Final Project: Multiple Linear Regression for Health Insurance Forecast

Author: Alejandro

Discussants:

https://www.r-graph-gallery.com/199-correlation-matrix-with-ggally.html

https://www.iffcotokio.co.in/health-insurance/10-factors-that-affect-your-health-insurance-premium-costs

https://www.investopedia.com/ask/answers/052015/what-main-business-model-insurance-companies.asp

Introduction

Over the last decades, there has been a significant rise in the number of people buying health insurance in the US, owing to the increasing cost of medical treatment. People are spending thousands of dollars every year to ensure that they and their family have a financial cushion in case of a medical emergency–which, in our world, has the potential to send the average person into bankruptcy.

However, like all private companies, medical insurers aim to make a profit. They do so by following a business model centered around assuming and diversifying risk, wherein individual payers are charged premiums in exchange for insurance coverage, and the premiums are then reinvested into other assets to generate even more revenue. The goal is that if a company prices its risk effectively, it should bring in more revenue than it spends on medical expenses for the insured population. To achieve this, an insurer would need to assess how likely a buyer is to trigger their conditional payment within the time frame outlined by the policy. This will depend on personal information about the insured population, which will often be limited (by law) to a few variables. Therefore, to put themselves in a position to make profits, insurance companies must figure out a way to model average medical care expenses based on limited population trends.

The aim of this project is to tackle this problem using multiple linear regression. I will be using a dataset called "Medical Cost Personal Datasets" which is freely available on Kaggle. In particular, I will use this dataset to create a model for predicting the medical costs billed by health insurance on an individual given some of the independent variables of the dataset. These variables will include: 1) *Age*: Health insurance premiums are based on the probability of you falling sick. Hence, it makes intuitive sense that older individuals are considered as high-risk customers, and will thereby be charged higher premiums. 2) *Body Mass Index (BMI)*: People with higher BMI values have a higher risk of getting affected by heart-related problems, diabetes,and other ailments. Therefore, the premium cost for such individuals tends to be higher compared to those with a normal BMI. 3) *Smoking*: Individuals who smoke tobacco are at an increase chance of developing cancer, heart disease, stroke, lung diseases, diabetes, among other health complications. Therefore, insurance companies will want to charge higher premiums to individuals who smoke compared to non-smokers. The dataset contains additional variables such as sex, location, and number of children. However, these variables will not be considered as per the scope of this project. Please see below for a sample of the dataset, which overall contains 1338 rows.

```
insurance_data <- read.csv("insurance.csv")
head(insurance_data)
```

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Data wrangling

To get the insurance dataset in shape for my analyses, I did a very mild cleansing process. All of the data was already in the format that I needed, so all I had to do was select the variables that I wanted to keep as predictors for my multiple linear regression model: two continuous (age and BMI) and one categorical (smoker), as well as my continous dependent variable (charges). I did this using the "select" function from the "dplyr" library.

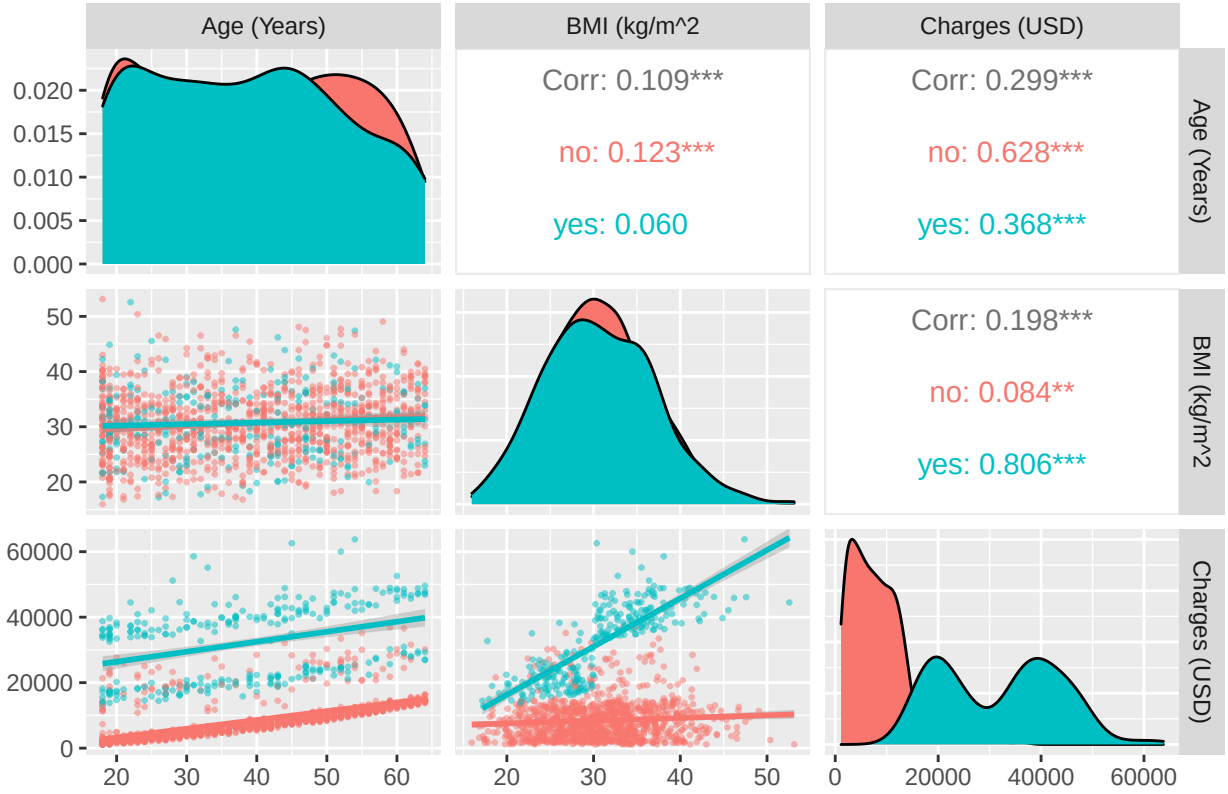
```
insurance_data <- select(insurance_data, age, bmi, smoker, charges)
```

Data Visualization

The first step in our analysis to to visualize the insurance dataset. Using the "ggpairs" function from the "GGally" library, I created a matrix of scatter plots, density curves and Pearson's product-moment correlation values to compare all continuous variables in the dataset (see Figure 1). I also separated the data categorically by color, with smokers represented in turquoise and non-smokers represented in red, and with corresponding simple linear regressions in each scatter plot (including confidence intervals).

```
ggpairs(insurance_data, columns = c(1, 2, 4), ggplot2::aes(colour=smoker),
        lower=list(continuous=GGally::wrap("smooth", alpha=0.5, size=0.5)),
        title = "Figure 1. Visualization of Insurance Dataset",
        columnLabels = c("Age (Years)", "BMI (kg/m^2", "Charges (USD)")) +
  theme(plot.title = element_text(hjust = 0.5))
```

Figure 1. Visualization of Insurance Dataset

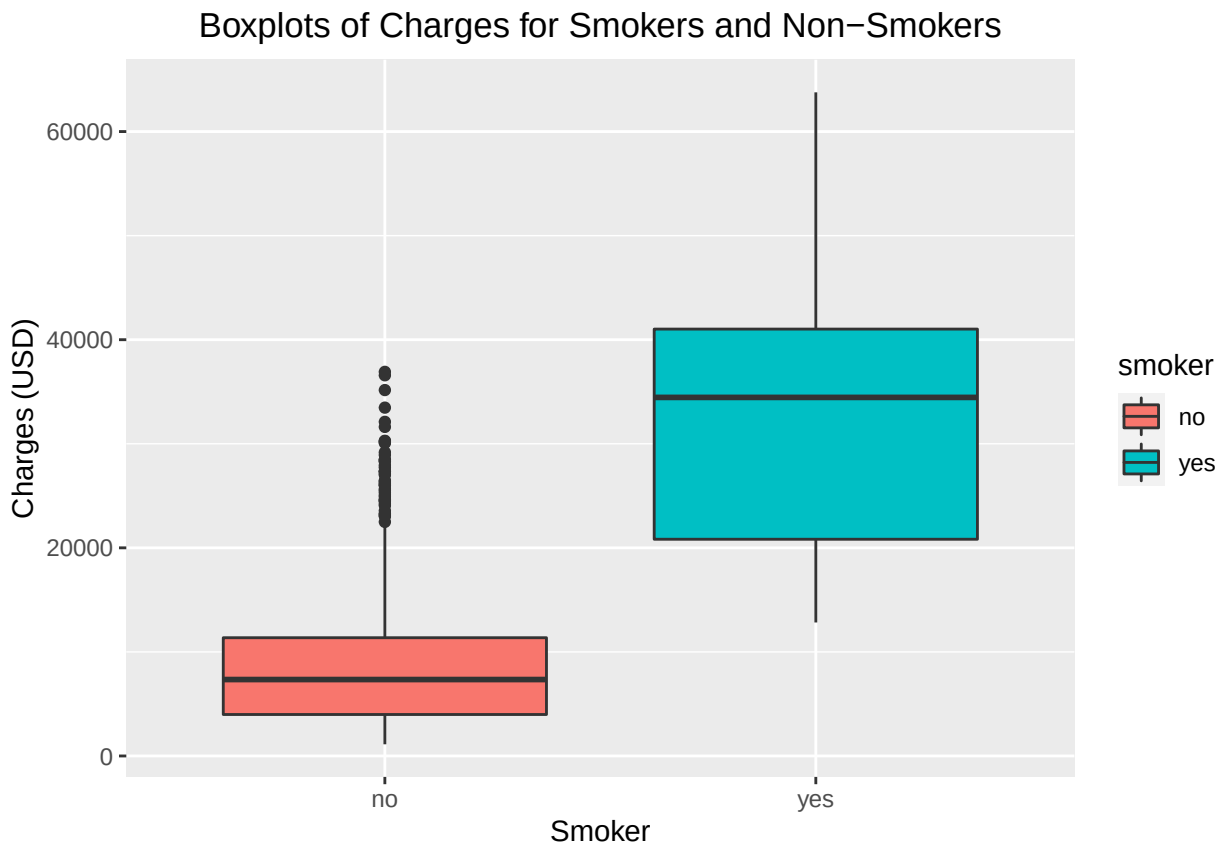


Looking at this visualization, we can begin to determine whether some of the assumptions that are required to make inferences about regression coefficients using parametric (t-statistic based) methods are being met. Firstly, we can check for linearity between each continuous predictor and the dependent variable. All scatter plots with charges as the y-variable show somewhat of a linear relationship. However, the strongest correlation seems occur when age is the x-variable (0.299), followed by BMI (0.198). Both of these correlation values were determined to be highly significant by a correlation test, which "ggpairs" computes automatically.

In addition, we can see somewhat of a linear relationship between our two continuous independent variables (age and BMI) looking at the scatter plot. Yet, the correlation between these two was very low (0.109). Generally, correlations of 0.8 and above suggest a strong relationship, in which case only one of the two variables is needed in the regression analysis. Therefore, it appears that we do not need to remove any variables from our analysis due to collinearity, although this will be assessed in further detail below.

Looking at the density curve for charges, we can also begin to predict the main effect of our categorical variable. There is a pretty clear division between the portion of the population containing non-smokers (on the left-hand side of the curve) and the smokers (on the right). This might hint at the fact that the average medical charges for non-smokers is larger than that for smokers which, as explained earlier, would make sense according to insurance companies' business models. To clarify this relationship, I also created a boxplot comparison of medical charges for smokers and non-smokers (see Figure 2). Looking at this plot, it becomes even more evident that there might be a significant difference. Both the average value and the entire interquartile range seem to be higher for smokers than non-smokers.

```
ggplot(insurance_data, aes(x = smoker, y = charges, fill = smoker)) +
  geom_boxplot() +
  labs(title = "Boxplots of Charges for Smokers and Non-Smokers",
       y = "Charges (USD)", x = "Smoker") +
  theme(plot.title = element_text(hjust = 0.5))
```



We are also interested in accounting for interactions between variables in our regression analysis. Looking back to the scatter plots in Figure 1, we can see that the linear fit line varies according to our categorical variable. When our independent continuous variable is age, the two lines seem virtually parallel, suggesting almost no interaction between predictors. However, if we use BMI as our x-axis, the lines are clearly not parallel. The slope is higher for smokers than non-smokers. Therefore, it makes sense to test for this interaction in our regression analysis.

Building the Model

To begin the regression analysis, I will look at the case where only the main effects of the prediction variables are considered.

```
lm_fit1 <- lm(charges ~ age + bmi + smoker, data = insurance_data)
summary(lm_fit1)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker, data = insurance_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12415.4  -2970.9   -980.5   1480.0  28971.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept) -11676.83     937.57  -12.45   <2e-16 ***
## age            259.55      11.93   21.75   <2e-16 ***
## bmi            322.62      27.49   11.74   <2e-16 ***
## smokeryes    23823.68     412.87   57.70   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6092 on 1334 degrees of freedom
## Multiple R-squared:  0.7475, Adjusted R-squared:  0.7469
## F-statistic:  1316 on 3 and 1334 DF,  p-value: < 2.2e-16
```

Looking at the table of coefficients (made using the "summary" function), it can be seen that the p-value of the F-statistic is $< 2.2e\text{-}16$, which is highly significant relative to a 0.05 significane level. This means that at least one of the predictor variables is significantly related to the outcome variable. Moreover, for every regression beta coefficient, the p-value of the t-statistic has a p-value of $< 2.2e\text{-}16$ which is, again, highly significant. This means that there is a significant association between the predictor and the outcome variable. That is, the beta coefficient of each predictor is significantly different from zero. Given our earlier observations, we also want to test interaction effects in addition to the main effects.

```
lm_fit2 <- lm(charges ~ age + bmi + smoker + age*bmi + age*smoker + bmi*smoker,
              data = insurance_data)
summary(lm_fit2)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + smoker + age * bmi + age *
##     smoker + bmi * smoker, data = insurance_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13940.5  -2021.5  -1327.2   -273.6  29337.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        52.787   2044.539   0.026    0.979
## age               205.842     49.608   4.149 3.55e-05 ***
## bmi               -69.678     65.865  -1.058    0.290
## smokeryes     -20126.832   1852.392 -10.865  < 2e-16 ***
## age:bmi            1.979      1.570   1.260    0.208
## age:smokeryes     -1.123     23.938  -0.047    0.963
## bmi:smokeryes   1433.743     53.411  26.843  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4908 on 1331 degrees of freedom
## Multiple R-squared:  0.8365, Adjusted R-squared:  0.8357
## F-statistic:  1135 on 6 and 1331 DF,  p-value: < 2.2e-16
```

In this case, the p-value of the F-statistic is once again $< 2.2e\text{-}16$, making the model highly significant. However, the y-intercept is no longer significant, and neither is the coefficient for BMI (with non-smokers), nor the interactions age:smokers and age:BMI, with p-values of 0.979, 0.290, 0.208, and 0.963 respectively. Since all these extra terms are only making the model more complex, I will chose to remove all of these terms in hopes for a simpler model

```
lm_fit3 <- lm(charges ~ age + smoker + bmi*smoker, data = insurance_data)
summary(lm_fit3)
```

```
##
## Call:
## lm(formula = charges ~ age + smoker + bmi * smoker, data = insurance_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14595.4  -2015.2  -1319.2   -290.5  29313.7
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -2290.008    831.999  -2.752    0.006 **
## age              266.758      9.617  27.739   <2e-16 ***
## smokeryes     -20093.508   1666.827 -12.055   <2e-16 ***
## bmi               7.109     25.058   0.284    0.777
## smokeryes:bmi   1430.920     53.217  26.888   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4907 on 1333 degrees of freedom
## Multiple R-squared:  0.8363, Adjusted R-squared:  0.8358
## F-statistic:  1702 on 4 and 1333 DF,  p-value: < 2.2e-16
```

This new model is also significant, given that the p-value of the F-statistic is $< 2.2e\text{-}16$, but it is simpler as it had fewer terms. It is worth noting that the beta coefficient for BMI in the case of non-smokers is still not significant, meaning that there might not be a significant difference between this value and zero. However, I will choose to keep the model as is given that the coefficient for BMI in the case of smokers is, in fact, statistically significant and, as we have seen in our visualization of the interaction between BMI and smoking (see again Figure 1), this interaction seems to be a real occurrence in the data.

Model Fit Assessment

In order assess which model is the best fit for our data, I will use two separate statistics: adjusted R-squared and BIC. Both of these will take into account how much of the total variance is explained by the model as well the number of predictor variables. That is, they incorporate a bias for "simpler" models.

```
BIC(lm_fit1)
```

```
## [1] 27149.83
```

```
BIC(lm_fit2)
```

```
## [1] 26590.04
```

```
BIC(lm_fit3)
```

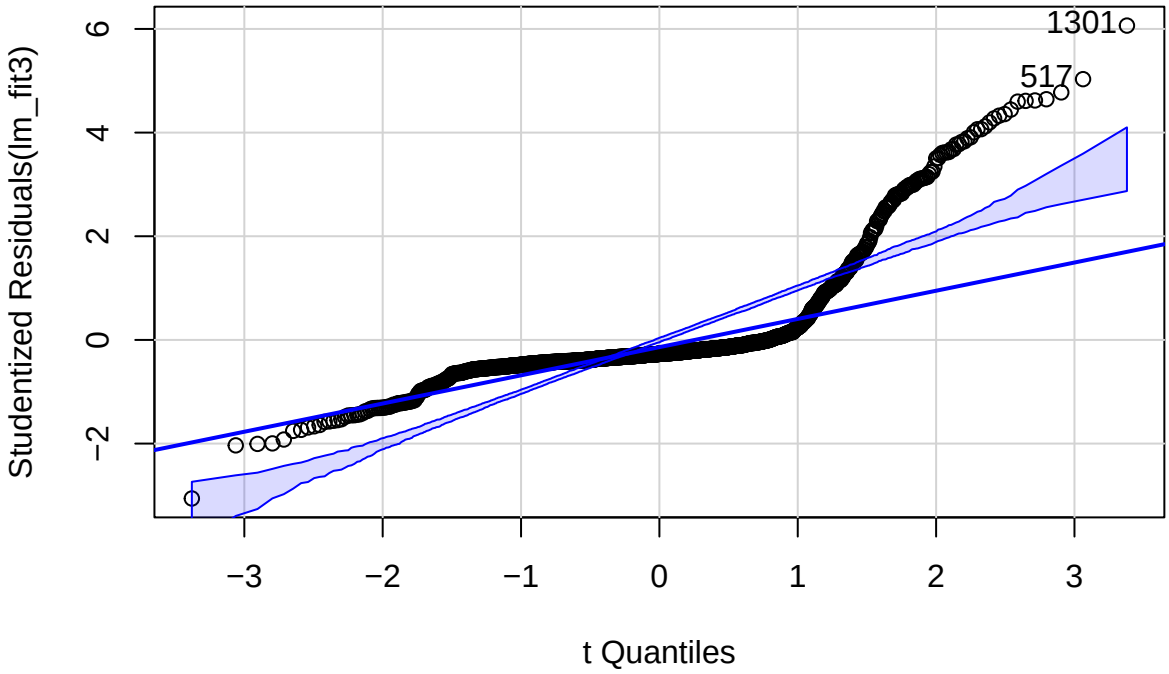
```
## [1] 26577.24
```

The adjusted R-squared values (given by the "summary" function in the last section), in order of first to last created model, are respectively 0.7469, 0.8357, 0.8358. Hence, the simpler model with BMI and smoking interaction is preferred by this metric. The BIC agrees with this assessment, with values of 27149.83, 26590.04, and 26577.24 respectively. All in all, it appears that the third model best describes our data.

Checking Parametric Assumptions

We have already discussed the linearity assumption with respect to Figure 1. Another assumption of linear regression is that of independence. Since every row in this particular data set corresponds to a separate person, with their own personal information (BMI, age, etc.), we may conclude that there is is no relation between groups. That is, one person's age will usually never impact how old another person is, and the same applies to the other variables. Another assumption of linear regression is that of normality. To asses this, I have created a Q-Q Plot between the t-quantiles and the studentized residuals.

```
qqPlot(lm_fit3)
```

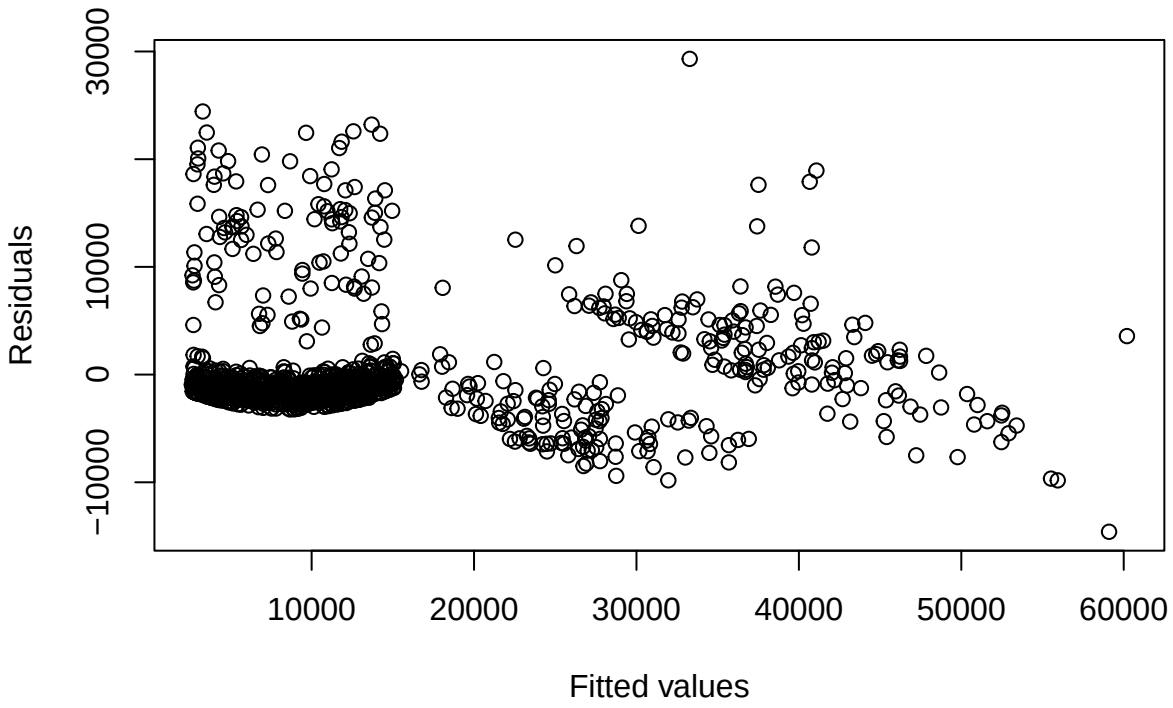



```
## [1]  517 1301
```

Looking at this plot, we do not seem to have a perfect linear relationship between the t-quantiles and the residuals. Yet, linear regression analysis is actually pretty robust for validity against nonnormality, so we may move on.

I am now interested in checking for homoscedasticity (equal variance in the residuals). To do this, I plotted a graph of the fitted values against the residuals.

```
plot(lm_fit3$fitted.values, lm_fit3$residuals,
     xlab = "Fitted values",
     ylab = "Residuals")
```



Based on this new plot, it seems as though homoscedasticity is not perfectly met. There is some left-hand side clustering in the residuals, indicating that the variance is not perfectly equal throughout. There also some indication of discrete or topologically different areas in the plot. This might be attributed to the inclusion of a categorical predictor in our model.

One last assumption of multiple linear regression is that there exist no multicollinearity. To test for this, I will use the VIF, which stands for variance inflation factor. The VIF is calculated as 1/tolerance, where tolerance is an indication of the percent of variance in the predictor that cannot be accounted for by the other predictors. A VIF score should be close to 1, but under 5 is fine and indicates no collinearity. Calculating this requires the "vif" function from the "cars" package. Note that I will using the first model as the input (the one containing only main effects) since only this will enable me to test for multicollinearity between the different groups.

```
vif(lm_fit1)
```

```
##      age      bmi   smoker
## 1.012747 1.012128 1.000669
```

The VIF values were all close to 1: 1.012747, 1.012128, and 1.000669 for age, BMI and smoking respectively. Therefore, we may conclude that there is no multicollinearity and this assumption is met.

Conclusion

In conclusion, I was indeed able to build a multiple linear model that predicts medical costs billed by health insurance given an individual's age, BMI, and whether or not they smoke tobacco. To do this, I explored three possible models: one involving only main effects, one involving main effects and all possible interactions, and a simpler model containing two main effects and one interaction. The latter was ultimately deemed the better fit using adjusted R-squared and BIC statistics.

This model accounts for an interaction between BMI and smoking. The higher the BMI, the higher the the difference between medical charges covered by insurance will be if that person smokes relative to if they did not smoke. This makes intuitive sense because people who have a high BMI and smoke have a much larger chance of needing medical assistance than a person with the same BMI who does not smoke. The presence of the two factors seems to multiply the amount risk. Therefore, I came to the conclusion that it was important for my model to reflect this relationship.

That being said, I also found that a couple assumptions of linear regression were not perfectly met by the model (normality and homoscedasticity). This indicates that there might be some improvements to be made in this analysis, perhaps including transformation of variables or further consideration of influential points. Moreover, perhaps another type of analysis is better suited for this problem such as polynomial regression analysis or elastic net regression analysis.

In terms of future directions, it might also be worthwhile to investigate the effects of the remaining variables in the dataset (location, sex, and number of children). It might be that a portion of the variance currently unaccounted for by my model would be accounted for by these three variables.

Reflection

Some things that went well with this project were my choice of a dataset, which had no missing values are required very minimal cleaning. I also believe to have been successful with some of my visualizations (the matrix comparison in Figure 1 in particular), and with my exploration of new packages and functions in R – namely the "ggpairs" function from the GGally" package and the "vif" function from the "car" package.

Something that could have gone better was constructing my actual model. One of the coefficients (corresponding to the BMI of non-smokers) was not statistically significant, which complicated the analysis because this model was a better fit for the data according to the adjusted R-squared and BIC values. There also seemed to be a clear interaction between BMI and smoking based on the visualization.

I spent approximately 9 hours working on this project.