**Statistical Question / Hypothesis**

**Final Project Write Up**

**Nilavan Duraisami**

Data Science, Bellevue University

DSC530-T301 2241: Data Exploration And Analysis

Professor Matthew Metzger

Nov 15, 2024

**Statistical Question / Hypothesis – Final Term Project – Write Up**

**Statistical/Hypothetical Question**

The central question explored in this analysis is whether “BMI is a significant predictor of medical expenses, regardless of smoking status.” This hypothesis aims to assess the impact of BMI on healthcare costs and investigate if its effect persists across different smoking statuses. Given the dataset's attributes, which include BMI, smoker status, age, and medical charges, the analysis explores whether BMI alone can account for variations in medical expenses.

**Outcome of Exploratory Data Analysis (EDA)**

The EDA results show that BMI has a statistically significant correlation with medical expenses, particularly for smokers, where the correlation is strong (0.81, p-value = 0.0000). For non-smokers, the correlation is weaker but still statistically significant (0.08, p-value = 0.0003). The regression analysis further supports the relationship between BMI and medical expenses, with the BMI coefficient suggesting that each unit increase in BMI is associated with an approximate increase of $396.20 in medical costs. However, the model’s R-squared value of 0.04 indicates that BMI alone explains only about 4% of the variability in medical expenses, suggesting other factors likely contribute to the costs. Additionally, the covariance analysis indicates that both smoker status and age influence medical expenses, especially for smokers, where the covariance with charges is notably high.

**What Was Missed During the Analysis?**

The dataset size was limited, and more comprehensive data could have improved the analysis. Additionally, the dataset lacked specific variables related to smoking habits, such as the length of the smoking habit or the number of cigarettes smoked per day. Including these variables could have provided deeper insights into the connection between smoking intensity and medical expenses.

**Additional Variables That Could Have Helped**

Variables such as the duration of smoking and the quantity of cigarettes smoked daily would have enriched the analysis, offering a more granular view of smoking’s impact on medical expenses. Socioeconomic factors, lifestyle choices, or health conditions (e.g., exercise frequency or pre-existing health conditions) could also have contributed to a more accurate prediction model for healthcare costs.

**Assumptions Made and Their Validity**

One assumption was that BMI alone could significantly predict medical expenses. However, the low R-squared value challenges this assumption, indicating that BMI, while statistically significant, is not a comprehensive predictor. Another assumption was that the dataset sufficiently represented the factors impacting medical costs, but missing variables like smoking intensity likely limited the analysis.

**Challenges Faced and Areas of Uncertainty**

A significant challenge was finding a more comprehensive dataset that could fully support the hypothesis. Additionally, understanding the limitations of a low R-squared value in interpreting BMI’s predictive power was initially complex, as it required recognizing that other, unmeasured variables play crucial roles in determining medical expenses.

**Conclusion**

In conclusion, the analysis reveals that BMI is a statistically significant predictor of medical expenses, particularly for smokers, who generally incur higher costs. However, BMI alone does not capture the full variability in medical expenses. Future analyses would benefit from a larger, more detailed dataset that includes additional variables, especially those related to smoking intensity and lifestyle factors. These insights underscore the importance of a multi-faceted approach to understanding healthcare costs in health insurance analytics.