Final Assignment

Group: Potato

GRA4157

1 Introduction

Health insurance premiums in the United States are influenced by various factors, including age, location, tobacco use, family size, and plan category. Notably, gender-specific differences in healthcare spending have been observed, with the average employd woman in the US having approximatly 18% more spending per year than a man, excluding pregnancy-related expenses[Edm23]. This disparity persists despite regulations mandating equal premium costs for men and women, suggesting that premiums alone do not fully capture the financial burden experienced by different genders. While existing research has explored many facets of health insurance, such as the increasing costs or the social issues related to decreasing coverage in the public[[DMC05], [Had07]], there is limited analysis on how specific variables, particularly gender, directly influence insurance charges. This report aims to fill this gap by utilizing the "Insurance Dataset for Predicting Health Insurance Premiums in the US" from Kaggle. The dataset comprises one million records with 12 variables that might have influence on insurance cost:

Variables	**Description**
Age	The age of the insured individual.
Gender	The gender of the insured individual.
BMI (Body Mass Index)	A measure of body fat based on height and weight.
Children	The number of children covered by the insurance plan.
Smoking Status	Indicates whether the individual is a smoker.
Region	The geographical region of the insured individual.
Medical History	Information about the individual's old medical problems.
Family Medical History	Information about the family's medical record.
Exercise Frequency	The frequency of the individual's exercise routine.
Occupation	The occupation of the insured individual.
Coverage Level	The type of insurance plan.
Charges	The health insurance charges for the individual.

Figure 1: Caption

By focusing on the 'charges' variable, this study will assess the impact of these factors, with a particular emphasis on gender, to understand potential disparities in insurance charging. Employing machine learning techniques, the analysis seeks to uncover patterns and relationships within the data, providing insights into the predictors of insurance charges. The goal is to enhance transparency in insurance cost allocation and inform strategies that address observed disparities, thereby contributing to a more equitable healthcare system.

2 Methodology

Our approach to uncover the underlying driving factors of medical insurance charges is to fit an appropriate model on the data. This statistical approach allowed us to quantify the relationships between the independent variables and the dependent variable, charges.Before training a model, we needed to pre-process our data. The data contains a mixture of quantitative and qualitative variables. For us to be able to draw information from our qualitative variables we used the Dummy coding method[VSMS16] to have the information represented numerically in our dataset. This method replaces categorical variables with

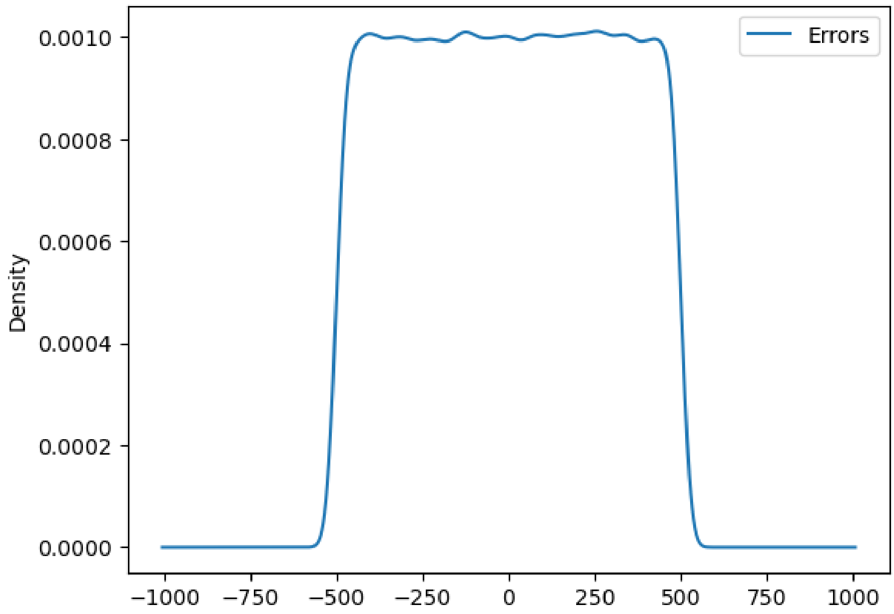

Figure 2: Distribution of the errors

binary variables. Continuing, split the data into training and test sets in order to run in-sample and out-of-sample on our models. With the training data we started off with an OLS model including all variables in the dataset. The multiple linear regression model takes the form:

$$\hat{y}_i = \hat{\alpha} + \hat{\beta}_1 age_i + \hat{\beta}_2 BMI_i + \hat{\beta}_3 1\{gender_i = "Male"\} + ... + \hat{\beta}_p 1\{smoker_i = "yes"\}$$

Running this regression resulted in an exceptional in-sample fit represented by an adjusted R^2 of 0,996. Additionally, according to the p-values of each estimate, all the variables seem to be significant. By making predictions on our test data and comparing these predictions to the real charges for the test data allows us to test the out-of-sample accuracy of our model as well. This resulted in a Mean Squared Error of 83 157, or 1,62% Mean Absolute Percentage Error. Our model exhibits exceptional out-of-sample prediction power as well, insinuating that the high R^2 is not due to overfitting. For further analysis of the model, we decided to look at the residuals of the model, which we calculate in the following way:

$$e = Y - X\hat{\beta}$$

Working with the OLS model, we have assumed that the residuals are normally distributed, but for our dataset it is uniformly distributed.

From the figure above we clearly see that the distribution of the errors resemble a uniform distribution. The uniform distribution is symmetric just as the normal distribution, thus the line that best fits the data will be the same. The problem however, is that the estimators are not t-distributed in this case because of the uniformly distributed residuals, thus their associated p-values are unreliable. With unreliable test statistics we can't easily perform feature selection. To circumvent this issue, we decide to run a Lasso regression to act as a sort of feature selection by pulling estimates deamed less important towards zero[Tib18]. The Lasso regression model modifies the traditional cost function by adding a penalty term proportional to the absolute sum of the coefficients:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \alpha \sum_{i=1}^{n} |\beta_i|$$

This penalty term encourages the coefficients of less significant variables to shrink towards zero, effectively removing them from the model. The Lasso regression model retains only the most important predictors, which helps simplify interpretation and improves generalization. To find the optimal hyper-parameter α for the Lasso model, we ran a 10-fold cross-validation on the model for α values in the range [0.01, 4] and compared the results. What we found was that the optimal α value was 0,01. This was the smallest value of α used in the cross validations and show that the model performs best when the penalty term is totally removed, i.e. running a pure OLS model. We also extended our analysis of regularization to the Ridge regression, which has a slightly altered penalty term more focused on reducing the magnitude of coefficients as opposed to removing. However, this approach yielded no better model than our original OLS model neither. Thus, we concluded that the OLS model was the best suited for analysis of medical charges. This model yielded the beta values:

	Coef
Intercept	1050.16
C(gender)[male]	999.51
C(smoker)[yes]	4999.55
C(region)	
northwest	-699.48
southeast	-498.84
southwest	-799.38
C(medical_history)	
Heart disease	2999.89
High blood pressure	-1000.87
No history	-1999.56
C(family_medical_history)	
Heart disease	3000.77
High blood pressure	-999.06
No history	-1999.66
C(exercise_frequency)	
Never	-2000.33
Occasionally	-998.40
Rarely	-1499.33
C(occupation)	
Student	-999.17
Unemployed	-1499.50
White collar	500.79
C(coverage_level)	
Premium	4999.77
Standard	1999.80
age	19.98
BMI	49.96
children	200.22

To interpret the coefficients, it is important to note that the intercept consists of:
[gender = 'female', smoker = 'no', region = 'northeast', medical_history = 'Diabetes', family_medical_history = 'Diabetes', exercise_frequency = 'Frequently', occupation = 'Blue collar', coverage_level = 'Basic']

3 Results

To analyze the factors influencing health insurance charges and explore potential gender disparities, we created a series of visualizations to provide an overview of the dataset and investigate key relationships. These visualizations not only reveal patterns within the data but also guide the focus for deeper analysis. A general inspection of the dataset reveals an equal distribution of genders, with approximately 500,000 records for both males and females (Figure 2). This balance ensures that any observed differences in charges are not a result of gender underrepresentation. Furthermore, the distribution of insurance charges is roughly normal, with most charges concentrated around $15,000 (Figure 3). This symmetry in the data provides a robust foundation for analysis and modeling.

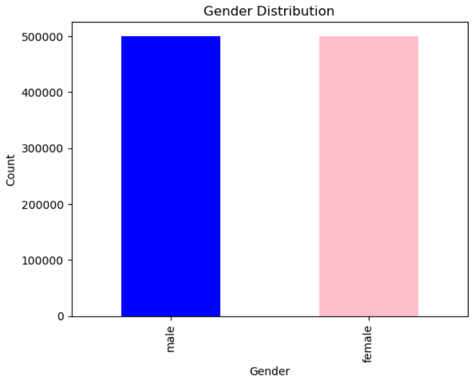

Figure 3: Representation of each gender

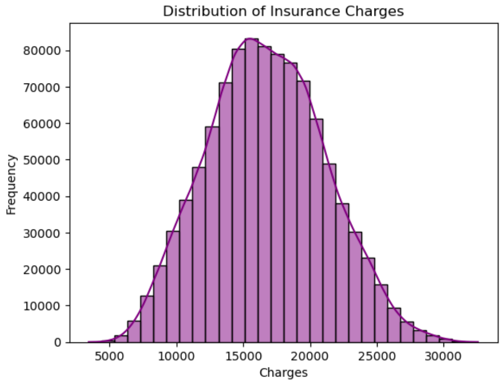

Figure 4: Distribution of charges

The correlation heatmap highlights weak but notable relationships between certain variables and insurance charges (Figure 4). Among them, the correlation between smoking and charges are very strong, suggesting that wether a person smokes or not has serious impact on how much they are charged. Other, less correlated variables that still seems to bear some significance include gender(0.11), BMI (0.1) and age (0.063). These preliminary observations emphasize the need to consider multiple variables when analyzing the dataset. Not including the information of, say smoker or non-smoker, would have serious effects on the estimated coefficients and result in omitted variable bias[WMWL21]. The effect of smoking on charges would be attributed to the other included variables and lead to erroneous conclusions.

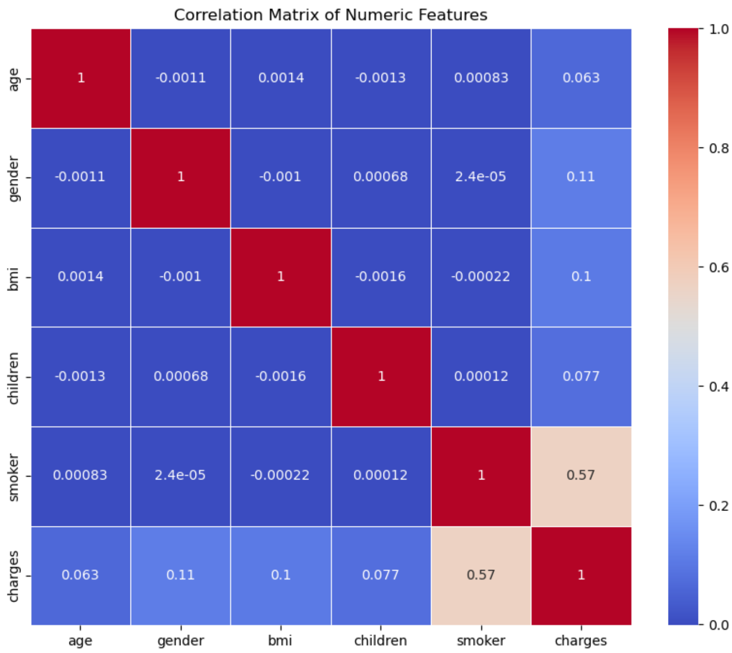

Figure 5: Correlation matrix for numeric and binary variables.

According to our model, gender has a strong effect on a persons medical insurances charges. Our model estimate that being of the gender male increase your charges with 999,51 from being a female.

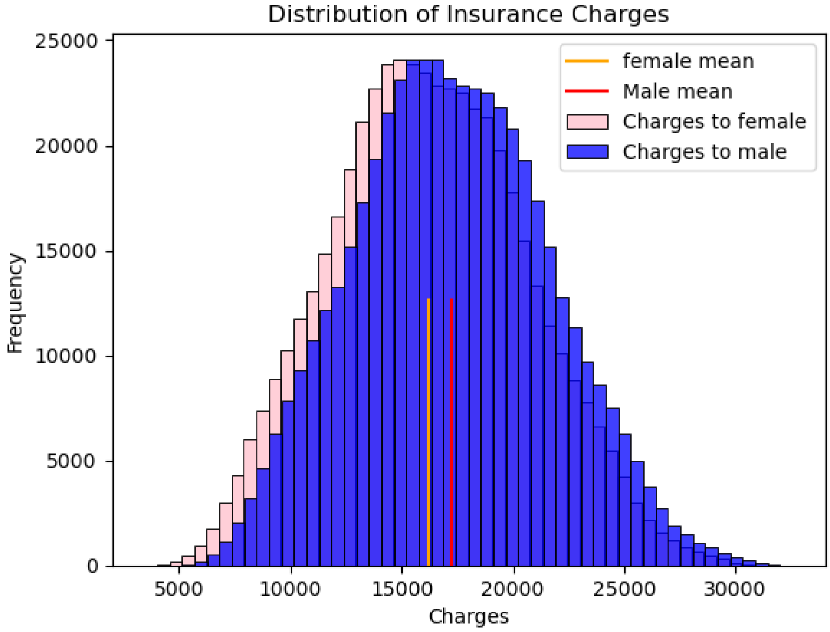

Figure 6: Distribution of charges; Male vs. Female

- Regulation violation to gender discriminate via prices?
- Ethical?

4 Conclusions

References

[DMC05] Michael Chernew Patricia Seliger Keenan David M. Cutler. Increasing health insurance costs and the decline in insurance coverage. *Health Services Research*, 40(4):1021–1039, 2005.

[Edm23] Charlotte Edmond. Us women are paying billions more for healthcare than men every year. *World Economic Forum*, 2023.

[Had07] Jack Hadley. Insurance coverage, medical care use, and short-term health changes following an unintentional injury or the onset of a chronic condition. *JAMA*, 297(10):1073–1084, 03 2007.

[Tib18] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 12 2018.

[VSMS16] M Venkataramana, M Subbarayudu, Rajani Meejuru, and K Sreenivasulu. Regression analysis with categorical variables. *International Journal of Statistics and Systems*, 11:135–143, 01 2016.

[WMWL21] R. Wilms, E. Mäthner, L. Winnen, and R. Lanwehr. Omitted variable bias: A threat to estimating causal relationships. *Methods in Psychology*, 5:100075, 2021.