Jae Yoon Cho
Anuv Gupta
Eric Liu
Angelina Rubio
Liaohan Wang

Health Insurance in the US

Introduction

In today's society, insurance is struggling to adapt and benefit from new technologies in comparison to other industries. However, despite the several challenges faced by insurance businesses, emergent technologies like AI and BlockChain have brought a radical change in insurance. Data Analytics, in particular, sits at the core of this transformation. 4 key factors behind the emergence of Analytics include Big Data, AI, Real Time Processing, and Increased Computing Power. Our dataset involved studying the US Health Insurance Dataset provided by Kaggle. In this study, a predictive model of the insurance data was constructed to study how various demographic variables, smoking, and medical costs affect the insurance premium. We modeled this dataset by noting the predictor and response variables, using R code to create several plots, and transformed certain variables to better fit our data set. An analysis of the final transformed model revealed that age, BMI and number of children were the most influential variables for an individual's cost of insurance.

Data Description

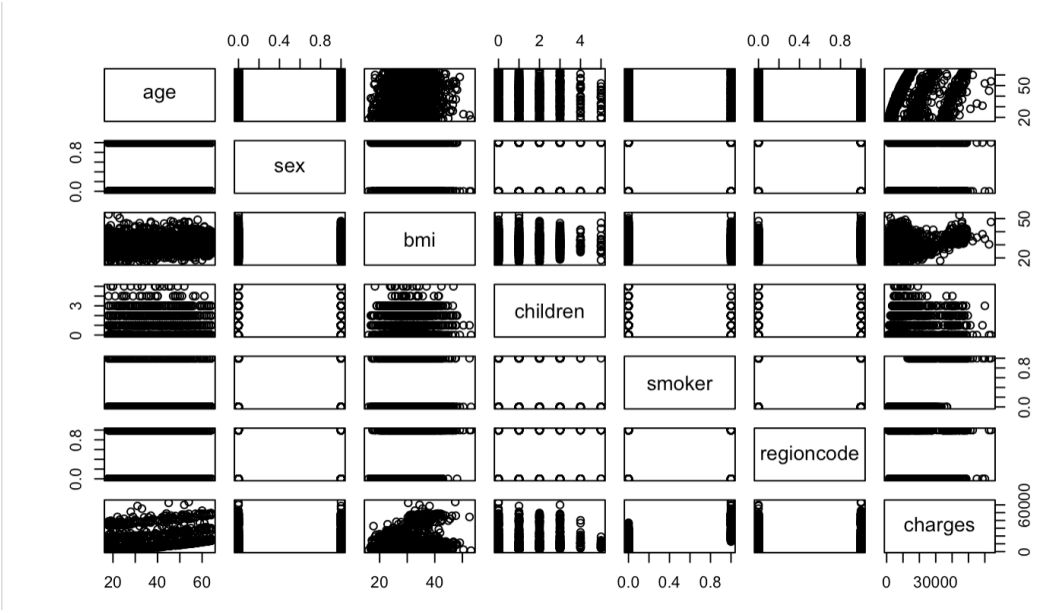
The original dataset contains 7 variables

1. Age (Predictor) : Numerical.
2. Sex (Predictor) : Character type with either "female" or "male". For convenience, we changed this to a dummy variable by letting "female" as 1 and "male" as 0.
3. Bmi (Predictor) : Numerical.
4. Children (Predictor) : Numerical.
5. Smoker (Predictor) : Character type with either "yes" or "no". For convenience, we changed this to a dummy variable by letting "yes" as 1 and "no" as 0.
6. Region (Predictor) : Character type with 4 categories. For convenience, we changed this to a dummy variable named "regioncode" by combining "southwest" and "southeast" to 1 and "northwest" and "northeast" to 0.
7. charges (Response) : Numerical.

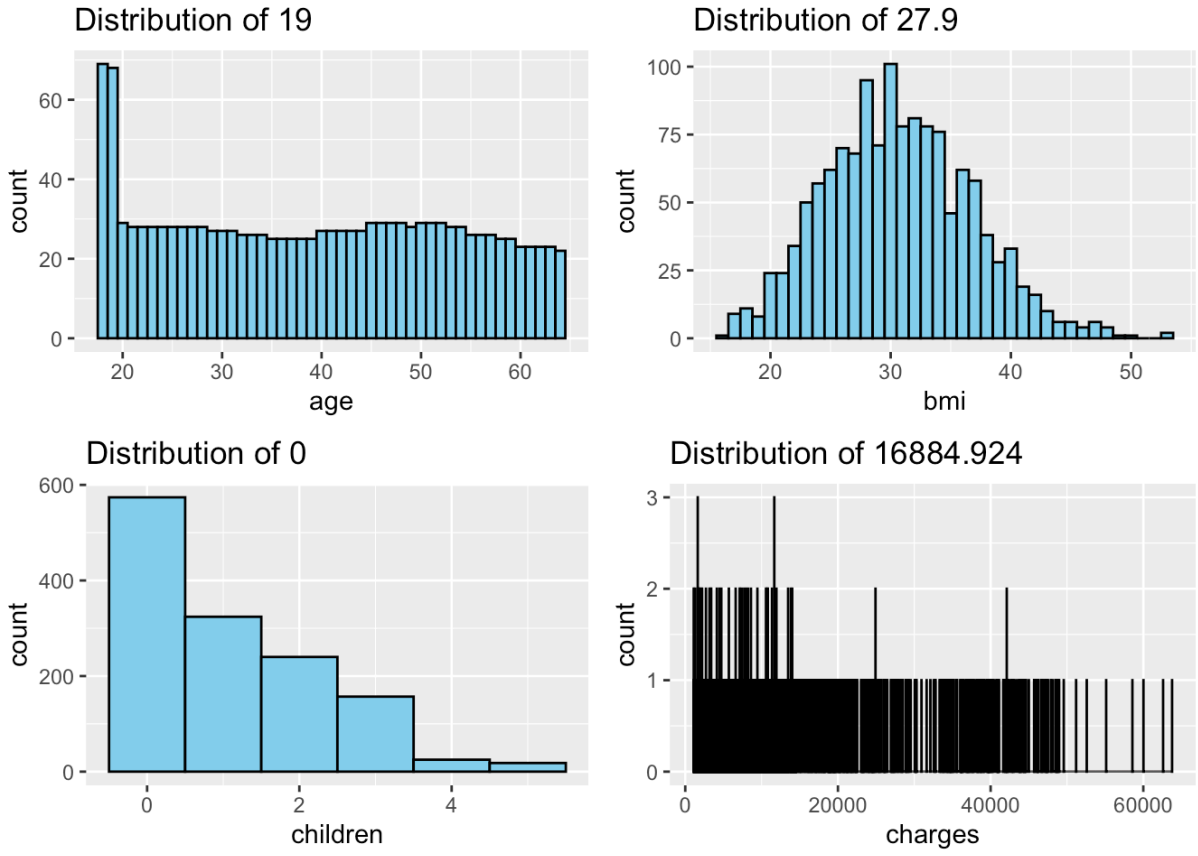
We are only interested in the summary statistics of numerical variables. The correlation coefficient among these predictors is low, thus, we may conclude there are no correlations between these 4 predictors.

```
age   mean:  39.20703  sd:  14.04996
bmi   mean:  30.6634  sd:  6.098187
children   mean:  1.094918  sd:  1.205493
charges   mean:  13270.42  sd:  12110.01
```

```
                age       bmi   children    charges
age      1.0000000 0.1092719 0.04246900 0.29900819
bmi      0.1092719 1.0000000 0.01275890 0.19834097
children 0.0424690 0.0127589 1.00000000 0.06799823
charges  0.2990082 0.1983410 0.06799823 1.00000000
```



The scatter plot matrix provides a visual representation of relationships between variables. Upon analysis, it becomes evident that charges exhibit linear patterns with the majority of predictors, except for BMI which appears to show a less pronounced linear trend.

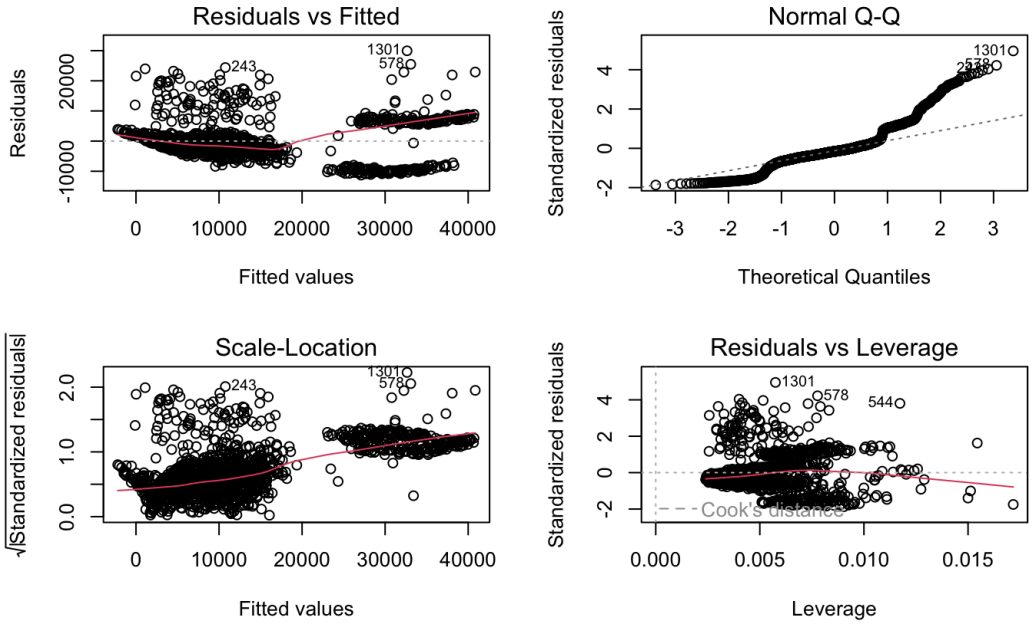


The 'age' variable demonstrates a uniform distribution, except for a notable concentration towards the lower end. This concentration indicates a substantial portion of the population being below the age of 20. The distribution of the 'bmi' variable exhibits the characteristic bell-shaped curve indicative of a

normal distribution, albeit with a subtle rightward skew. This skewness implies a relatively higher proportion of the population possessing a BMI below 30. The 'children' variable, denoting the number of children in the population, conforms to a normal distribution characterized by a prolonged tail and pronounced right skew. Notably, as the count of children increases, the frequency of occurrence diminishes. This pattern aligns with the distribution observed in prior variables, reinforcing the prevailing trend of a considerable portion of the population being below the age of 20. Given the wide range of values within the 'charge' variable, the distribution plot's bins appear relatively thin, potentially impacting visual clarity. Nevertheless, it is discernible that the 'charges' variable maintains a normal distribution, albeit skewed towards the right. This skewness implies that a significant proportion of the population exhibits charges around the vicinity of $20,000.

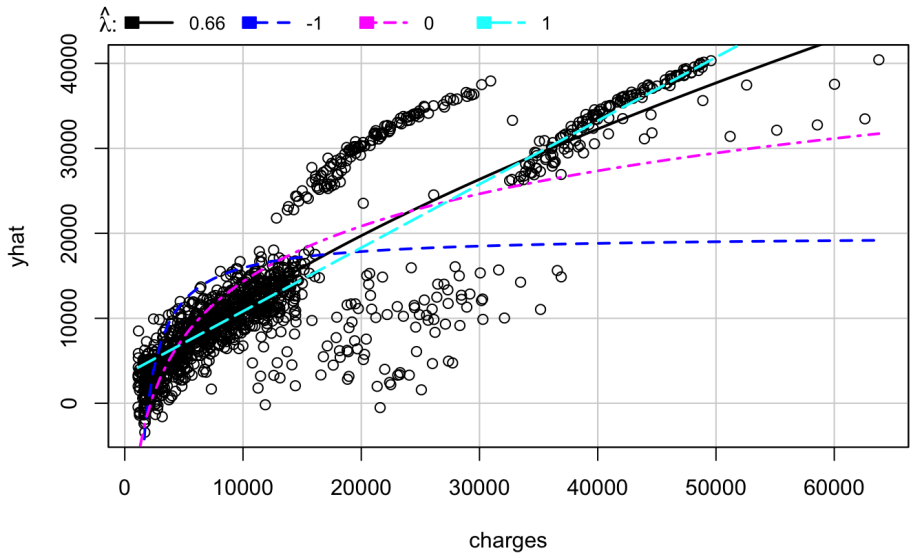
Results and Interpretation

model1 <- lm(charges ~ age + sex + bmi + children + smoker + regioncode)

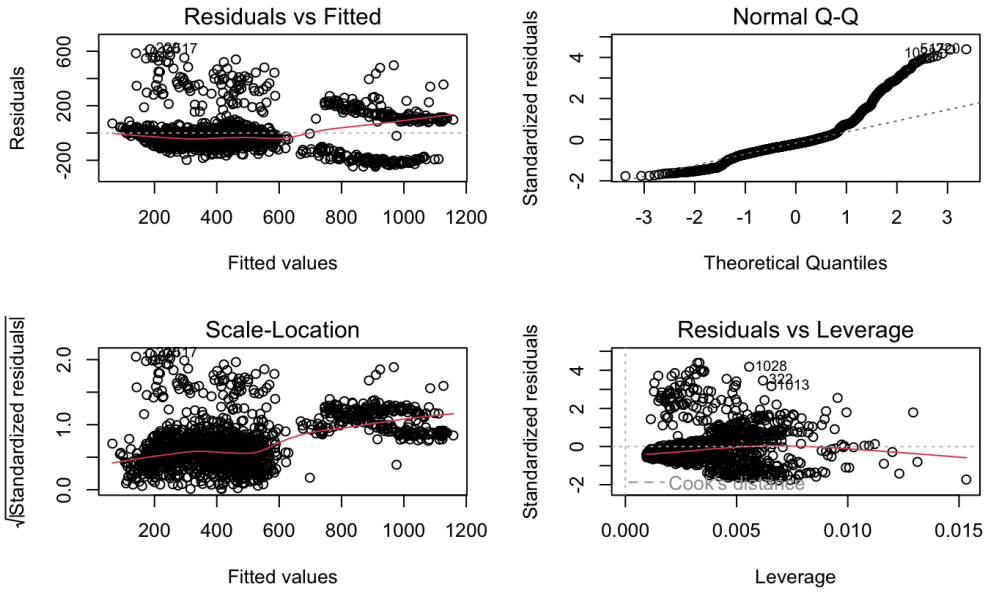



```
Call:
lm(formula = charges ~ age + sex + bmi + children + smoker +
    regioncode, data = dataset)

Residuals:
     Min       1Q   Median       3Q      Max
-11281.1  -2825.0   -988.2   1336.0  29949.9

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -12224.05     962.34 -12.702  < 2e-16 ***
age            256.95      11.89  21.616  < 2e-16 ***
sex            130.29     332.76   0.392 0.695459
bmi            338.38      28.17  12.013  < 2e-16 ***
children       473.12     137.61   3.438 0.000604 ***
smoker       23851.76     411.95  57.899  < 2e-16 ***
regioncode    -820.68     341.26  -2.405 0.016317 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6059 on 1331 degrees of freedom
Multiple R-squared:  0.7508,    Adjusted R-squared:  0.7497
F-statistic: 668.4 on 6 and 1331 DF,  p-value: < 2.2e-16
```

```
> summary(powerTransform(cbind(age,bmi)~1))
bcPower Transformations to Multinormality
    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
age    0.6296         0.5       0.4712       0.7879
bmi    0.4466         0.5       0.2202       0.6730

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)
                              LRT df       pval
LR test, lambda = (0 0) 76.83678  2 < 2.22e-16

Likelihood ratio test that no transformations are needed
                              LRT df       pval
LR test, lambda = (1 1) 43.47758  2 3.6221e-10
```

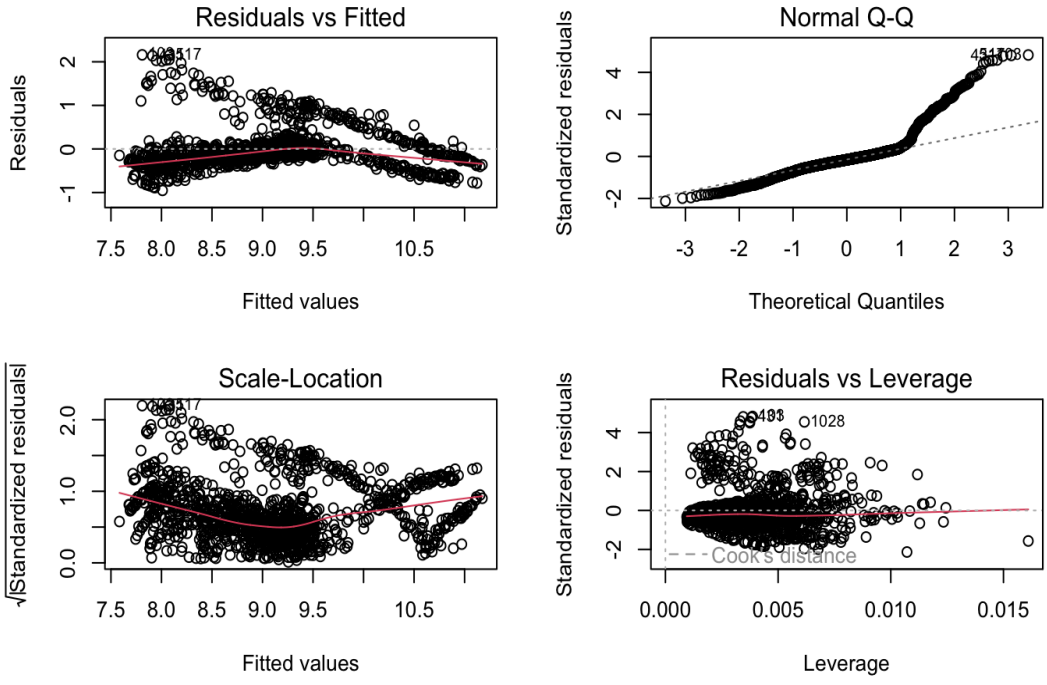


model2 <- lm(I(charges^0.66) ~ I(age^0.5) + I(bmi^0.5) + children + smoker)



```
Call:
lm(formula = I(charges^0.66) ~ I(age^0.5) + I(bmi^0.5) + children +
    smoker)

Residuals:
    Min      1Q  Median      3Q     Max
-247.37  -73.61  -29.09   28.08  615.10

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -619.674     41.900 -14.789  < 2e-16 ***
I(age^0.5)    93.316      3.349  27.864  < 2e-16 ***
I(bmi^0.5)    72.300      6.991  10.342  < 2e-16 ***
children      14.532      3.188   4.559 5.61e-06 ***
smoker       563.058      9.500  59.270  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140.2 on 1333 degrees of freedom
Multiple R-squared:  0.7692,    Adjusted R-squared:  0.7686
F-statistic:  1111 on 4 and 1333 DF,  p-value: < 2.2e-16
```

model3 <- lm(log(charges) ~ log(age) + log(bmi) + children + smoker)



```
Call:
lm(formula = log(charges) ~ log(age) + log(bmi) + children +
    smoker)

Residuals:
     Min       1Q   Median       3Q      Max
-0.95107 -0.22029 -0.07225  0.08745  2.15954

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.98160    0.22566  13.213  < 2e-16 ***
log(age)     1.25266    0.03171  39.504  < 2e-16 ***
log(bmi)     0.35441    0.06091   5.819 7.41e-09 ***
children     0.08155    0.01021   7.986 2.99e-15 ***
smoker       1.54135    0.03038  50.742  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4482 on 1333 degrees of freedom
Multiple R-squared:  0.7631,    Adjusted R-squared:  0.7624
F-statistic:  1073 on 4 and 1333 DF,  p-value: < 2.2e-16
```

The initial linear model included all predictor variables. Interpreting the summary results first, it showed that the “sex” variable is not statistically significant. Moreover, looking at the diagnostic plots, apparent non-linearity and non-constant variance issues were shown. Thus, transformation for the response and predictor variables should be considered. (Moving on, we removed the “region code” variable for better results.) We then used transformation for multiple linear regression, transforming X first with power Transform function in R. The results suggested square root of age and bmi is the optimal choice, and then we used invResPlot to transform Y. The results showed lambda = 0.66 as the best fit and the final model is shown above in model 2. Finally, we tried using the log transformation for all variables, and the model is shown above in model 3. We decided that model 3, the log model, is the “best” predictive model because its diagnostic plots were the most satisfactory and its interpretation is the easiest to understand.

The log(age) , log(bmi), children, and smoker predictors are all statistically significantly greater than zero. Moreover, the coefficient of 0.08155 and 1.54135 for the children and smoker variables highlight that the presence of these aspects in one’s life correlates with an increase in insurance charges. Furthermore, increases in age and body mass index are also associated with increases in insurance charges. The positive values of the variables make sense in the context of the study and in the real world, The review of the diagnostic plots show that the assumptions of linear regression are satisfied, albeit not perfectly. The plot of residuals vs. fitted depicts a red line that is nearly horizontally centered about y=0 value, meaning that there is a linear relationship. The normal qq plot shows almost a perfect straight line from theoretical quantiles -3 to 1 but does not continue the pattern after 1. The standardized residual vs. fitted values plot shows that points are for the most part having constant variance.

Discussion

The results show that almost all the predictor variables involved have a significant impact on the insurance charges. The only one that doesn’t have a significant impact, with 95% level confidence, is the “sex” variable. Most of this data makes sense when applied to the real world. For example, “older individuals can pay up to three times as much as younger folks for coverage” (healthcare.gov). Smoking can also significantly increase premium charges as they “could be up to 50% higher than those paid by someone who doesn't smoke” (Kaiser Family Foundation). One thing to note about this data is that while this data was taken from individuals within the US, there is no information on where this dataset came from specifically, This dataset was only brought up from Kaggle alone. For this dataset to become more credible, multiple sites and sources should be used to make a claim about demographic variables, smoking, and medical costs affecting the insurance premium.