

STAT 547M Project

Diana Lin & Nima Jamshidi

14/03/2020

Contents

| | | |
|----------|------------------------------|----------|
| 1 | Introduction | 1 |
| 2 | Research Question | 1 |
| 3 | Data Description | 2 |
| 4 | Exploring the Dataset | 2 |
| 5 | Methods | 5 |
| 6 | Results | 7 |
| 7 | Discussion | 7 |
| 8 | Conclusion | 8 |

1 Introduction

The dataset we have chosen to work with is the “Medical Expenses” dataset used in the book Machine Learning with R, by Brett Lantz. This dataset was extracted from Kaggle by Github user @meperezcuello. The information about this dataset has been extracted from their GitHub Gist.

This dataset is very interesting as the USA does not have universal healthcare, and is known for bankrupting its citizens with hospital visits despite having insurance. It will be interesting to see the relationship between characteristics of a beneficiary, such as **BMI** and **Smoking** status, and the **charges** incurred.

2 Research Question

In this study, we are analyzing the data to find a relationship between the features and the amount of insurance cost.

Does having an increased BMI increase your insurance costs? What about age? Number of dependents? Smoking status? Are certain areas of the USA associated with higher insurance costs?

In order to answer the questions above we’re planning to perform a linear regression analysis and plot the regression line and relevant variables. The variables need to be normalized before performing the regression analysis.

Table 2: summary of the dataset

| age | sex | bmi | children | smoker | region | charges |
|---------------|------------------|---------------|---------------|------------------|------------------|--------------|
| Min. :18.00 | Length:1338 | Min. :15.96 | Min. :0.000 | Length:1338 | Length:1338 | Min. : 1122 |
| 1st Qu.:27.00 | Class :character | 1st Qu.:26.30 | 1st Qu.:0.000 | Class :character | Class :character | 1st Qu.: 474 |
| Median :39.00 | Mode :character | Median :30.40 | Median :1.000 | Mode :character | Mode :character | Median : 938 |
| Mean :39.21 | | Mean :30.66 | Mean :1.095 | | | Mean :13270 |
| 3rd Qu.:51.00 | | 3rd Qu.:34.69 | 3rd Qu.:2.000 | | | 3rd Qu.:1664 |
| Max. :64.00 | | Max. :53.13 | Max. :5.000 | | | Max. :63770 |

3 Data Description

This dataset explains the medical insurance costs of a small sample of the USA population. Each row corresponds to a beneficiary. Various metadata was recorded as well.

The columns (except the last one) in this dataset correspond to metadata, where the last column is the monetary charges of medical insurance. Here are the possible values for each of the columns:

| Variable | Type | Description |
|----------|---------|---|
| Age | integer | the primary beneficiary's age in years |
| Sex | factor | the beneficiary's sex: female or male |
| BMI | double | the beneficiary's Body Mass Index, a measure of their body fat based on height and weight (measured in kg/m2), an ideal range of 18.5 to 24.9 |
| Children | integer | the number of dependents on the primary beneficiary's insurance policy |
| Smoker | factor | whether or not the beneficiary is a smoker: yes or no |
| Region | factor | the beneficiary's residential area in the USA: southwest , southeast , northwest , or northeast |
| Charges | double | the monetary charges the beneficiary was billed by health insurance |

4 Exploring the Dataset

Here is a summary of the dataset, and the values of each variable (Table 2):

Next, we want to inspect the data set to see if there is any correlation between the variables. From now on we want to consider charges as our dependent variable. In order to analyze correlation between variables, the ones that are categorical with two categories, are translated into binary vectors. The only categorical variable with more than two categories, is region. We split this variable into four different binary vectors, each indicating if the sample data has category (1) or not (0).

After using dummy variables for sex, smoker, and region, according to the correlogram show in Figure 1, smoker and charges has the strongest correlation of 0.79. No high collinearity between independent variables is observed.

In order to check if there is any cluster of data points, we use faceted plot (Figure 2). While the data between regions and sex does not appear to vary much, the smokers vs nonsmokers of each facet appear to cluster together, with the non-smokers having an overall lower medical cost.

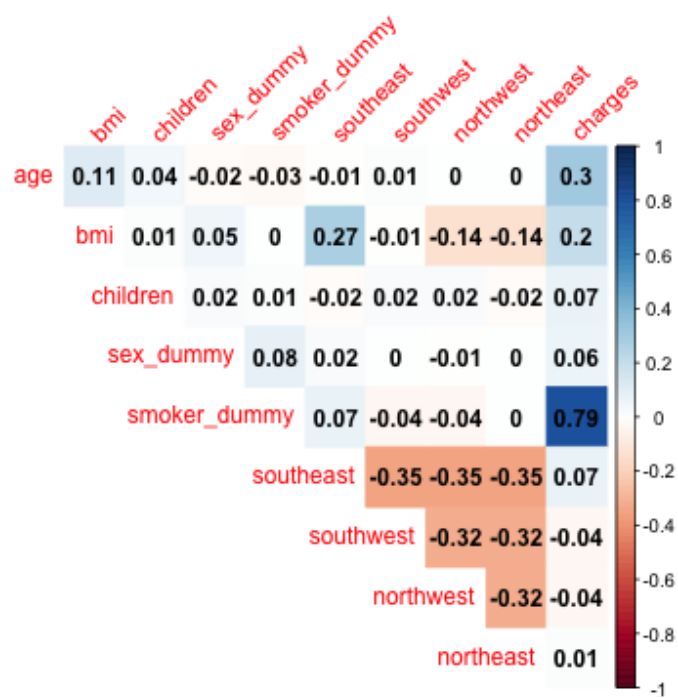


Figure 1: Correlation plot

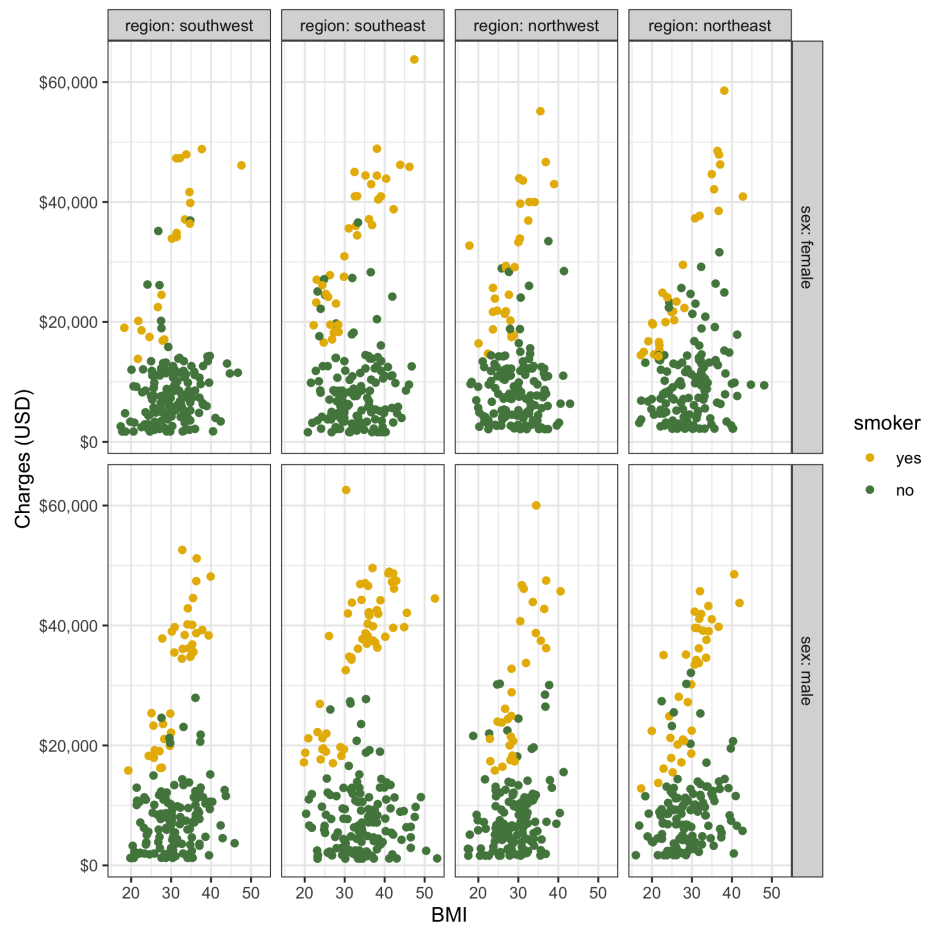


Figure 2: Exploring the medical costs dataset

How is the distribution of sex among different age groups? Looking at Figure 3, there appears to be more beneficiaries in the 20-60 age range. The biggest difference in the number of beneficiaries from different sex is seen in the 20-30 bracket.

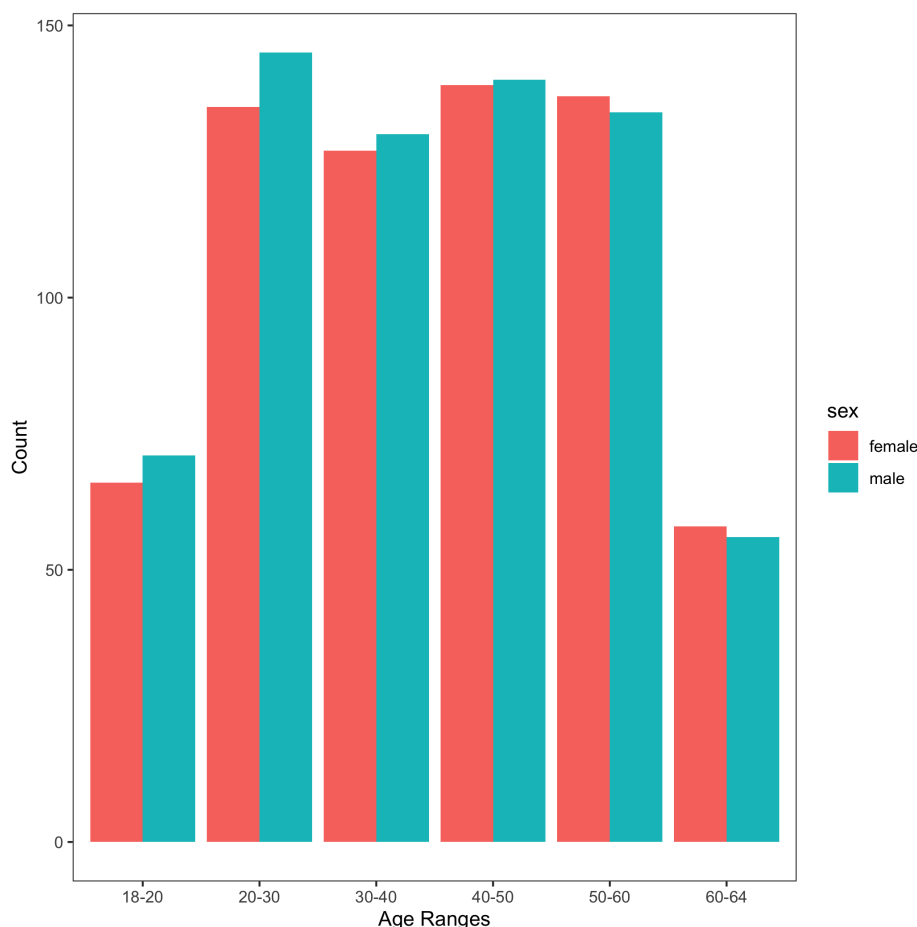


Figure 3: Distribution of age ranges

How about the distribution of sex among the regions? Figure 4 shows the distribution of sex in each of the four regions. At a glance, the dataset looks very even when it comes to sex, but there are slightly more beneficiaries in the southeast.

5 Methods

Here we use multiple linear regression to study the relations between the independent variables and the dependent one, charges. Below you can find the results of the regression in Table 3. `lm` function in R transforms a categorical variable with n levels into $n-1$ variables each with two levels to insure the variables are independent. Here we can see that variables age, bmi, children, and smoker are significantly important in the regression. Sex is an insignificant factor in the model.

In Table 4 we can see that the r-squared value is 0.75. Figure 5 shows the diagnostics plots of the regression model.

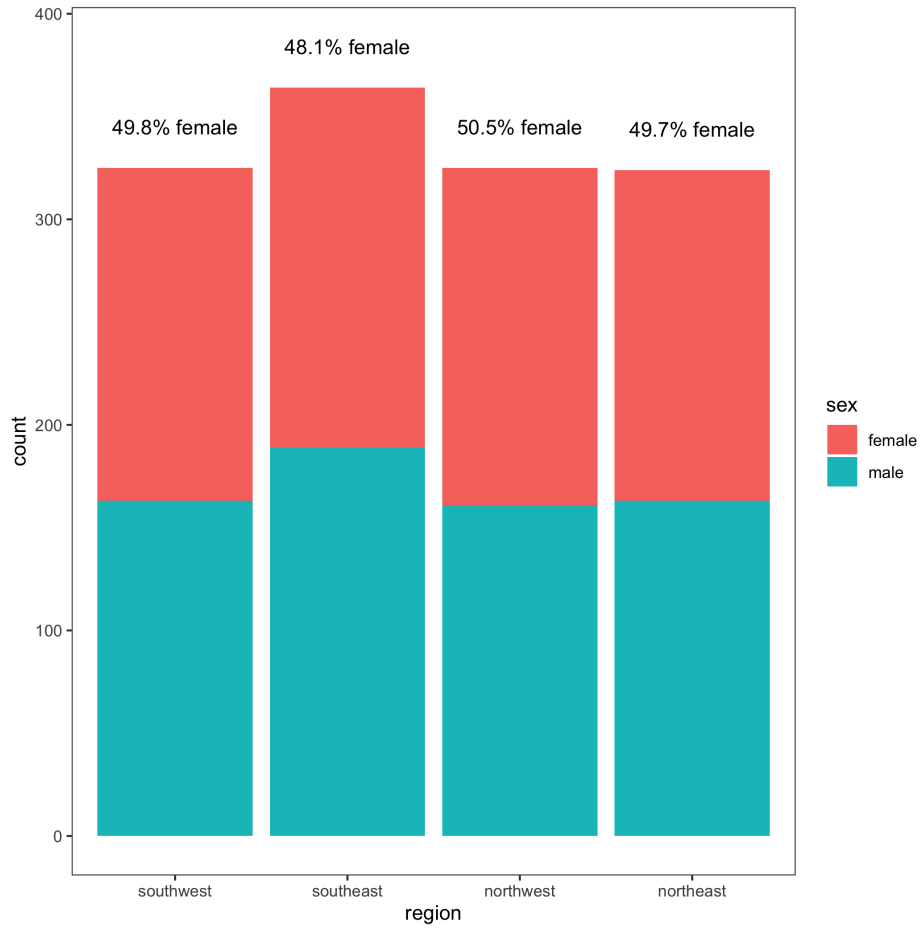


Figure 4: Sex distribution across four regions

Table 3: Summary of the model's variables and their respective coefficients

| term | estimate | std.error | statistic | p.value |
|-----------------|-------------|-----------|-------------|-----------|
| (Intercept) | -11938.5386 | 987.81918 | -12.0857530 | 0.0000000 |
| age | 256.8564 | 11.89885 | 21.5866552 | 0.0000000 |
| sexmale | -131.3144 | 332.94544 | -0.3944020 | 0.6933475 |
| bmi | 339.1935 | 28.59947 | 11.8601306 | 0.0000000 |
| children | 475.5005 | 137.80409 | 3.4505546 | 0.0005770 |
| smokeryes | 23848.5345 | 413.15335 | 57.7232020 | 0.0000000 |
| regionnorthwest | -352.9639 | 476.27579 | -0.7410914 | 0.4587689 |
| regionsoutheast | -1035.0220 | 478.69221 | -2.1621870 | 0.0307817 |
| regionsouthwest | -960.0510 | 477.93302 | -2.0087563 | 0.0447649 |

Table 4: Model summary

| r.squared | adj.r.squared | sigma | statistic | p.value | df | logLik | AIC | BIC | deviance | df.residu |
|-----------|---------------|----------|-----------|---------|----|-----------|----------|---------|-------------|-----------|
| 0.750913 | 0.7494136 | 6062.102 | 500.8107 | 0 | 9 | -13547.75 | 27115.51 | 27167.5 | 48839532844 | 132 |

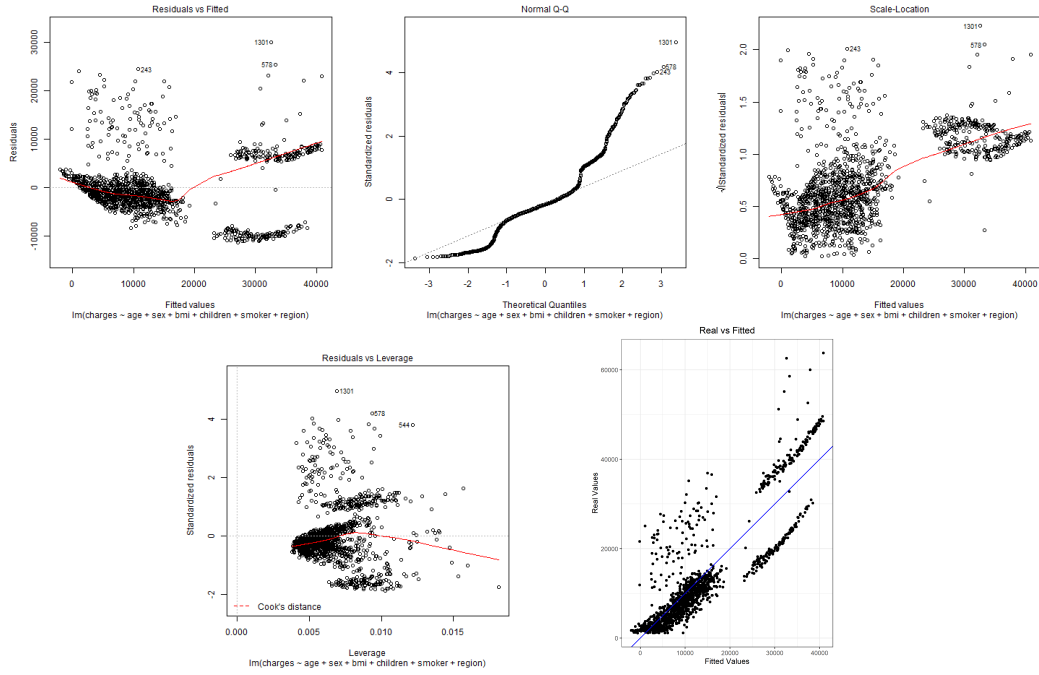


Figure 5: regression diagnostics plots

Table 5: Estimated values their statistics

| charges | age | sex | bmi | children | smoker | region | .fitted | .se.fit | .resid | .hat | .adj.r.squared |
|-----------|-----|--------|--------|----------|--------|-----------|-----------|----------|------------|-----------|----------------|
| 16884.924 | 19 | female | 27.900 | 0 | yes | southwest | 25293.713 | 586.0725 | -8408.7890 | 0.0093467 | 60.0 |
| 1725.552 | 18 | male | 33.770 | 1 | no | southeast | 3448.603 | 448.8438 | -1723.0505 | 0.0054821 | 60.0 |
| 4449.462 | 28 | male | 33.000 | 3 | no | southeast | 6706.988 | 480.0578 | -2257.5265 | 0.0062711 | 60.0 |
| 21984.471 | 33 | male | 22.705 | 0 | no | northwest | 3754.830 | 460.7247 | 18229.6404 | 0.0057761 | 60.0 |
| 3866.855 | 32 | male | 28.880 | 0 | no | northwest | 5592.493 | 424.3699 | -1725.6382 | 0.0049005 | 60.0 |
| 3756.622 | 31 | female | 25.740 | 0 | no | southeast | 3719.826 | 454.5231 | 36.7958 | 0.0056217 | 60.0 |

6 Results

In Table 5 you can find a number of examples of the data with their fitted value.

7 Discussion

Based on the “Residuals vs Fitted” and “Real vs Fitted” graphs, we can see that the model fairly works for charges under 2000\$. There are three clusters in these graphs with similar slopes. There is a gap between charges under and over 2000\$ which might be relevant to the weak estimates of the model over 2000\$. If we apply linear regression on each cluster we will get similar coefficients for the variables with different intercepts. Each cluster might be attributed to a different disease group and in each of them the impacts of age, smoking, bmi and etc. are similar.

8 Conclusion

We were able to do a linear regression on our dataset. The results show that there is an association relationship between age, bmi, number of children, and smoking with medical charges. interestingly, gender does not affect medical charges. Diagnostic plots reveal that the data is not completely normally distributed. Moreover, three clusters of records are present in the dataset, which might be representative of different types of diseases.