Final Assignment

Group: Potato

GRA4157

GitHub Repo: https://github.com/s2110880/DataCuration-Project

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 employed woman in the US having approximately 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. Research on medical insurance and gender has mostly been centered around aggregate spend differences, health care serivce usage[KDB00] or insurance enrollment differences. 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: Dataset description

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 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.

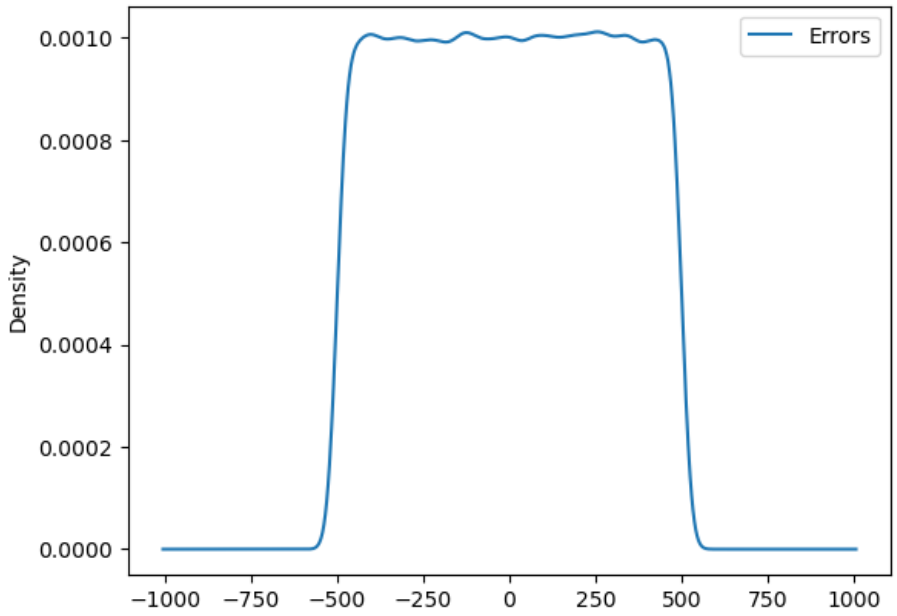

Figure 2: Distribution of the errors

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 coefficients. 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']

Lastly, we needed to validate an important assumption within OLS regression, namely that the variance of the error terms is constant for all observations(Homoscedassticity). This needs to be checked as the presence of heteroscedasticity would imply that OLS is not the minimum variance estimator for this data. To confirm that the errors are homoskedastistic, we plot the residuals versus the fitted values to see if it displays a pattern as the fitted values increase.

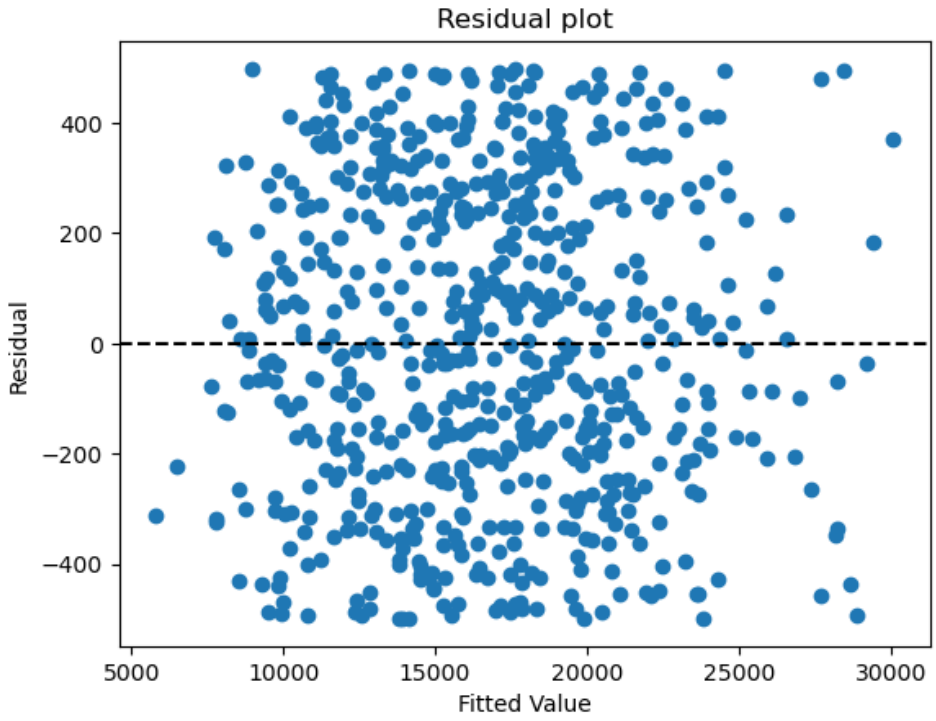

Figure 3: Residual plot

The plot does not show any discernible pattern of increasing variance in error terms as the fitted value increase, justifying our assumption that homoskedasticity.

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 4). 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 5). This symmetry in the data provides a robust foundation for analysis and modeling.

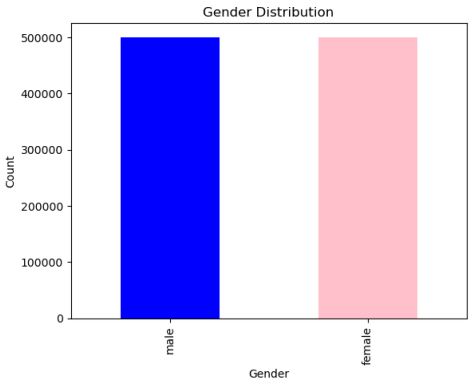

Figure 4: Representation of each gender

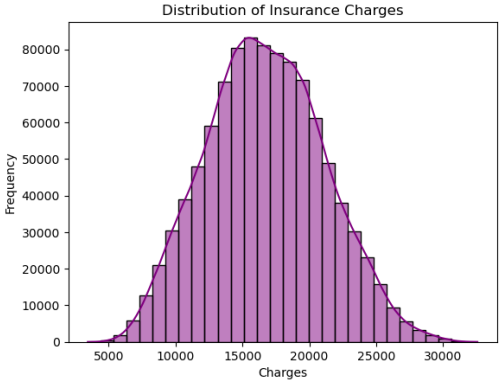

Figure 5: Distribution of charges

The correlation heatmap highlights weak but notable relationships between certain variables and insurance charges (Figure 6). 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.

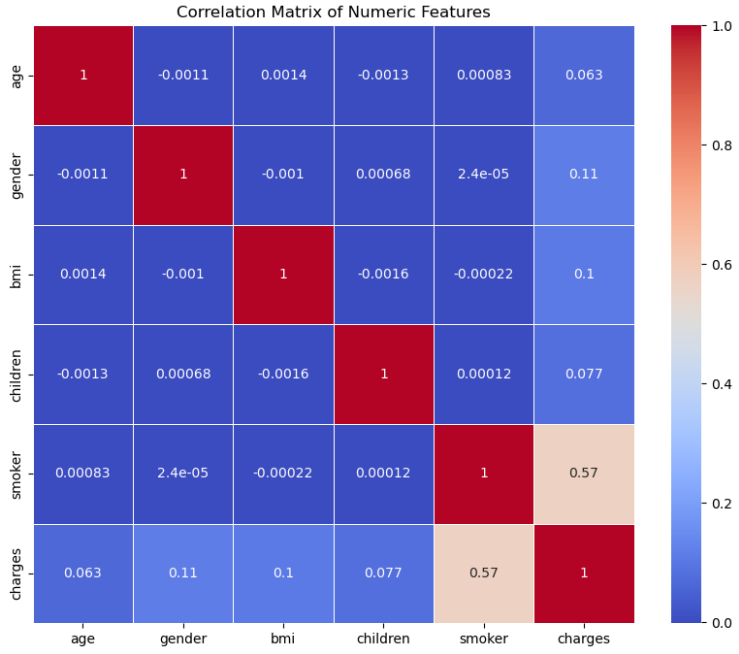

Figure 6: Correlation matrix for numeric and binary variables.

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. For EU contries, this would be illegal as the Eurpoean Commisson rules that insurance pricing must be gender-neutral[Com12]. In the US, were this data is from, according to the Civil Rights Division of the U.S. Department of Justice, federal laws prohibit discrimination based on a person's national origin, race, color, religion, disability, sex, and family status. The Civil Rights Department enforces the federal laws that prohibit discrimination in several sectors, including Public Accommodations, which covers the insurance industry[oJ00]. Although the general law prohibits sex discrimination, it is mostly under the U.S. Equal Employment Opportunity Commission it is clearly defined what constitutes discrimination based on gender and what is breaking the law[Com]. Thus, although prohibited, it is poorly defined within the regulations on what constitutes gender discrimination within public accommodations. Its been known for a long time that products marketed towards women have a tendency to be more expensive than those marketed towards men. This has been theorized to be due to gender discrimination. However, since this approach to pricing does not restrict one gender to one price, it has remained legal as it is viewed as firms grouping consumers into their willingness to buy and charging accordingly. However, this excuse is not relevant for the insurance industry, as insurance companies offer personalized prices for their consumers based on their predictors of risk. Whether or not firms use gender as a predictor for insurance charges is hard to prove, but given the results of our model, there is evidence to suggest that gender has a significant impact on prices.

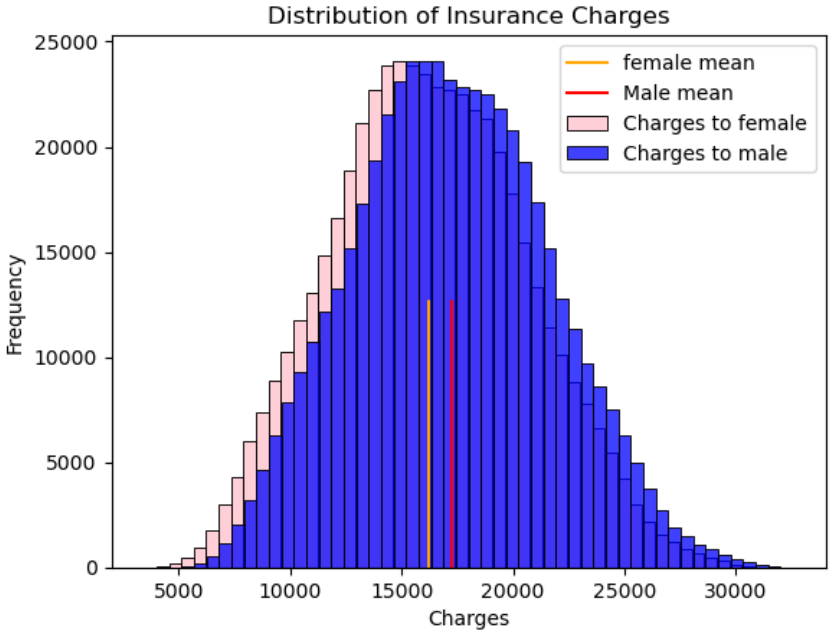

Figure 7: Distribution of charges; Male vs. Female

Gender is not the only determining factor of the insurance charges. External factors such as lifestyle and health behaviors exert greater influence according to our model. Most notably is the coefficient for smoking status, which indicate that charges for smokers are substantially higher than those for non-smokers, regardless of gender.

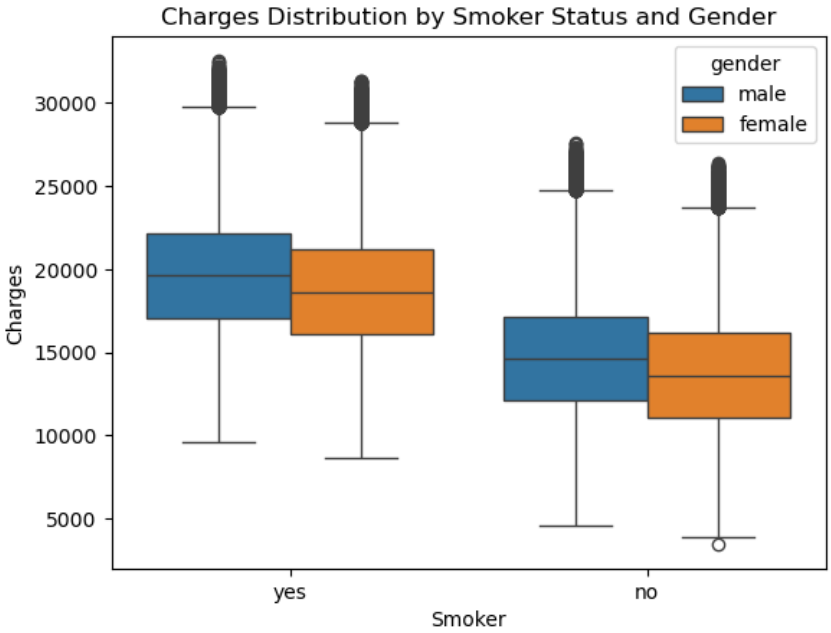

Figure 8: Charges Vs. Smoker status

Smoking status combined with age and BMI are all justifiable predictors of risk for the insurance companies and hence it is reasonable that there variables correlate positively with charges. In our model, each additional year increases charges by $19.98, indication a steady rise in perceived health risks with age. A one-unit increase in BMI raise charges by $49.96, reflecting the link between body mass and health care costs.

Lastly, we also conducted an error analysis to evaluate the accuracy and fairness of the model predictions. This analysis identifies where the model predictions deviate most significantly from the actual values, focusing on patterns in the errors. After generating predictions on the dataset, the following error metrics were calculated:

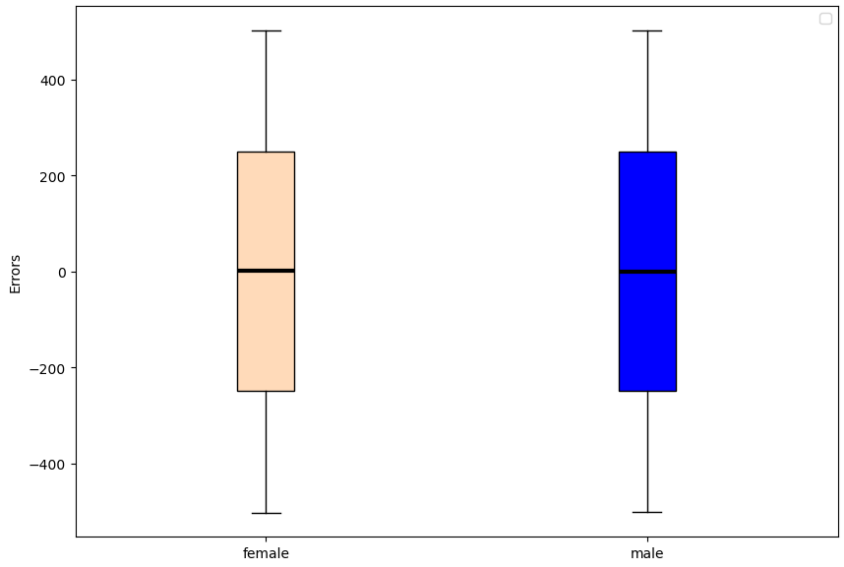

Figure 9: Boxplot of error distribution

- **Error**: The difference between the actual charges and the predicted charges $y - \hat{y}$.
- **Absolute Error**: The absolute value of the difference $|y - \hat{y}|$.
- **Absolute Error Percentage**: The percentage error relative to the actual charges $\frac{|y-\hat{y}|}{y} * 100$.

The dataset was then sorted based on the absolute error percentage in descending order, and the top 1,000 largest errors were extracted for further investigation. By looking for discernible patterns, we found that no data point much higher error than the rest. Grouping the errors into groups based on gender, we find that the errors are almost identical with both median 0.

This result from our error analysis shows that our model is exceptional at capturing the patterns in the data set, adding to the credibility of its coefficient estimates.

Despite the robustness of the analysis, several limitations must be acknowledged; The dataset seems to be simulated, which, while comprehensive, may not fully capture real-world complexities. Our analysis assumed all predictors were equally reliable, which may not be true in real-world scenarios where data quality varies.

4 Conclusions

This analysis revealed key factors influencing health insurance charges, with a particular emphasis on the role of gender. While gender was statistically significant, with males incurring higher charges ($999.51) than females, it was not the dominant factor. Variables such as smoking status and medical history had far more substantial effects on charges, with smoking alone increasing cost by nearly $5,000. The model demonstrated high accuracy, achieving a Mean Absolute Percentage Error(MAPE) of 1.62%, which underscores its reliability in capturing the underlying patterns determining insurance costs. Furthermore, an error analysis suggested that there was no major systematic bias in the overall predictions for males and females.

The findings of this study hold significant implications. Since our model has such high in-sample and out-of-sample fit, where almost all the variance in the dependent variable can be explained by the independent variables, it is very likely that there is no omitted variable bias and the coefficient for gender is the true correlation between medical charges and gender. Thus, our research would suggest that the medical insurance industry in the US is not acting in accordance with the federal laws.

References

[Com] U.S. Equal Employment Opportunity Commision. Your rights.

[Com12] European Commission. Factsheet: Eu rules on gender-neutral pricing in insurance, 12 2012.

[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.

[KDB00] Jay L. Helms Edward J. Callahan John A. Robbins Klea D. Bertakis, Rahman Azari. Gender differences in the utilization of health care services. *The Journal of Family Practice*, 49(2):147–152, 2000.

[oJ00] U.S. Department of Justice. Federal protections against national origin discrimination, 2000.

[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.