurance claims paid
- However, we still have several variables that have coefficient estimates, which are significant.
- Bonferroni adjusted 95% confidence intervals

Variable Name	Adjusted Confidence Interval
BMI	($1.008, $1.013)
Current_smoker	($1.521, $1.626)
Former_smoker	($1.042, $1.089)

Possible follow-ups

- Poisson regression on claims made in the year
- Linear regression on total medical costs (insurance + individual)
- Logistic regression including lifestyle choices AND prior medical history

References

- Thalla, Mohan. (2025, September). Medical Insurance Cost Prediction, Version 1. Retrieved October 21, 2025 from https://www.kaggle.com/datasets/mohankrishnathalla/medical-insurance-cost-prediction