Insurance Charges Prediction

Luo Chen, Huu Nguyen, Hyeongwan Gwak

Agenda

1. Research Question
2. Introduction
3. Data
4. Methods
5. Results
6. Summary

Research Question

Which predictors have a significant influence on the Insurance Premium?

Introduction and Data Description and Variables

Insurance Premium Data

- Data Source: Insurance Premium Data | Kaggle
- Dataset Description: Various individual medical costs bill by health insurance under various individual personal information
- Initial Observation: This data contained total of 9,366 observations (1,339 rows) and 7 attributes (7 columns)

Dependent Variable:

- Charges: Individual medical costs bill by health insurance (Quantitative)

Quantitative Predictors:

- Age: Age of primary beneficiary
- BMI: Body mass index, providing an understanding of the body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
- Children: Number of children covered by health insurance / Number of dependents

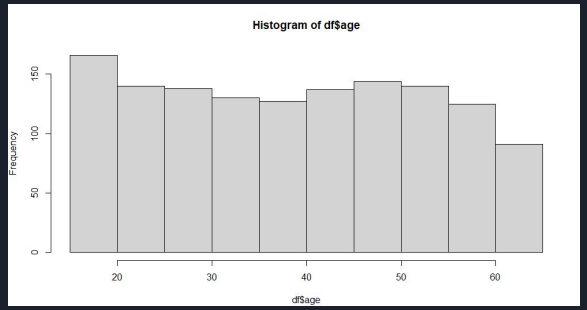
Qualitative Predictors:

- Sex: Insurance contractor gender (Male or Female)
- Smoker: Smoking Status (Yes or No)
- Region: The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.

Data: Table 1: Variables In The Dataset

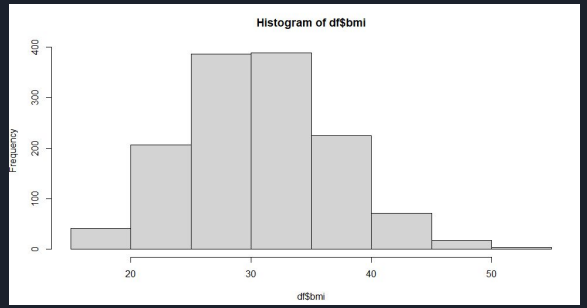
Name	Type	Description
Age	Integral (18 - 64)	Age of primary beneficiary
Sex	Binary	Insurance contractor gender, female, male
BMI	Integral (15.96 - 53.13)	Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9
Children	Integral (0 - 5)	Number of children covered by health insurance / Number of dependents
Smoker	Binary	Smoking; Yes or No
Region	Character	The beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
Charges	Integral (1,122 - 63,770)	Individual medical costs billed by health insurance

Data: Histogram Figures



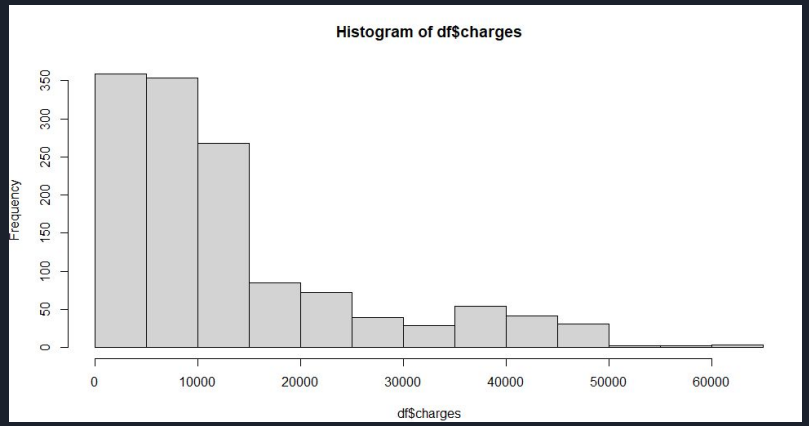
Frequency of Age Histogram

- The age histogram chart has very even frequencies of age in this dataset.



Frequency of BMI Histogram

- Most people do not have an ideal BMI (18.5 to 24.9).



Frequency of Individual Medical Cost Bill Histogram

- The majority of the Individual medical costs billed by health insurance in the dataset are below 15,000 dollars

Methods used to determine the association between variable

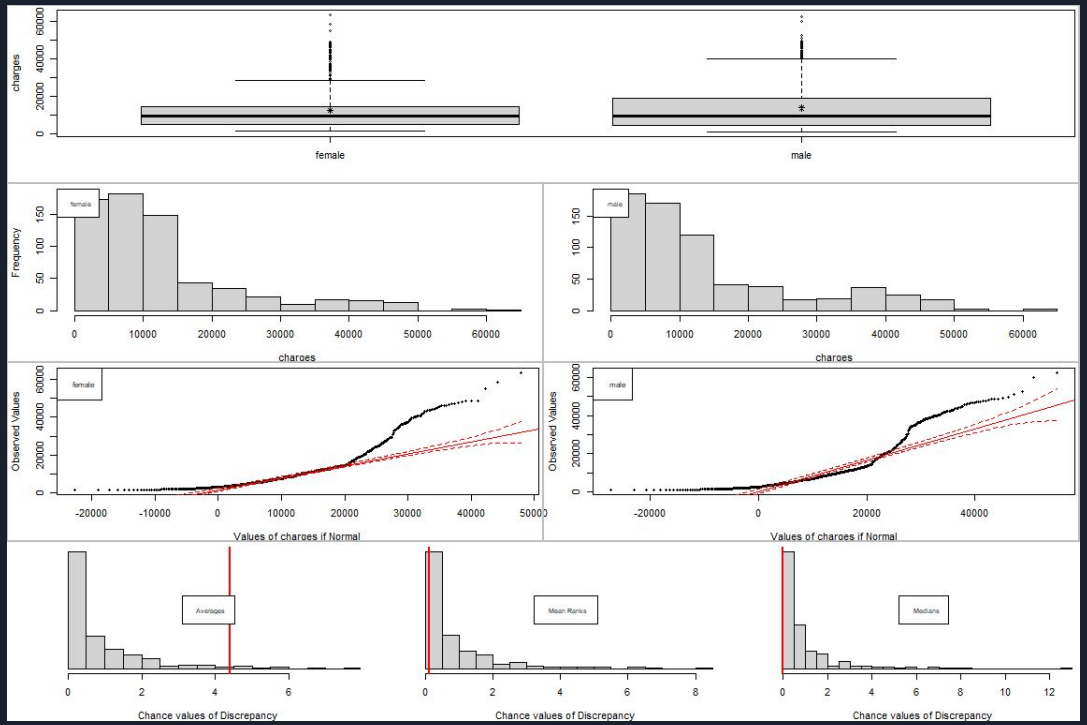
- Three main approaches:

1. Association Between Categorical and Quantitative Variables
2. Association Between Two Quantitative Variables
3. Linear Regression Models Interpretation

- The process will involve using ANOVA test, p-value, box plots, scatterplots, linear regression with Spearman's correlation.

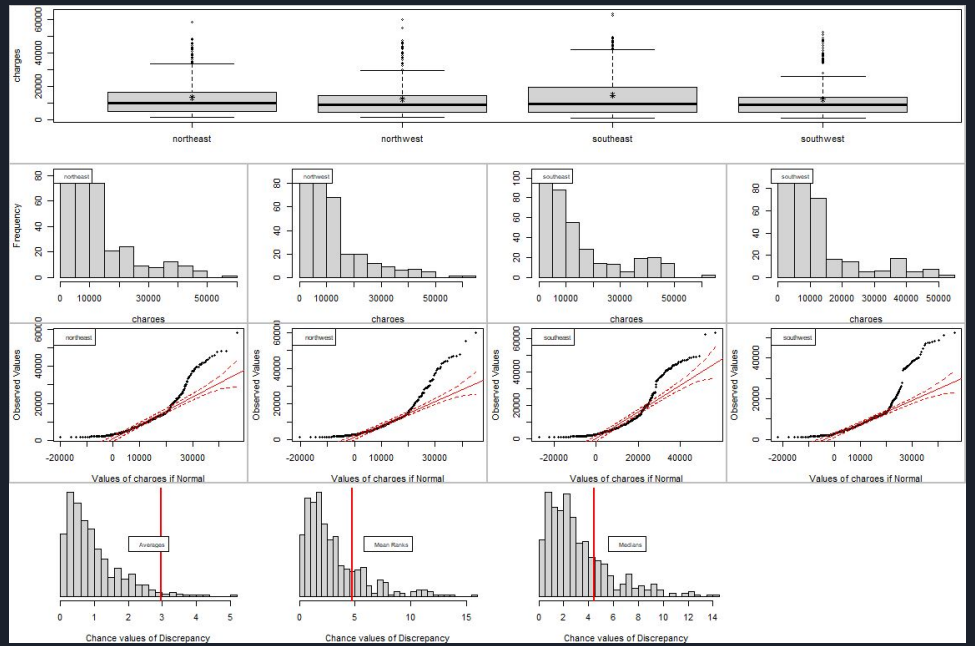
Methods used to determine the association between variable (cont.)

1. Association Between Categorical and Quantitative Variables
- Will run model to test if there is a correlation between “sex”; “smoker”; “age” with “charges”
- Test with p-value of ANOVA and look at ggplot, histogram for distribution
- Median for p-value to determine if the association is statistically significant or not



Methods used to determine the association between variable (cont.)

2. Association Between Two Quantitative Variables
   - Run test to check the association between “age”, “BMI”, “children” vs. “charges”
   - Look for trend and pattern on the scatterplot to have better understand of the relationship
   - Based on the trend result, we will use Spearman’s correlation to see if the p-value is between 0 and 0.007



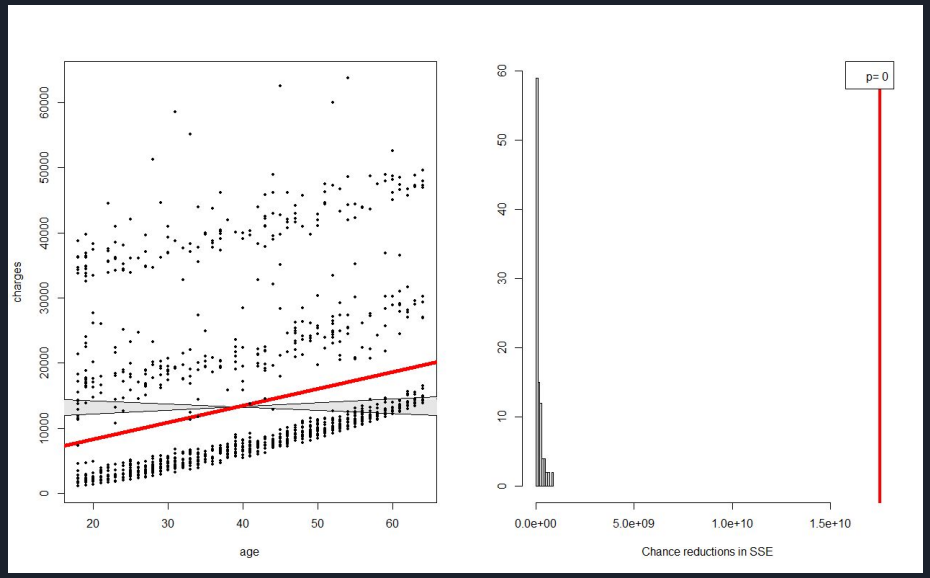
Methods used to determine the association between variable (cont.)

3.3) Linear Regression Models Interpretation

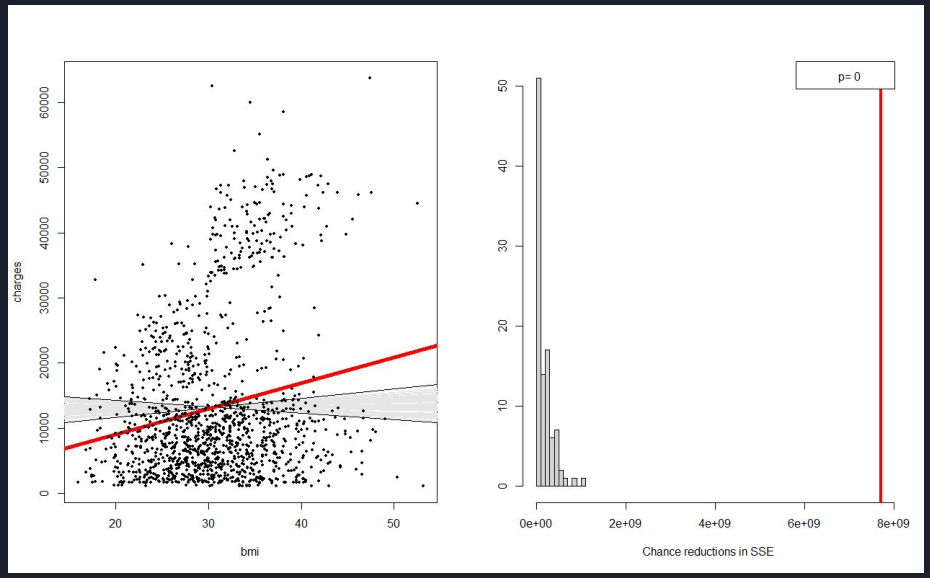
- Only work between quantitative variables
- Utilize lm() function and summary() function to indicate the model fitting performance
- Additional information to look at : RMSE and R^2 when compared to "charges" variable

➢ Draw conclusion on which variable has the strongest association with target variable "charges" and reveal other information

Methods used to determine the association between variable (cont.)



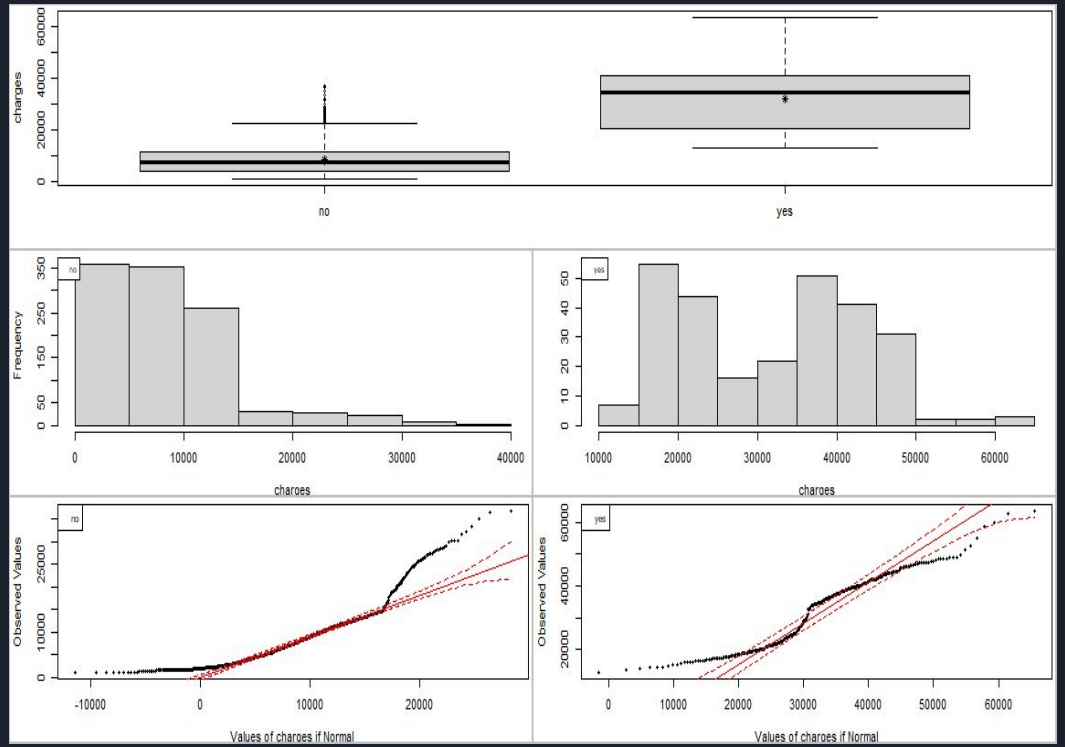
"Possible Regression Age"



"Possible Regression BMI"

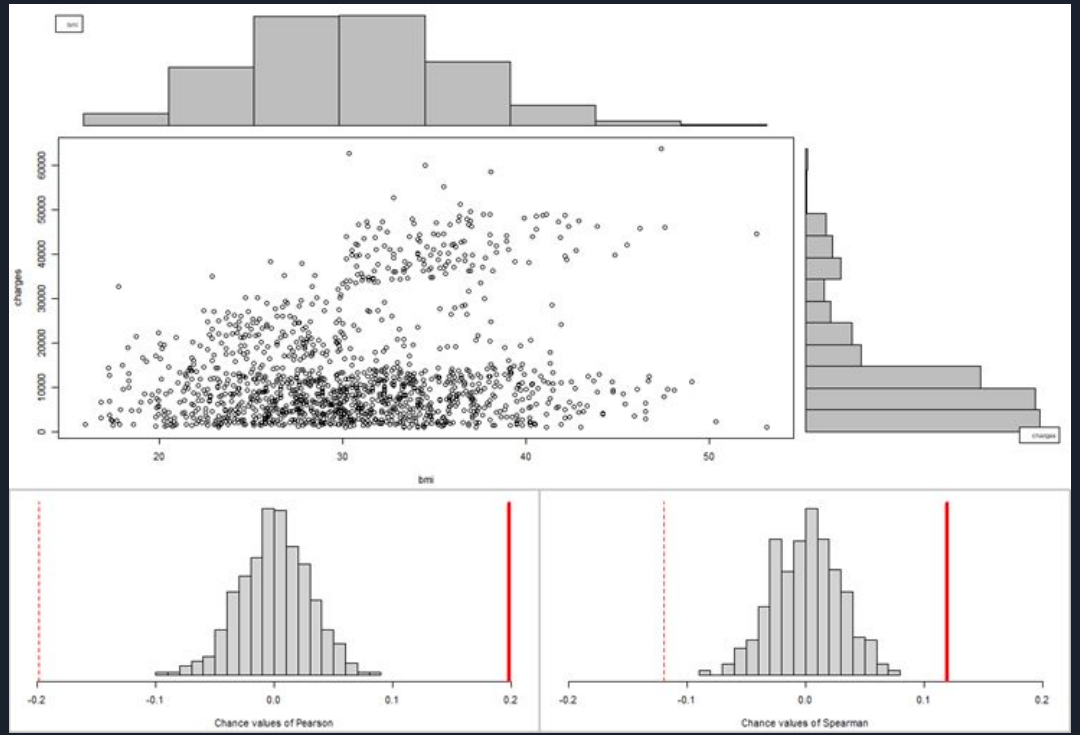
Results - (1) Association Between Categorical and Quantitative Variable

- Associate test performed on three categorical variables:
  - Sex, Smoker, Region
- Results Evaluated Based on P-Value
- The “Smoker” variable has a statistically significant correlation with the “Charges” variable (target)
  - **p-value** of median test is between 0 and 0.007
  - Sex: Between 0.19 and 0.265
  - Region: Between 0.929 and 0.969



Results - (2) Association Between Two Quantitative Variables

- Associate test performed on three quantitative variables:
  - Age, BMI, Children
- Results Evaluated Based on P-Value
- All three variables show a statistically significant correlation with the “Charges” variable
  - Age: Spearman’s correlation
  - BMI: Pearson’s correlation
  - Children: Spearman’s correlation



Results - (3) Linear Regression Model Interpretation

- Linear Regression Model performed on three quantitative variables:
  - Age, BMI, Children
- Results Evaluated Based on Slope, RMSE & R^2
- All three variables show a statistically significant correlation with the “Charges” variable
  - Age: RMSE = 11560, R^2 = 0.08941
  - BMI: RMSE = 11870, R^2 = 0.03934
  - Children: RMSE = 12090, R^2 = 0.004624

Charges vs. Age

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3165.9       937.1   3.378 0.000751 ***
age            257.7        22.5  11.453  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11560 on 1336 degrees of freedom
Multiple R-squared:  0.08941,    Adjusted R-squared:  0.08872
F-statistic: 131.2 on 1 and 1336 DF,  p-value: < 2.2e-16
```

Charges vs. BMI

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1192.94    1664.80   0.717    0.474
bmi           393.87      53.25   7.397 2.46e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 11870 on 1336 degrees of freedom
Multiple R-squared:  0.03934,    Adjusted R-squared:  0.03862
F-statistic: 54.71 on 1 and 1336 DF,  p-value: 2.459e-13
```

Charges vs. Children

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  12522.5       446.5  28.049   <2e-16 ***
children       683.1       274.2   2.491   0.0129 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12090 on 1336 degrees of freedom
Multiple R-squared:  0.004624,   Adjusted R-squared:  0.003879
F-statistic: 6.206 on 1 and 1336 DF,  p-value: 0.01285
```

Summary

- Associate Test:
  - Categorical: “Smoker” variable has a statistically significant correlation with the “Charges” variable
  - Quantitative: All three quantitative variables have a statistically significant correlation with the “Charges” variable
- Linear Regression:
  - “Age” variable performs best in the linear model fitting test, it has the highest R^2 value, and the lowest RMSE value
- Future Work:
  - Explore the rationale behind the behavior of the quantitative variables having a significant correlation with the target variable in the associate test, but bad performance in predicting the data points in the linear model
  - Change the “Children” variable to be a categorical variable instead of a quantitative variable

Thank You :)