STAT 170 Final Project

Group 16

Nicole Carter
Jonathan Emmons
Joonho Han
Dylan Nunes

Topic and Motivation

Given data about Medical Cost Personal Datasets (Insurance)

Project Research Question: What factors influence individual medical costs billed by health insurance?

Insurance Dataset

```
      age            sex                bmi            children        smoker             region
 Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000   Length:1338        Length:1338
 1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000   Class :character   Class :character
 Median :39.00   Mode  :character   Median :30.40   Median :1.000   Mode  :character   Mode  :character
 Mean   :39.21                      Mean   :30.66   Mean   :1.095
 3rd Qu.:51.00                      3rd Qu.:34.69   3rd Qu.:2.000
 Max.   :64.00                      Max.   :53.13   Max.   :5.000
    charges
 Min.   : 1122
 1st Qu.: 4740
 Median : 9382
 Mean   :13270
 3rd Qu.:16640
 Max.   :63770
```

Overview:

- Sample of ~1300 adults with health insurance
- Contains basic information about each person and the amount of medical charges they have accumulated
- 6 predictors and 1 response (medical charges)

3 categorical variables:

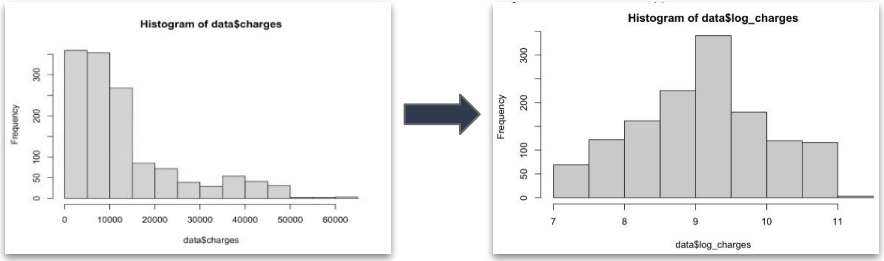
- Sex (Male or Female), Smoker (Yes,No), Region (Northeast, Northwest, Southeast, Southwest)

4 quantitative variables:

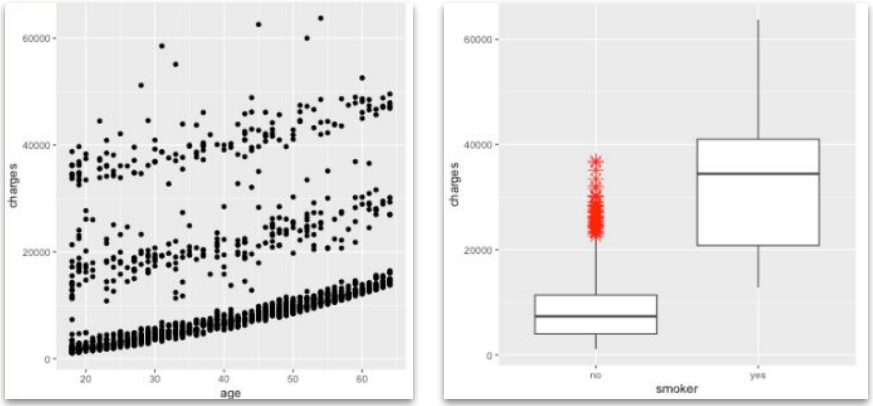
- Age, BMI, Children, Charges

Highlights from EDA

- Discovered that our response variable was not normally distributed and required log transformation



- Plotted each predictor vs response. Age and Smoker seem to have the strongest correlation with Medical Charges.



Final Model

- Stepwise vs All-Possible-Regression Best Subset Models
- Assumptions met?

x_1 : age

x_2 : children

x_3 : bmi

x_4 : smoker (yes)

x_5 : region (northwest)

x_6 : region (southeast)

x_7 : region (southwest)

x_8 : sex (male)

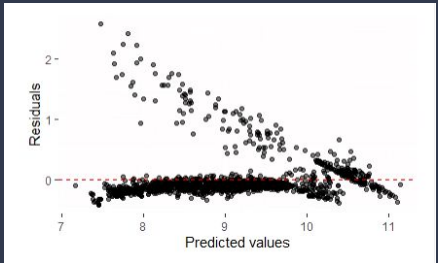
$$\hat{y} = 6.341 + 0.05238x_1 + 0.3477x_2 + 0.04569x_3 + 1.306x_4 - 0.1856x_5 - 0.4361x_6 - 0.4779x_7 - 0.2975x_8 - 0.0001832x_1^2 - 0.0176x_2^2 - 0.0007x_3^2 - 0.03286(x_1x_4) + 0.04959(x_3x_4) - 0.004225(x_1x_2) - 0.127(x_2x_4) + 0.002682(x_1x_5) + 0.007159(x_1x_6) + 0.0073(x_1x_7) + 0.004858(x_1x_8) + 0.07285(x_4x_5) + 0.09043(x_4x_6) + 0.1928(x_4x_7) + 0.09103(x_4x_8)$$

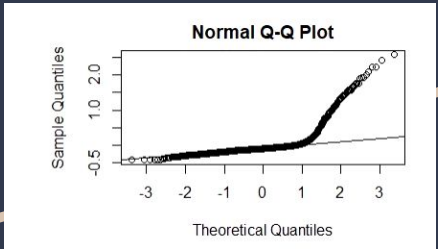
Interesting Findings From Model

- Predictors that didn't make the model
- Unexpected coefficients from terms and interactions

```
> summary(second_step_model1)

Call:
lm(formula = log_charges ~ smoker + age + children + bmi + region +
    sex + I(age^2) + I(bmi^2) + I(children^2) + smoker:age +
    smoker:bmi + age:children + smoker:children + age:region +
    age:sex + smoker:sex + smoker:region, data = insurance_no_charges)

Residuals:
     Min       1Q   Median       3Q      Max
-0.41691 -0.14947 -0.09516 -0.02579  2.57603

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              6.341e+00  2.254e-01  28.137  < 2e-16 ***
smokeryes                1.306e+00  1.460e-01   8.940  < 2e-16 ***
age                      5.238e-02  5.279e-03   9.923  < 2e-16 ***
children                 3.477e-01  3.607e-02   9.639  < 2e-16 ***
bmi                      4.569e-02  1.284e-02   3.557 0.000388 ***
regionnorthwest         -1.856e-01  8.763e-02  -2.117 0.034415 *
regionsoutheast         -4.361e-01  8.572e-02  -5.087 4.16e-07 ***
regionsouthwest         -4.779e-01  8.904e-02  -5.367 9.45e-08 ***
sexmale                 -2.975e-01  6.119e-02  -4.862 1.30e-06 ***
I(age^2)                -1.832e-04  6.315e-05  -2.900 0.003791 **
I(bmi^2)                -7.000e-04  2.018e-04  -3.468 0.000541 ***
I(children^2)           -1.796e-02  6.273e-03  -2.863 0.004258 **
smokeryes:age           -3.286e-02  1.821e-03 -18.041  < 2e-16 ***
smokeryes:bmi            4.959e-02  4.241e-03  11.693  < 2e-16 ***
age:children            -4.225e-03  6.446e-04  -6.554 8.02e-11 ***
smokeryes:children      -1.270e-01  2.191e-02  -5.795 8.52e-09 ***
age:regionnorthwest      2.682e-03  2.069e-03   1.296 0.195131
age:regionsoutheast      7.159e-03  2.017e-03   3.550 0.000398 ***
age:regionsouthwest      7.300e-03  2.082e-03   3.507 0.000469 ***
age:sexmale              4.858e-03  1.444e-03   3.365 0.000788 ***
smokeryes:sexmale        9.103e-02  5.141e-02   1.771 0.076860 .
smokeryes:regionnorthwest 7.285e-02  7.407e-02   0.984 0.325517
smokeryes:regionsoutheast 9.043e-02  7.019e-02   1.288 0.197849
smokeryes:regionsouthwest 1.928e-01  7.474e-02   2.580 0.009989 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3689 on 1314 degrees of freedom
Multiple R-squared:  0.8419,    Adjusted R-squared:  0.8391
F-statistic: 304.1 on 23 and 1314 DF,  p-value: < 2.2e-16
```

Conclusions, Limitations/Future Work





- High R^2, but…
  - Two distinct groups in Residual plot means at the very least constant variance assumption is violated.
  - Non-normal Q-Q Plot means normality assumption is also violated
- What does this mean?
  - Our data is not generalizable to the population, but still appears to at least fit our training data.
  - There appears to be a hidden variable.
- In the future…
  - Investigate into the hidden variable further
  - Re-fit our model with the hidden variable taken into account.