STAT 170 Final Report

Brenda Lorenzo, Iqra Naz, Neha Matthews, Vishal Gondi

2023-12-14

Contents

List of Figures

Introduction

For our final project, the dataset we decided to explore was the Insurance dataset. We focused on the main question: Can we successfully predict insurance charges? To gain a better understanding on our main objective, we narrowed our focus to address three key sub-questions:

- Do older individuals tend to have higher insurance charges?
- Does smoker status have a significant impact on insurance charges?
- Do smokers or non-smokers pay more insurance charges depending on their BMI?

Our motivation for this research comes from the concern that insurance costs are very expensive these days. We want to investigate the specific factors leading to an increase in insurance costs. Our goal is to advocate for a healthcare system that is both accessible and affordable for all.

Data Description

In our analysis, we considered a set of independent variables: 'Age,' 'Sex,' 'BMI,' 'Children,' 'Smoker,' and 'Region.' These include continuous quantitative variables such as 'Age,' 'BMI,' and 'Charges,' as well as categorical variables like 'Sex,' 'Children,' 'Smoker,' and 'Region.' The dependent variable, 'Charges,' are the insurance charges in USD.

The summary statistics of the dataset provide insights into the central tendency and spread of each quantitative variable. For instance, the mean insurance charge is `$13,270`, ranging from a minimum of `$1,122` to a maximum of `$63,770`.

Upon examining scatter plots of relationship between the response variable and each independent variable, we noticed interesting associations. For example, the scatter plots for 'Smoker,' 'Region,' 'Sex,' and 'Children' suggest that a bar plot may better represent the data since they are all categorical. Furthermore, the scatter plots for 'Age' and 'BMI' indicate a very weak to no association with 'Charges'.

Exploratory Data Analysis

age	sex	bmi	children	smoker	region	charges
19	female	27.900	0	yes	southwest	16884.924
18	male	33.770	1	no	southeast	1725.552
28	male	33.000	3	no	southeast	4449.462
33	male	22.705	0	no	northwest	21984.471
32	male	28.880	0	no	northwest	3866.855
31	female	25.740	0	no	southeast	3756.622

Dependent Variable

Moving on to the Exploratory Data Analysis (EDA), we notice that the histogram for the 'charges' variable appears to have a right-skewed distribution. Meaning the data's distribution, mostly consists of data points on the left-hand side, showing that the majority of observations are smaller, while a few larger observations contribute to the right tail.

To decrease the right-skewness of the data, we applied a log transformation to the 'charges' variable. This transformation aiding with the spread of the values, making the distribution more symmetrical.

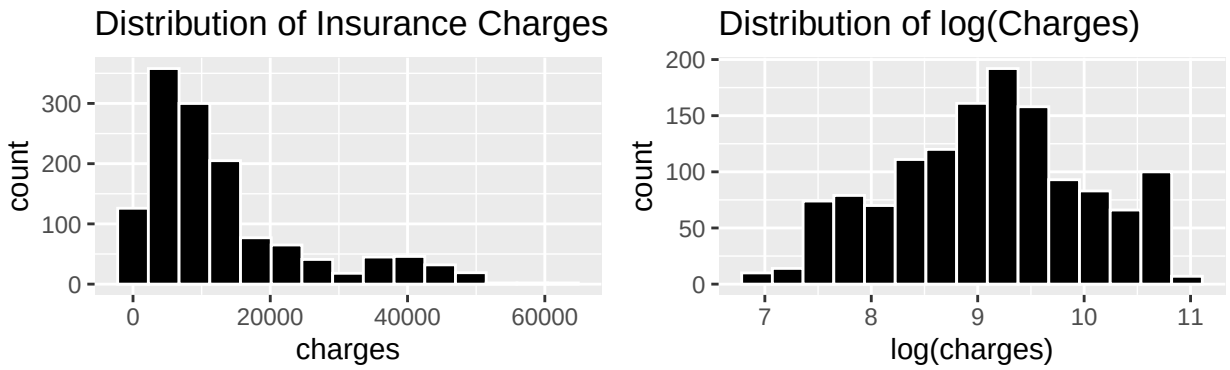

Figure 1: Charges Bar Plot

Independent Variables

When examining the graphs of the predictors, we can see that the age histogram follows a uniform distribution with a few exceptions. This suggests that the dataset contains individuals across various age groups. The histogram for the number of children reveals a right-skewed distribution, indicating that a majority of individuals have fewer children. As you can see, BMI displays a relatively normal distribution.

Examining the distribution of sexes one can see similar frequencies, which is a crucial factor to prevent bias in the predictive model. Similarly, consistent frequencies across different regions imply a balanced representation of the geographical areas. Additionally, the dataset appears to consists mainly of non-smokers, indicating a class imbalance. It is important to point out that we did not find sex and region to be statistically useful so it is not included in our final model.

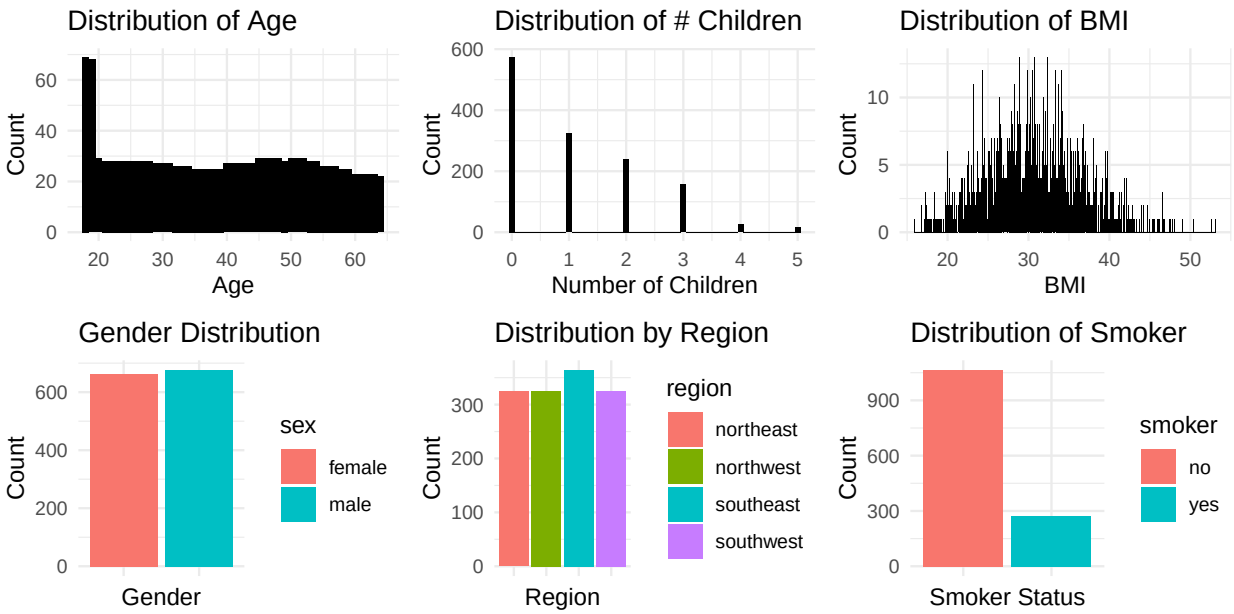

Figure 2: Distribution of Predictor Variables

Independent Variables vs Dependent Variable

When examining the relationship between predictor and response variables, the age vs. charges scatter plot reveals three distinct sets of points, each following a separate positive linear pattern. The topmost group consists predominantly of smokers, the middle group is a mix of smokers and non-smokers, while the bottommost group consists only of non-smokers. Hence it may be ideal to make three different models if time permits. This suggests that there may be interaction effects between age and another variable.

The BMI vs. charges scatter plot indicates no obvious relationship between them. The sex vs charges box plot appears to be skewed and shows that there could be a disparity in the distribution of the sexes. The region vs charges box plot also appears to be skewed with several outliers.

When examining the smoker vs. charges box plot it provides sufficient evidence that smokers seem to face higher charges in comparison to non-smokers. Moreover, these observations illustrate the significance of interactions,variable relationships, and potential disparity in the distributions included in the dataset, which are factors that will help us build a solid model.

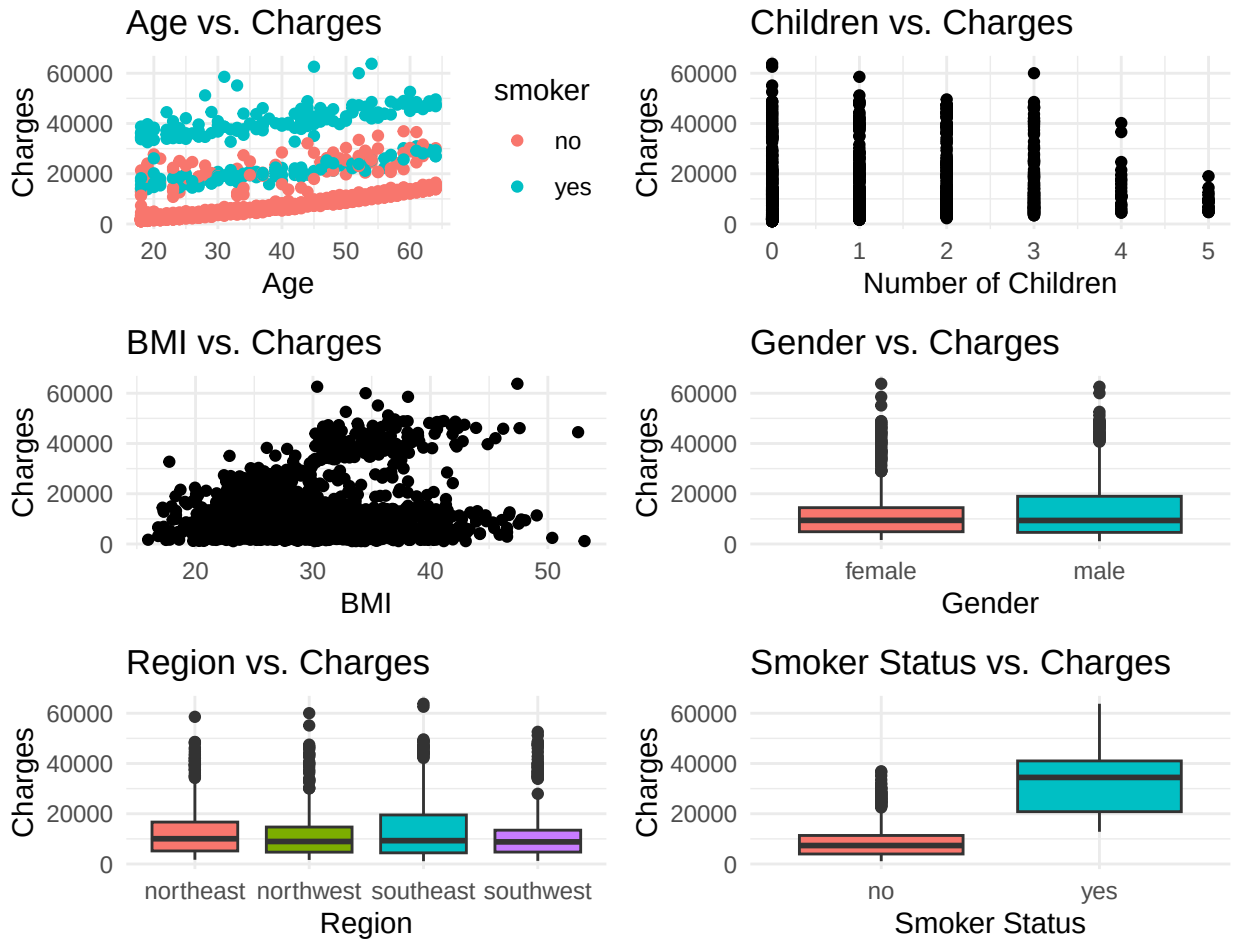

Figure 3: Predictors vs Charges

Regression Analysis

In order to find our final model we first sought to find the functional form of our base model which included all the numeric variables: age, bmi, and children. This would allow us to see if any transformations or interactions needed to be added to the model in order to meet our assumptions. The first step we took in trying to find the functional form was by removing the contribution of variables that have a strong relationship with Y. However, this step did not improve the normality of Y, which led us to believe transformations would not be necessary for our model. The next step we took was to see if there were any interactions between variables. One interaction we found was between bmi and smoker.

This led us to add an interaction term between bmi and smoker to our model. Another addition we made to our model is by adding an intercept of 0 to the model. The reason for this is that if age was set to 0 meaning someone is not alive, it does not make sense for a dead person to incur insurance charges.

With these additions to the model we saw that the normality and linearity assumptions were met. The only assumption we were not able to meet was constant variance. We believe that there are hidden variables that we do not have access to that are causing this issue.

We then executed step wise regression to find the best possible model with our new base model. However, the stepwise regression algorithm removes the intercept of 0 which is crucial to our model. Due to the addition, of the y-intercept, which does not have a meaningful interpreation, the R^2 was decreased greatly. We concluded that step wise regression would not improve our model based on this.

Our final model ends up being: $log(charges) = 0 + 0.071486(age) + 0.245349(children) + 0.229045(bmi) - 0.050251(bmi)(smoker) + \epsilon$

Coefficient for age: holding other variables constant, a one unit increase in age is associated with a 0.071486 increase in the logarithm of charges. (p-value < 2e-16) Coefficient for children: holding other variables constant, a one unit increase in children is associated with a 0.245349 increase in the logarithm of charges. (p-value 3.19e-15) Coefficient for bmi: holding other variables constant, a one unit increase in bmi is associated with a 0.229045 increase in the logarithm of charges. (p-value <2e-16) Coefficient for interaction between being a smoker and bmi: each one unit increase of BMI is associated with a decrease of 0.050251 units in log-transformed charges, holding other variables constant. (p-value < 2e-16)

The R^2 value of our model is 0.9777. This means that 97.77% of the variability in insurance charges can be explained by our model.

The residual standard error of our model is 1.366 on 1334 degrees of freedom. Our RSE is pretty close to 1 which is why we feel that our model is quite accurate in predicting in insurance charges. # Conclusion ## Model Overview Despite our efforts to achieve a positive trend in our data along with constant variance; through the trial and error of several different models, we encountered challenges in satisfying both. Despite this, our final model showcases a high R-squared value of 0.9777, indicating that a large portion of the variability in the logarithm of charges can be explained by the selected predictors.

Final model:

$log(charges) = 0 + \beta_1 * age + \beta_2 * children + \beta_3 * bmi + \beta_{4,5} * bmi : smoker + \epsilon$

where:

- $\beta_1, \beta_2, ..., \beta_n$ are the coefficients for the selected predictors
- ϵ represents the random error term

The F-statistic of 1.465e+04 and a low p-value (< 2.2e-16), provides strong evidence for the overall significance of our model in accurately predicting the logarithm of charges. While we weren't able to satisfy the constant variance assumption, the high metrics show our model's ability to provide valuable insights into the relationship between the predictors and the logarithm of charges.

Interpretation

Coeff	Std.Err	t.value	Pr(>|t|)	Resid.Std.Err	Multiple R squared	Adjusted R squared	F statistic	DF	P value
0.071	0.0025	29.014	0	1.366481	0.9777391	0.9776724	14647.95	4	1334
0.245	0.0308	7.978	0	1.366481	0.9777391	0.9776724	14647.95	4	1334
0.229	0.0040	56.753	0	1.366481	0.9777391	0.9776724	14647.95	4	1334
-	0.0030	-	0	1.366481	0.9777391	0.9776724	14647.95	4	1334
0.050		16.996							

The positive coefficient for age suggests that, holding other variables constant, a one-unit increase in age is associated with a 0.071486 increase in the logarithm of charges. This aligns with the expectation that older individuals tend to experience higher insurance expenses due to increased health risks and more costly treatments.The positive coefficients for both children and bmi suggest that individuals with more children or higher BMI tend to have higher insurance charges.

The negative coefficient for bmi:smoker ($\beta_{bmi:smoker} = 0.050251$) suggests that the effect of BMI on charges is modified by smoking status. Specifically, for smokers, the relationship between BMI and charges differs compared to non-smokers. While BMI still positively influences charges for smokers, the effect is less pronounced.

The residuals provide insight into the model's performance, indicating the differences between observed and predicted values. The model overall demonstrates high significance, as indicated by the low p-values for the coefficients.

Multicollinearity

When running VIF we found that the values were all under 10. We removed the BMI factor since it was not outputting a VIF value since the addition of the BMI was causing there to be a almost perfect multicollinearity in our model.

Residual Diagnostics

	Test.Statistic	P.value
W	0.998352	0.2231491

The QQ plot and the results from the Shapiro-Wilk normality tests suggest that the residuals of the model demonstrate a approximate normal distribution. This is good, as it implies that the assumption of normality for the model residuals is satisfied.

However, an interesting observation can be seen from the residual vs. fitted plot, where a negative slope is evident. This negative trend causes concern for the assumption of equal variance, potentially showing the presence of heteroscedasticity. We attempted many different transformation and addition of other independent variables into the model, however we were not able to find a proper solution to the negative trend in the plot. The heteroscedasticity in the model implies that there are unaccounted factors that impact the variability of the residuals.

On the other hand, the assumption of independence, tested through the Durbin-Watson test, yields a p-value of 0.82. Given this high p-value, we accept the null hypothesis, indicating that the residuals show independence.

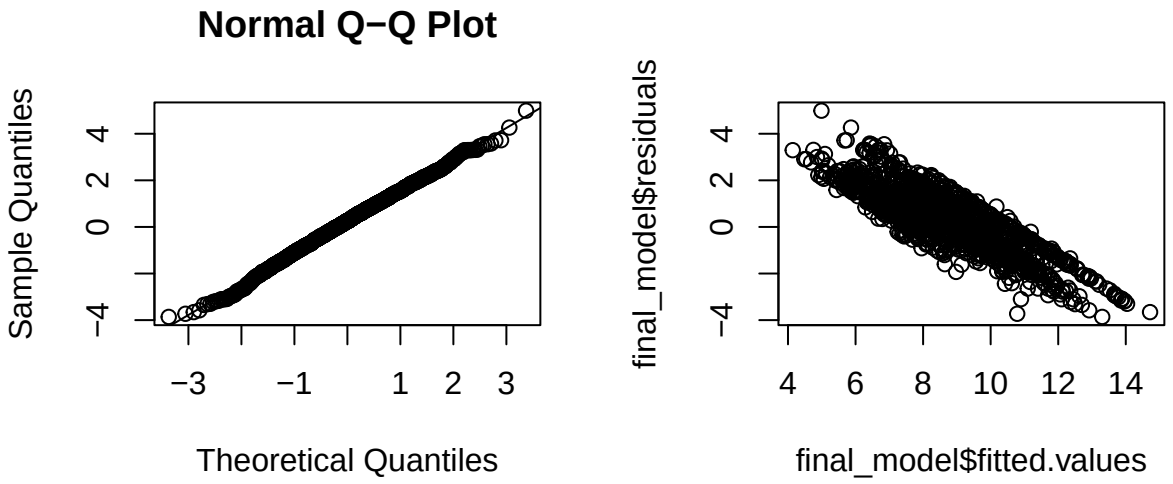

Figure 4: Assumptions

For the linearity assumption, it seems that the model is linear but has a negative slope.

In summary, while the model residuals does pass normality,linearity,and independence, the observed negative slope in the residual vs. fitted plot hints at potential heteroscedasticity. Despite multiple attempts to fix this issue, we were not able to find a proper solution. Recognizing this as the model's limitations provides insights on model improvement in the future.

Limitations

After trying different transformation methods, we were not able to fix the negative slope in the residuals vs fitted plot. This indicates that there might be some missing variables or interactions that cause of the downward effect in our plot.

A solution to this is to add more relevant variables or interaction terms could possibly improve the model's fit and address the downward pattern in the residuals vs. fitted plot. Some potential variables that we thought may be helpful are credit history or health history since these are valid predictors that can help predict charges.

Another way to improve our model is to create three separate models. We can see from the age versus charges plot that there seems to be three different subsets so we can create individual models for each subset which may help with the negative pattern that we observe in the model. Next, having more comprehensive data will also help since we saw that in our current data set there is data missing.

Overall, we were able to take a look at potential factors that are increasing the charges for healthcare. We would like to further explore factors that contribute to this increased insurance charges and futher improve our model.