Milestone 3 (Report Rough Draft)

By: James Christensen

Introduction

Health is the prerequisite for all human endeavors. Being healthy is something most of us take for granted, yet we are all internally aware that it is constantly at risk. As of 2023, 92% of Americans had health insurance (public or private) (Keisler-Starkey, K. & Bunch, L. N.). This insurance is a tremendous financial burden for most. In fact, the average health insurance premium for family coverage in 2024 was $25,572. An increase of 24% since 2019 (KFF). These costs sum to an insurance market with a volume of $318.4 billion dollars a year (Health Insurance - worldwide: Statista market forecast).

Since health insurance is both a necessity and expensive, it then begs the question, *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?*

Methods

To answer this question, we analyzed the Medical Insurance Cost Prediction dataset from Kaggle. This dataset includes information about 100,000 individuals including their demographics, socioeconomic status, health conditions, lifestyle factors, insurance plans, and medical expenditures. Specifically, we would like to analyze how lifestyle and habits impact medical expenditures.

The dataset has a CC0: Public Domain License and was "compiled from publicly available health surveys, insurance research studies, and anonymized online healthcare data" (Thalla). It covers individuals around the globe from the years 2014 through 2024.

To answer the question, we selected the total_claims_paid (Y) column. As our predictor variables we selected bmi, smoker, alcohol_freq, urban_rural, employment_status, and income. The only variable which had missing values was alcohol frequency, 30083 missing values to be exact. Since there are 100,000 values in the dataset, a possible option was to remove all of the missing value rows as there would still be substantial power in the model.

In order to see if this was proper, we conducted two sample t-tests for difference in means between the values that were missing and the remaining data for numerical variables, and Chi-Sq tests for difference in proportions for the categorical variables. *Fig 1.* displays that no factor in the model is entangled with alcohol_freq at an alpha level of 0.05. In other words, the values of alcohol_freq are missing likely due to some of the original data sources not containing this variable or completely at random. It thus seemed appropriate to remove these missing values.

After removing the missing values, each variable was individually analyzed. Total_claims_paid is very strongly left skewed *fig 2.* This indicated that a transformation may be necessary, however we waited to see what model assumptions were violated before proceeding with this measure.

The distribution for bmi matched what is found in others' research. Our distribution for bmi had a mean of 27 and a standard deviation of 5, which matches the research of (insert source here).

The smoker variable had a mode of "never" smokers. This level was deemed the baseline for the model and dummy variables for current and former smokers were created. Similarly, the mode for alcohol_freq was "occasional" drinkers, which was assigned as the baseline with dummy variables for weekly and daily. The same procedure was conducted with the urban_rural variable. Urban_rural had a mode of "urban", and dummy variables were created for suburban and rural. The employment status variable was conducted differently. The employment_status variable had levels of employed, retired, unemployed, and self-employed. It seemed fitting to combine the levels of employed and self-employed as they will likely provide feelings of fulfillment that come from working (insert source for claim here). This combined level simply became "employed". It is the baseline with dummy variables unemployed and retired.

The income variable is not normally distributed. It has a mean of $49907 and a standard deviation of $46959. From (insert source here) we know that income follows the pareto distribution, so its distribution is not a surprise *fig 4.* Before performing regularization on this variable, we waited to see if the model's assumptions were violated and remediated with transformations on Y first.

Multicolinearity was checked for between bmi and income. The two variables have a correlation of 0, so there is no multicolinearity in the model.

Model Selection

The first model that was fit was a basic linear regression model with no transformations on Y.

$$Y_{totalClaimsPaid}(\text{Finish equation later})$$

Due to the way the data was collected, it is possible that the assumption of independence is violated. It is possible that a few individuals in the dataset are duplicates from different time periods which would affect the data. However, the probability that this happened even once, or more than once, is so low, that for all intents and purposes, it is appropriate to conclude that the assumption of independence is met as each row represents a difference individual. This model did violate the assumptions of normally distributed errors, homoscedastic errors, (*fig 5.)* and a linear relationship between the variables (*fig 6.)*.

As such, box-cox was conducted and it found that the optimal lambda for Y is 0.067. Since this is very close to zero, a log transformation was applied to Y. A second model was fit with this log transformation on Y.

$$\log(Y_{totalClaimsPaid}(\text{Finish equation later})$$

This model had some very strange behavior, which was evidenced in both the QQ-plot (*fig 7.)* and the fitted vs. residual plot (*fig 8.)*. The same three assumptions were violated, but this

time two clear groups formed. Testing some data points, it became clear that the two groups were individuals who didn't have anything spent on them and those who did.

The natural remedy to this was to try and use logistic regression. Logistic regression would still answer our research question, but instead of quantifying savings for the insurance company, it would give us a probability of savings. Unfortunately, the pseudo R^2 for this model is 0.00009. There also wasn't one variable that was statistically significant, so this model was not helpful in answering our research question.

The last model we tried was deemed "the magnitude model." This model was only on the individuals who had a claim paid. It still answers the original research question, but in a different light. This model answers if a claim is made, how much does lifestyle choices affect how much is spent?

The model met all of the assumptions of linear regression as seen in Figs. 10 and 11. As such this was our final model.

Results

The resulting model still had fairly low predictive power. The model had a an R^2 value of 0.029. This means that the model accounts for 2.9% of the variation in total_claims_paid. The model still had a few statistically significant variables. To make inference on all three simultaneously, we adjusted the confidence intervals with the bonferroni adjustment. The three statistically significant variables are smoker_current, smoker_former, and bmi.

The 95% bonferroni adjusted confidence interval:

Smoker_current: ($1.521, $1.626)

Smoker_former: ($1.042, $1.089)

BMI: ($1.009, $1.013)

This means that we are 95% confident that the true coefficients for smoker_current, smoker_former, and bmi all lie upon their respective interval.

Future Research

Some future research that could be done is conducting poisson regression on the number of claims made in a year. Linear regression could be done on total medical costs (costs paid by both the individual and the insurance). Logistic regression on medical history, lifestyle choices, and other available variables in the data.

Conclusion

Our conclusion is that there while there is statistically significant evidence that the "lifestyle choices" of currently smoking, formerly smoking, and BMI increase the amount that health insurance spends on individuals increases with these choices, the coefficients are so small, that there is no practical savings that could be afforded to the individual subscriber.

References

- *Health Insurance - worldwide: Statista market forecast*. (n. d.). Statista. https://www.statista.com/outlook/fmo/insurances/non-life-insurances/health-insurance/worldwide?srsltid=AfmBOorCm_d8dYNvj6Lg4gyHFabogYVBwQh880-epbVoX38UPX8tfnW7
- Keisler-Starkey, K., & Bunch, L. N. (2024, September 11). *Health insurance coverage in the United States: 2023*. Census.gov. https://www.census.gov/library/publications/2024/demo/p60-284.html#:~:text=In%202023%2C%20most%20people%2C%2092.0%20percent%20or,for%20some%20or%20all%20of%20the%20year.
- Kff. (2025, October 14). *2024 employer health benefits survey*. KFF. https://www.kff.org/health-costs/2024-employer-health-benefits-survey/#e3efa8b3-48d2-458b-a2f7-c4d5add1983b--h-section-1-cost-of-health-insurance
- 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

Generative AI

ChatGPT was utilized during the coding portion of this project. It was consulted on how to efficiently perform different multilinear regression operations.

Appendix

All figures will be inserted later

Fig 1. Table displaying the p-value for each test between its rows associated with the missing alcohol frequency values and the not missing values

Fig 2. Boxplot of total_claims_paid

Fig 3. Distribution of BMI

Fig 4. Distribution of Income

Fig 5. QQ-plot of first model

Fig 6. Fitted-residual plot of first model

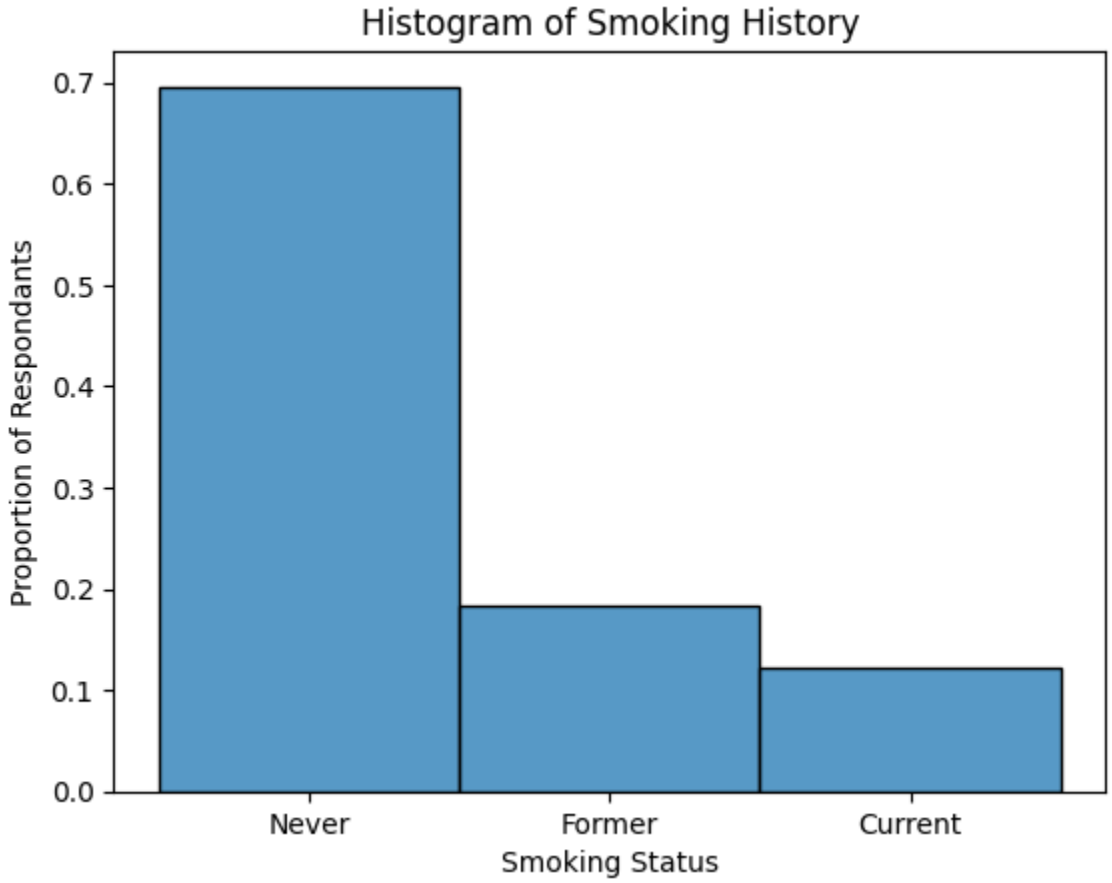
Fig 7. QQ-plot of the second model

Fig 8. Fitted-residual plot of the second model

Fig 9. Logistic Regression output

Fig 10. QQ-plot magnitude model

Fig 11. Fitted-residual plot Magnitude model



Bio-Sketch

James Christensen, M.S. Data Science University of Arizona, August 2025-December 2026, contributed to the project by writing the report, identifying the data, preprocessing the data, fitting the model, and validating the model.