finalproject

2024-04-07

Section 1: Introduction Provide a brief introduction of the goal of this final project. What is it all about? Where did you get the data from? What is the data framework? What are the main questions you want to answer with this data analysis?

The data comes from the Kaggle data sets. The link can be found here: https://www.kaggle.com/datasets/mirichoi0218/insurance/data

Section 2: Exploratory Data Analysis Include some graphical displays and numerical summaries of the data. Also comment on any patterns/characteristics of the data which you find interesting or anything relevant to your later analysis.

Provide a brief explanation/summary of variables you plan to include in your analysis. Here are some question you might ask:

Which variables are categorical (when applicable) and which are numerical? Should we remove any unusual observations? Should we add or remove some variables in our analysis? For categorical variables (when applicable), should we include any interactions? For numerical variables, any evidence supporting nonlinear trends?

Section 2: Exploratory Data Analysis

Here is a brief overview of all of the variables of this dataset:

Numerical Variables:

- age: age of primary beneficiary
- bmi: body mass index
- children: number of children/dependents covered by health insurance
- charges (response variable): medical costs billed by health insurance

Categorical Variables:

- sex: male or female
- smoker: yes or no
- region: residential area in the US, can be northeast, southeast, southwest, or northwest

The first step is to check for unusual observations by performing an outlier test. To do so, we obtained the studentized residuals for our data, and compared the largest values with the Bonferroni critical value.

This is the Bonferroni critical value we calculated:

```
## [1] -4.137174
```

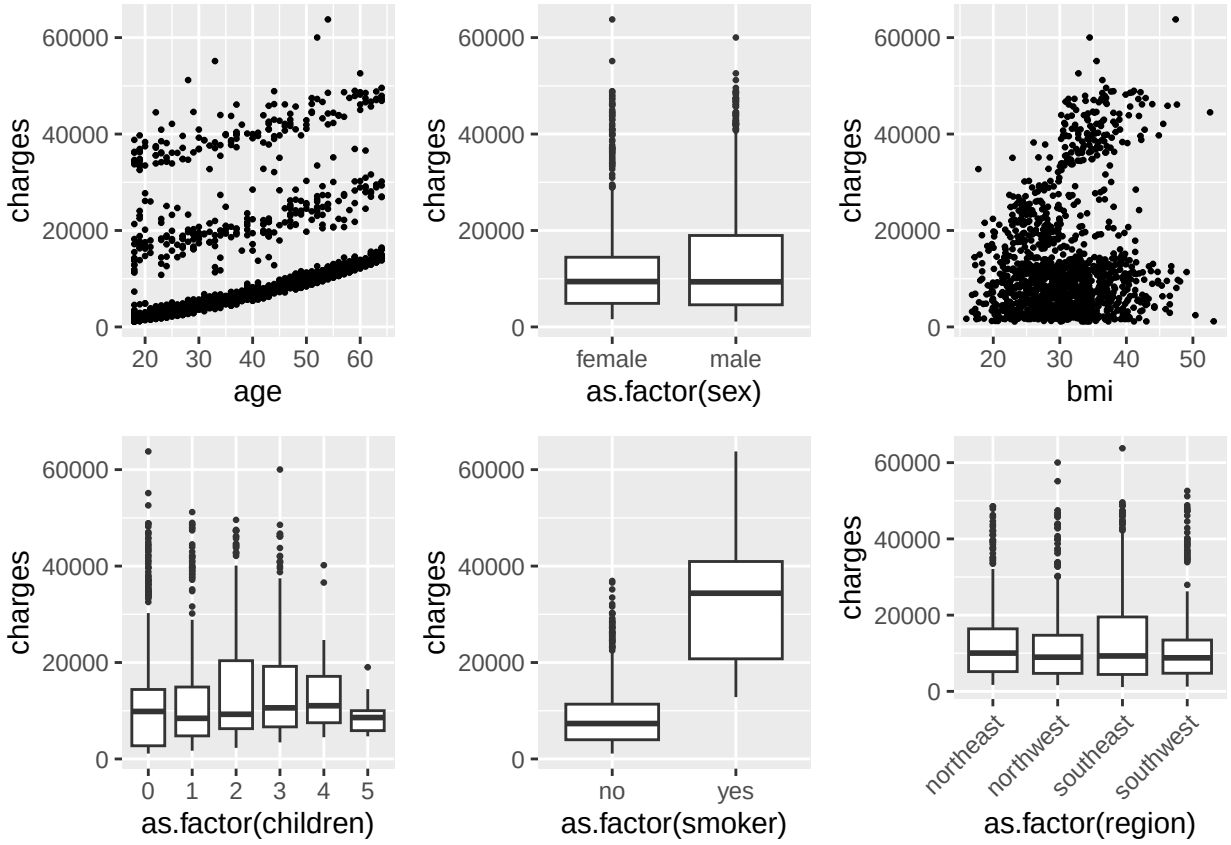
Next, here are the five largest studentized residuals, which we will be comparing to the Bonferroni critical value:

```
##     1301      578      243      220      517
## 5.009599 4.219800 4.053228 3.998326 3.863878
```

Since the absolute value of their respective studentized residual is greater than the Bonferroni critical value, we can conclude that Observation #1301 and Observation #578 are outliers.

We also checked Cook's distance, however we found that there is no point with Cook's distance greater than 1. We will simply remove the two outliers we found earlier.

After removing these two points, we will be analyzing the predictors to see if there are any variables we should add or remove.



From the plot of charges vs age, we notice a trend that the charges increase as age increases. We will not be removing 'age' as a predictor since it is clear that it has a significant impact on the response variable 'charges'. This makes sense because as people grow older, their health declines, and the insurance costs increase.

The charges vs sex boxplot tells us that males tend to have higher charges than females. We will conduct a t-test to test this difference.

We obtained this p-value from the t-test:

```
## $p.value
## [1] 0.03460436
```

Given that the p-value is less than 0.05, we conclude that 'sex' is a significant predictor, with males incurring higher insurance costs than females. A possible explanation for this is that males are usually more at risk of health conditions that result in increased charges compared to females.

From the plot of charges vs bmi, we can see a general trend that as bmi increases, the charges also increase. As a result, we will not be removing 'bmi' as a predictor since it appears to have an impact on the response variable. The graph also makes sense logically since people with a higher bmi tend to be overweight and subsequently, have worse health than people with a lower bmi, causing their insurance charges to be greater as well.

The boxplot of charges vs children conveys to us that charges tend to increase as the number of children grows from 0 or 1 to 2 or 3. While there are some data points for people with 4 or 5 children, it appears that these groups have a very low amount of people, and it may be challenging to draw meaningful conclusions from it. We perform an ANOVA test to test the significance of 'children' as a predictor.

This is the p-value we obtain:

```
## [1] 0.008620164
```

This p-value is less than 0.05, thus signifying that the predictor 'children' is significant. We will be keeping as a predictor in our model. A larger number of children/dependents covered leads to higher charges, as there are more individuals who need to be covered.

From the boxplot of charges vs smoker, it is evident that smokers generally face significantly higher charges than non-smokers. Consequently, we will not be removing 'smoker' as a predictor. People who smoke typically incur higher insurance costs due to the health risks posed by smoking.

The boxplot of charges vs region shows us that the southeast region seemingly has higher charges than the other three regions. We will conduct an ANOVA test to make sure that the predictor 'region' is significant.

This is the p-value we obtained from the ANOVA test:

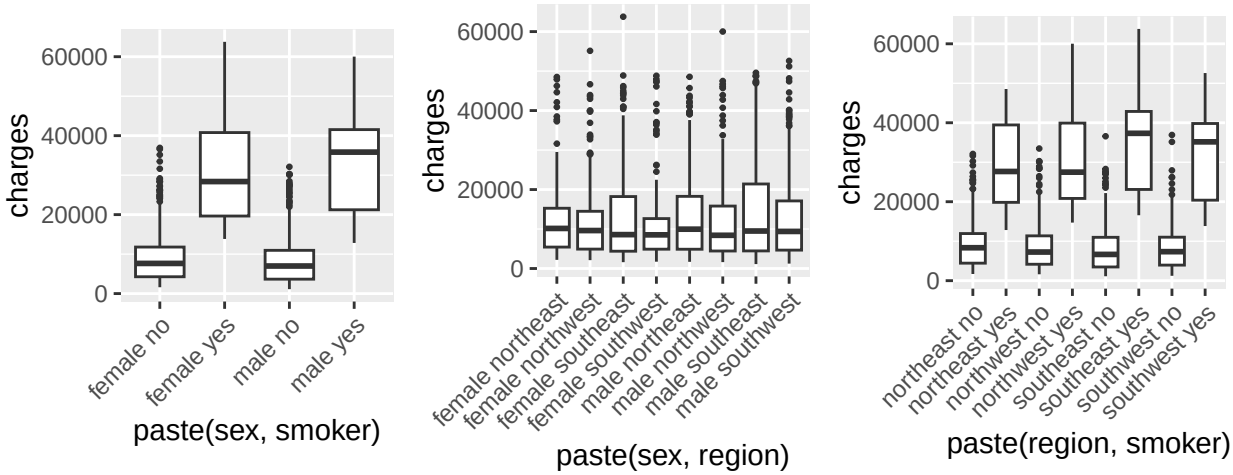
```
## [1] 0.04525914
```

The p-value is less than 0.05, indicating that the predictor 'region' is significant, and thus we will retain it as a predictor. This confirms that there are regional differences in the insurance costs.

Now, we will explore whether or not we should include any interactions between categorical variables. We will be analyzing all possible first-order and second-order interactions between our categorical predictors, including:

- 'sex' and 'smoker'
- 'sex' and 'region'
- 'region' and 'smoker'
- 'sex', 'smoker', and 'region'

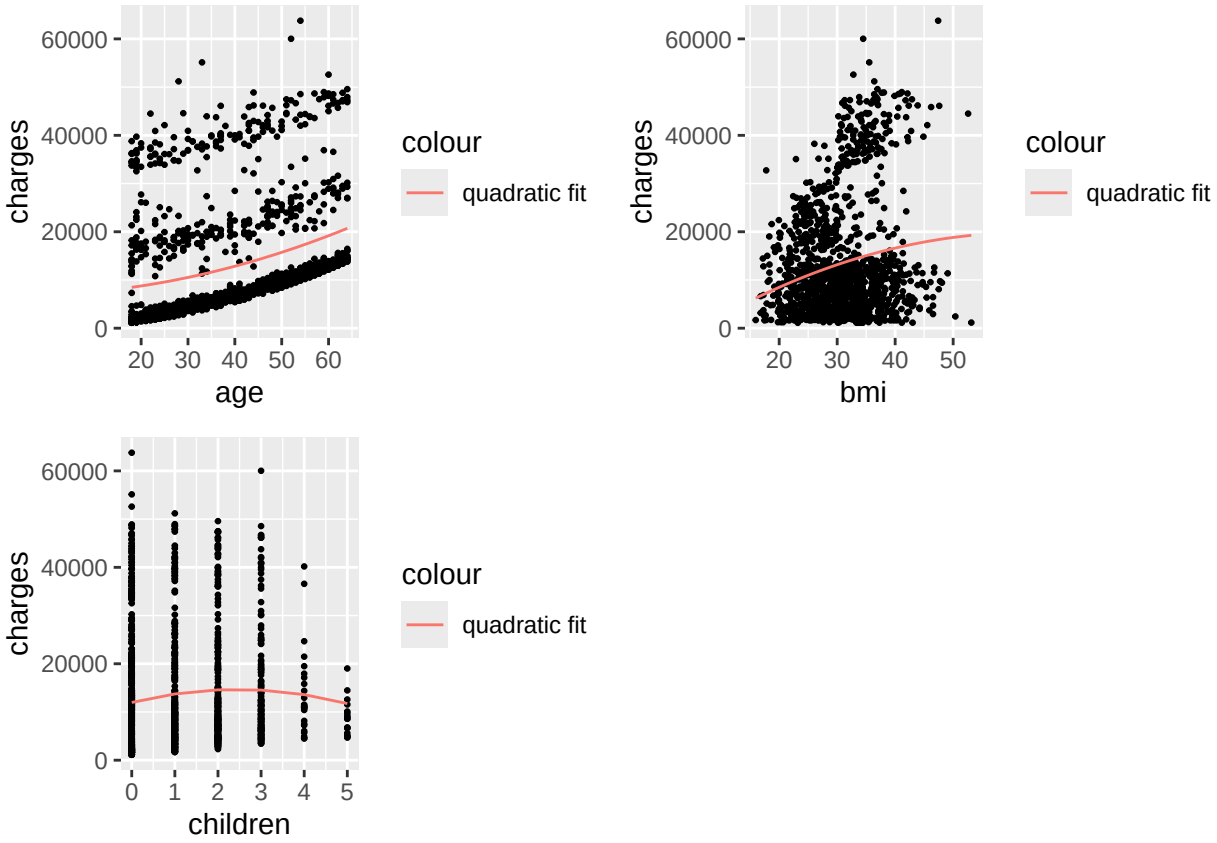
Here are some plots that show the interactions between these predictors.



From the plots, it appears as though smoking increases the charges more for males than females. Additionally, it seems like smoking has a greater impact on charges in the southeast and southwest, when compared to northeast and northwest. However, it does not appear that sex and region is a significant interaction.

We applied sequential F-tests with the anova function to assess the significance of the interaction terms in our model. Initially, we included all potential interactions among the three variables 'sex', 'smoker', and 'region'. We first found that the second-order interaction sex:smoker:region was not significant, and removed it from the model. Subsequently, we tested the first-order interactions and found that the interaction sex:region was not significant. Finally, after removing sex:region as a predictor, we concluded that only the interactions sex:smoker and smoker:region are statistically significant, and we will retain them in the final model.

Now, we will examine the numerical variables and see if there is evidence of any nonlinear trends.



Based on the plot of the quadratic fit for charges vs age, we see there is evidence of non-linearity for the predictor 'age' since the quadratic fit has some upward curvature.

Similarly, based on the plot of the quadratic fit for charges vs bmi, we see there is some evidence of non-linearity for the predictor 'bmi', although maybe not as clear as for 'age'. This time, the quadratic fit curves downwards.

Finally, the plot of charges vs children also suggests some non-linearity. However, the non-linearity is not as evident as for 'age and 'bmi' since although the quadratic fit does appear to curve downwards as the number of children increases, there is not many data points for 4 and 5 children, so we may not be able to draw a strong conclusion.

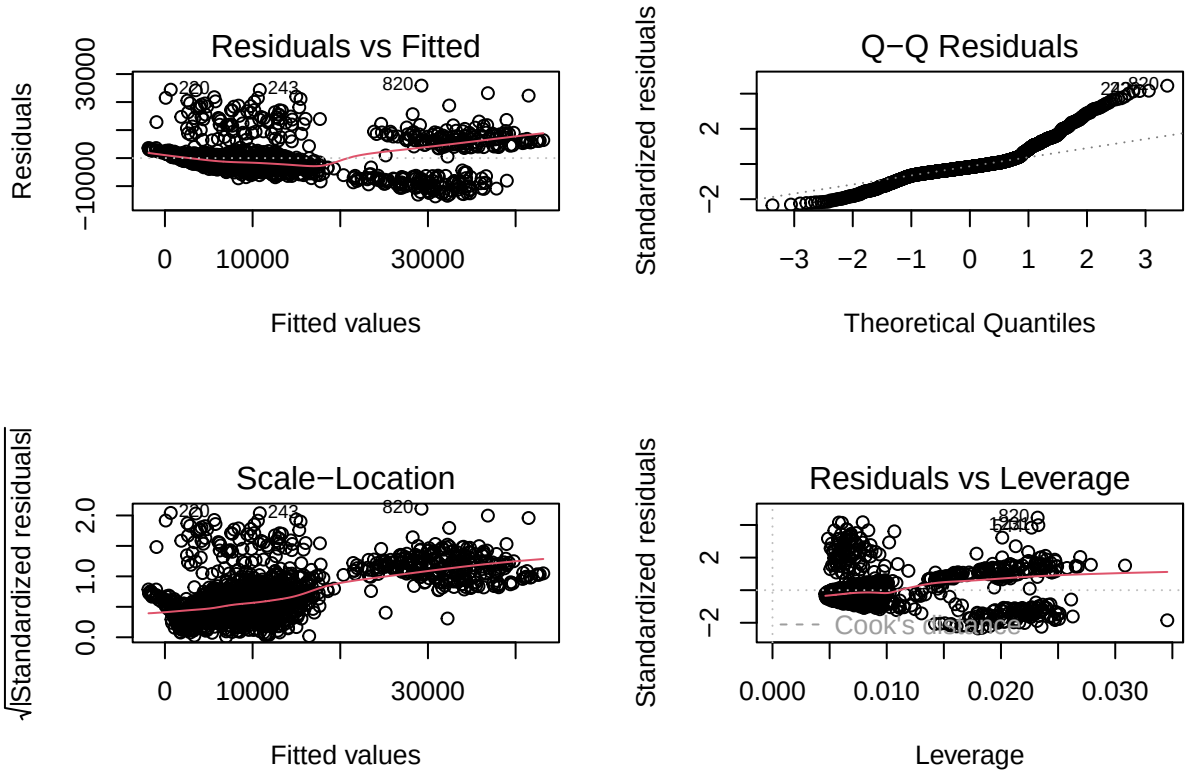
Section 3: Methodology

Section 3.1: Simple Model Model Selection The choice of linear regression for this analysis is based on its efficacy in providing a clear and straightforward interpretation of how various predictors affect insurance charges. Linear regression is particularly valued for its ability to establish a baseline understanding of the relationships between variables, offering insights into the direct impacts of age, BMI, children count, smoking status,sex and regional differences on insurance costs. Variable Selection and Diagnostics Age: A continuous variable expected to positively correlate with insurance charges. BMI: Another continuous variable, which is hypothesized to influence insurance costs due to its association with health risks. Children: A discrete variable that represents the number of dependents,potentially affecting insurance premiums. Smoker Status: A categorical variable that significantly impacts insurance costs, as identified in the exploratory analysis. Sex:A categorical variable, often considered in insurance cost analysis because gender may influence health risks and insurance premiums. Region: As a categorical variable, the region captures variations in insurance costs across different geographical locations. Sex and Smoker Interaction: This interaction term examines how gender and smoking status jointly influence insurance costs. Smokers of different genders may face different health risks, which can be reflected in their insurance premiums. Region and Smoker Interaction: This interaction term was included to examine if the impact of smoking on insurance charges varies by region, which could suggest regional variations in healthcare costs or lifestyle patterns.

```
##              GVIF Df GVIF^(1/(2*Df))
## age      1.017141  1        1.008534
## bmi      1.108041  1        1.052635
## children 1.004025  1        1.002011
## smoker   1.011905  1        1.005935
## sex      1.009150  1        1.004565
## region   1.099632  3        1.015955
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + sex +
##     region + sex:smoker + smoker:region, data = insurance2)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -13669  -2817  -1035   1269  25874
##
## Coefficients:
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)               -10872.78     980.81 -11.085  < 2e-16 ***
## age                          256.73      11.57  22.197  < 2e-16 ***
## bmi                          329.46      27.84  11.836  < 2e-16 ***
## children                     493.41     133.85   3.686 0.000237 ***
## smokeryes                  19484.22     931.12  20.926  < 2e-16 ***
## sexmale                     -600.93     360.92  -1.665 0.096153 .
## regionnorthwest             -506.43     514.05  -0.985 0.324713
## regionsoutheast            -2272.77     524.14  -4.336 1.56e-05 ***
## regionsouthwest            -1748.10     515.09  -3.394 0.000710 ***
## smokeryes:sexmale           2163.42     812.25   2.663 0.007827 **
## smokeryes:regionnorthwest    978.98    1178.49   0.831 0.406291
## smokeryes:regionsoutheast   5661.02    1082.02   5.232 1.95e-07 ***
## smokeryes:regionsouthwest   4329.60    1178.69   3.673 0.000249 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5881 on 1323 degrees of freedom
```

```
## Multiple R-squared:  0.7612, Adjusted R-squared:  0.759
## F-statistic: 351.4 on 12 and 1323 DF,  p-value: < 2.2e-16
```



Model Diagnostics The diagnostics conducted include: Multicollinearity Check: The Variance Inflation Factor (VIF) was calculated for all predictors, with results indicating minimal multicollinearity among the variables used. VIF values were as follows: age (1.00), BMI(1.05), children (1.00), smoker status (1.01), sex (1.00) and region(1.02). Model Fit: The linear model was fitted with charges as the dependent variable. The R-squared value of 0.7612 suggests that about 76.12% of the variability in insurance charges is explained by the model, which is quite substantial for a simple linear model.

Residual Analysis: Residuals: The residuals' plot indicated that while there is some spread, major deviations from normality or constant variance (homoscedasticity) were not apparent, suggesting that the model assumptions are reasonably met.

Coefficients: Intercept: The model's intercept is significantly negative, which might indicate that the baseline insurance charges are negative when all other variables are zero. This could suggest that further adjustments to the model might be necessary.

Age: The coefficient for age is significant and positive, indicating that insurance charges increase with age. This aligns with the general logic of insurance pricing, where increased age is associated with higher health risks.

BMI: The coefficient for BMI is also significantly positive, suggesting that higher BMI is associated with higher insurance charges, likely due to increased health risks associated with higher BMI.

Children: The coefficient for the number of children is significant and positive, indicating that having more children leads to higher insurance charges, possibly reflecting the greater healthcare needs of larger families.

Significant Impact of Smoking Status: The coefficient for smokers is significantly higher than for non-smokers, consistent with previous studies indicating that smoking is a significant factor affecting insurance costs.

Interactions of Region and Gender: Region Coefficients: Some regional categories, such as southeast and southwest, are significant, suggesting that insurance costs in these regions are higher compared to the baseline region.

Interaction of Smoking Status and Gender: This interaction term is significant, indicating that the combination of smoking status and gender has a specific impact on insurance charges, particularly for male smokers who may face higher charges.

These conclusions suggest that while the overall model performs well, the impact of certain variables like regions and interaction terms may require more complex models to more finely interpret these factors' effects on insurance costs. Further model validation and adjustments are also necessary to ensure the model's generalization capability and predictive accuracy across different datasets.

Section 3.2: Predictions from Simple Model Model Implementation The linear regression model was implemented using the lm() function in R, focusing on charges as the dependent variable. This model incorporates several predictors, including age, BMI, number of children, smoker status, and the interaction between smoker status and region, which were identified as significant in influencing insurance costs during the exploratory analysis.

Testing and Predictions To validate the effectiveness of the linear regression model, the dataset was split into training and testing subsets. Specifically, 80% of the data was used for training the model, and the remaining 20% served as the testing set. This division ensures that the model is tested on unseen data, providing a fair assessment of its predictive accuracy.

```
##       Actual Predicted
## 5   3866.855  5801.680
## 9   6406.411  8756.868
## 14 11090.718 14035.788
## 19 10602.385 14611.437
## 25  6203.902  7658.894
## 28 12268.632 14258.066
```

Predictive Performance Evaluation The model was fitted on the training data and then used to make predictions on the testing set. Here are the results of the predictions for the first few observations:

As observed, the model provides reasonable estimates for the charges, although there are noticeable differences between the actual and predicted values, suggesting areas for improvement.

The initial testing indicates that while the linear regression model captures general trends and is able to provide ballpark estimates of insurance charges, discrepancies remain between predicted and actual charges. These differences highlight the potential need for more complex models that could better account for nuances in the data not captured by this simple linear model. The following sections will explore such models, comparing their predictions with those of the baseline linear model to identify the most effective approach for predicting insurance charges.

Section 3.3: Advanced Models Model Types: Regularized Regression To enhance the predictive accuracy and manage potential issues of multicollinearity and overfitting present in our linear regression model, we employed advanced modeling techniques such as Ridge and Lasso regression. These methods are forms of regularized regression that include a penalty term to the loss function:

Ridge Regression (L2 regularization): Adds a penalty equal to the square of the magnitude of coefficients. This method is particularly effective in reducing the model complexity while still allowing the use of all variables. It helps in handling multicollinearity, preventing overfitting by shrinking the coefficients. Lasso Regression (L1 regularization): Adds a penalty equivalent to the absolute value of the magnitude of coefficients. Unlike Ridge, Lasso can completely eliminate the weight of less important variables by setting their coefficients to zero. This results in feature selection within the model, which is beneficial when we need a sparse model with fewer variables.

Implementation and Predictions Both models were implemented using the glmnet package in R, which efficiently handles large datasets and complex models with its capabilities for both Lasso and Ridge regression. The matrix of predictors x was derived from the training data, excluding the intercept to allow glmnet to handle regularization properly. The regularization parameter s was set to 0.01 for both models. This choice

is typically subject to tuning through methods like cross-validation, but for simplicity, we've chosen a small but non-zero value, which demonstrates regularization without overly constraining the models.

Predictive Performance Evaluation The predictions made by the Ridge and Lasso models on the testing set were compared with the actual charges and those predicted by the simple linear model.

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-8
```

```
##       Actual    Linear     Ridge     Lasso
## 5   3866.855  5801.680  5976.579  5894.958
## 9   6406.411  8756.868  8931.156  8638.090
## 14 11090.718 14035.788 14047.500 14047.158
## 19 10602.385 14611.437 14110.493 14658.892
## 25  6203.902  7658.894  7695.461  7725.347
## 28 12268.632 14258.066 14087.683 14248.724
```

Comparison and Conclusions The Ridge and Lasso models both adjusted the predictions closer to actual values in some cases, compared to the basic linear model. However, they also introduced some variability, likely due to the influence of the regularization term balancing bias and variance differently:

Ridge Model: It generally increased the estimates, potentially due to its nature of shrinking coefficients but not setting them to zero, thus still considering all input features.

Lasso Model: It showed a trend similar to the linear model but with slight adjustments, indicating that it might be dropping some less influential variables from the model.

These results suggest that while both Ridge and Lasso offer advantages in terms of handling overfitting and multicollinearity, the choice between them would depend on the specific characteristics of the dataset and the goal of the analysis, with Lasso providing a sparser solution and Ridge offering stability across a broader range of data conditions. Further model tuning and cross-validation would be necessary to optimize their performances and fully assess their effectiveness compared to the simple linear model.

##Section 4: Discussion and Conclusions Summary of Results In this analysis, we employed various statistical models to predict health insurance costs using a dataset that included variables such as age, BMI, number of children, smoker status, sex and regional interactions. The results can be summarized as follows:

Linear Regression Model: Provided a solid baseline for understanding how different variables impact insurance costs. Significant predictors included age, BMI, and especially smoker status, which greatly increased predicted charges. However, the model sometimes overestimated or underestimated charges significantly. Ridge Regression Model: Adjusted predictions closer to actual values in some instances, suggesting its effectiveness in handling multicollinearity by shrinking coefficients but considering all variables. Lasso Regression Model: Offered predictions similar to the linear model but with adjustments likely due to its feature selection capability, which dropped less influential predictors.

For the predictions, we apply to the test data set all the transformations made to the train data set. For each model, the predictions are obtained using the same command predict, and we evaluate the residual sum of squares, RSS.

```
##               Model         RSS
## 1 Linear Regression 10162756224
## 2  Ridge Regression 10163886527
## 3  Lasso Regression 10159739488
```

Comparative analysis showed that while the linear model was useful for understanding direct relationships and setting a baseline, both Ridge and Lasso provided nuanced insights by addressing overfitting and highlighting essential predictors, but the Lasso regression model performs the best on the testing dataset.However, specific model selection should also consider other factors such as model interpretability, computational complexity, and so on.

Impact of Analysis This analysis significantly enhances our understanding of factors influencing health insurance costs:

Smoker Status: Emerged as a critical determinant of insurance costs, with smokers incurring significantly higher charges. This highlights the potential for insurance companies to adjust premiums or offer programs encouraging smoking cessation.

Age and BMI: Both variables were also strong predictors of costs, aligning with expected health risk increases with age and higher BMI.

Regional Differences: The interaction terms suggested that regional factors might also play a role in insurance costs, potentially due to environmental, policy, or lifestyle differences.

Conclusions Smoking status is the most potent predictor of health insurance charges, suggesting targeted health initiatives could be beneficial. Age and BMI are important factors in predicting insurance costs, indicating that personalized insurance plans could be more effective. Regularized regression models (Ridge and Lasso) are valuable for refining predictions and identifying key predictors, particularly in datasets prone to multicollinearity.