Insurance Data

BY: GROUP 7 - ANDREW, EVELYN, FLORENCIA, RAIYEN, SHARON

BACKGROUND

Based on the increase in insurance charges that are happening, we want to know what makes the insurance rate go up. By using the Insurance dataset taken from Kaggle, we try to indicate **what are the factors that make individual beneficiary's medical costs (Predictor) increased or decreased** based on other variables that are provided such as beneficiary's age, gender, body mass index, number of children, smoker or non-smoker, as well as region of origin.

APPLIED METHODS EXPLANATION

The approach that we use to solve our problem is by using multiple linear regression from insurance data with charge as dependent variable and the other attribute, like age, bmi, smoking status, sex, region as independent variable. Then, we do a few fitting models to see what attributes are deeply correlated to dependent variables(charge).

fit1

```
Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)      7.0286350  0.0817876  85.938 < 0.0000000000000002 ***
age              0.0344775  0.0009975  34.563 < 0.0000000000000002 ***
sexmale         -0.0630094  0.0276554  -2.278              0.02290 *
bmi              0.0135283  0.0023754   5.695          0.0000000159 ***
children         0.1003293  0.0116411   8.619 < 0.0000000000000002 ***
smokeryes        1.5547441  0.0341976  45.464 < 0.0000000000000002 ***
regionnorthwest -0.0593295  0.0392754  -1.511              0.13119
regionsoutheast -0.1741843  0.0394688  -4.413          0.0000112225 ***
regionsouthwest -0.1221596  0.0392828  -3.110              0.00192 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4495 on 1063 degrees of freedom
Multiple R-squared:  0.7617,    Adjusted R-squared:  0.7599
F-statistic: 424.7 on 8 and 1063 DF,  p-value: < 0.00000000000000022
```

fit2

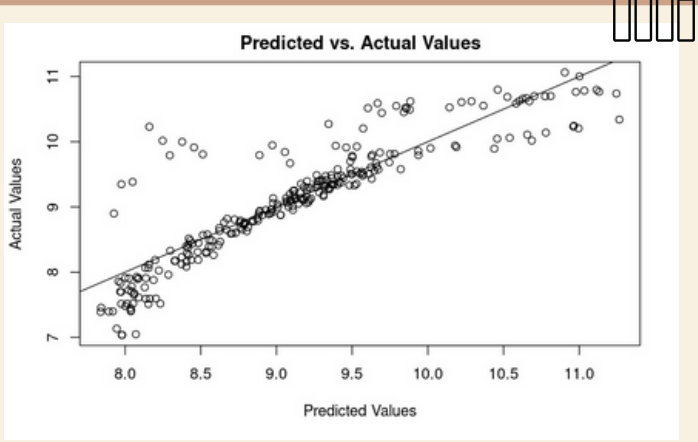
```
Coefficients:
                  Estimate Std. Error t value             Pr(>|t|)
(Intercept)      7.0072634  0.0814077  86.076 < 0.0000000000000002 ***
age              0.0344869  0.0009995  34.505 < 0.0000000000000002 ***
bmi              0.0132067  0.0023759   5.559          0.0000000344 ***
children         0.1000999  0.0116636   8.582 < 0.0000000000000002 ***
smokeryes        1.5475627  0.0341190  45.358 < 0.0000000000000002 ***
regionnorthwest -0.0580901  0.0393489  -1.476              0.14016
regionsoutheast -0.1748504  0.0395454  -4.422          0.0000108065 ***
regionsouthwest -0.1205816  0.0393539  -3.064              0.00224 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4504 on 1064 degrees of freedom
Multiple R-squared:  0.7605,    Adjusted R-squared:  0.759
F-statistic: 482.8 on 7 and 1064 DF,  p-value: < 0.00000000000000022
```

fit3

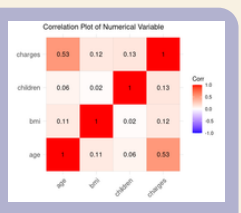
```
Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 6.998181   0.079478  88.052 < 0.0000000000000002 ***
age         0.034593   0.001008  34.335 < 0.0000000000000002 ***
bmi         0.010461   0.002307   4.535          0.00000643 ***
children    0.100353   0.011747   8.543 < 0.0000000000000002 ***
smokeryes   1.542850   0.034382  44.874 < 0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4544 on 1067 degrees of freedom
Multiple R-squared:  0.7556,    Adjusted R-squared:  0.7547
F-statistic: 824.8 on 4 and 1067 DF,  p-value: < 0.00000000000000022
```

RESULTS & DISCUSSION

After finished 3 model fitting, we found that **fit 3 is the most suitable** fit that we can use as our final model, with age, bmi, amount of children and smoking status as the final dependent variable that have strong correlation with dependent variable in the regression model.



PREDICTION MODEL EXPLANATION

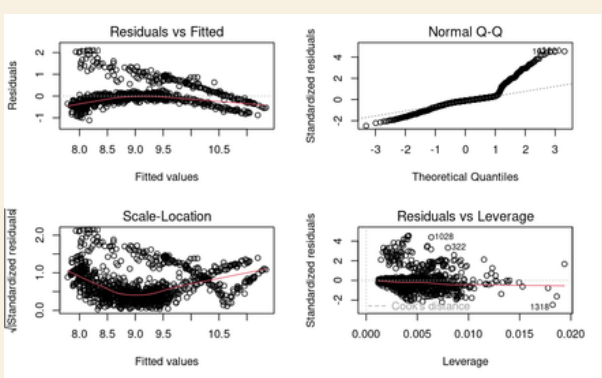


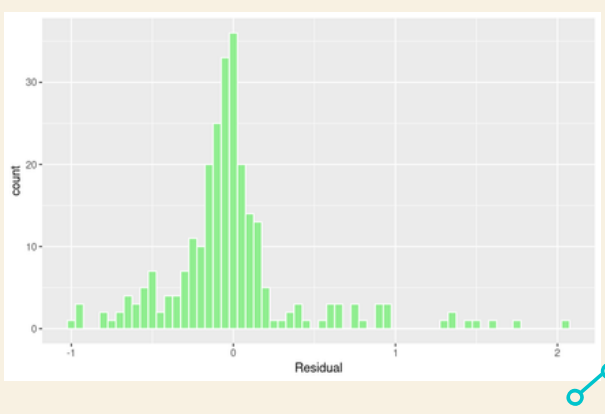
The final regression equation model is :
log(Charges) = 7 + 0.035*age + 0.01*bmi + 0.1*children + 1.54*smokeryes

It means that charges is the response variables. While age, bmi, children, and smokeryes are the predictor variables.

The equation shows that :

- For each **1 age increases**, it **increases the log of charges by 0.035**
- For each **1 BMI increases**, it **increases the log of charges by 0.01**
- For each **1 children** , it **increases the log of charges by 0.1**
- If you **smoke**, it **increases the log of charges by 1.54**





CONCLUSION

For the final result we can conclude that the factors which affect the increasing or decreasing of the individual medical costs(Predictor) are Primary beneficiary’s age, the Body mass index, number of children covered by health insurance, and the status of a person as a smoker or not. This result was also tested and it gets an **accuracy** of **88.7%** which was a good result.