Medical Insurance Regression Analysis

Zachery Gebreab, Kyle Keshmeshian,
John Moro Dutra, Kenneth Nguyen

Introduction

2.7k

Observations

07

Variables

Categorical: smoker, region, sex
Numerical: Cost, age, BMI, children

Analysis Goal

Our goal is to better understand which factors affect medical insurance costs

Data Cleaning

1.3k

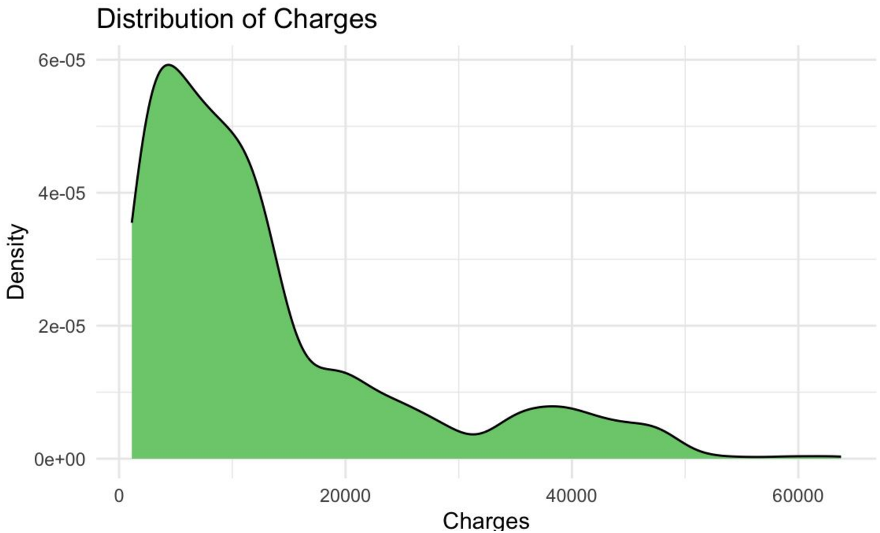
Observations After Removing
Duplicates and Locating Null Values

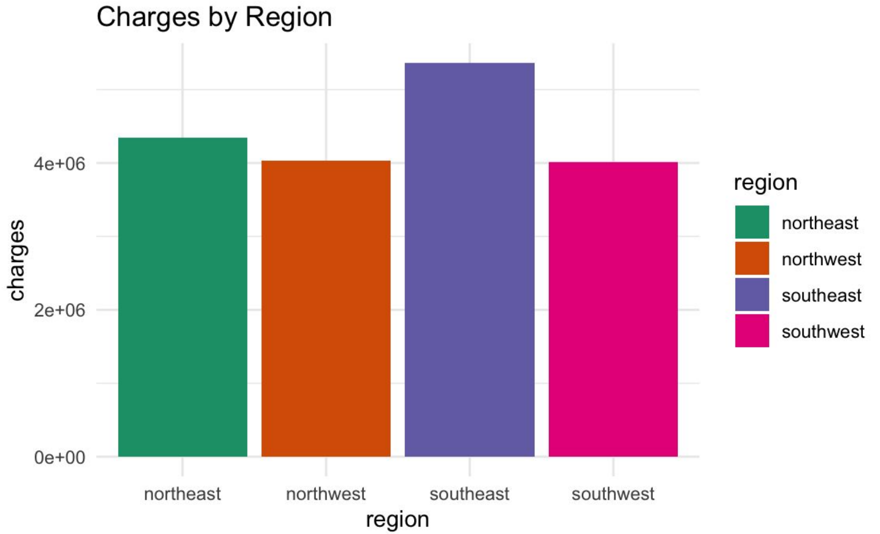
Variables

Regressors: Age, Sex, BMI, Children, Smoker, Region
Response: Charges

01

Exploratory Data Analysis

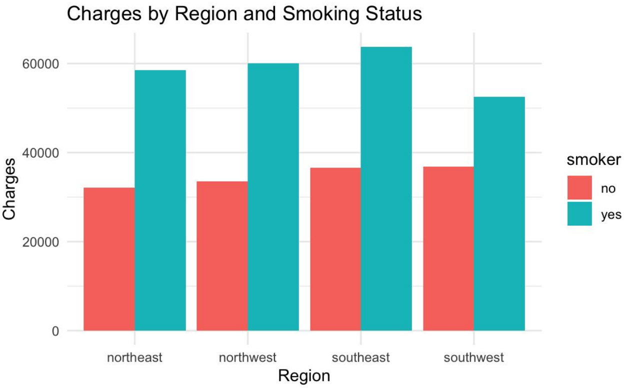


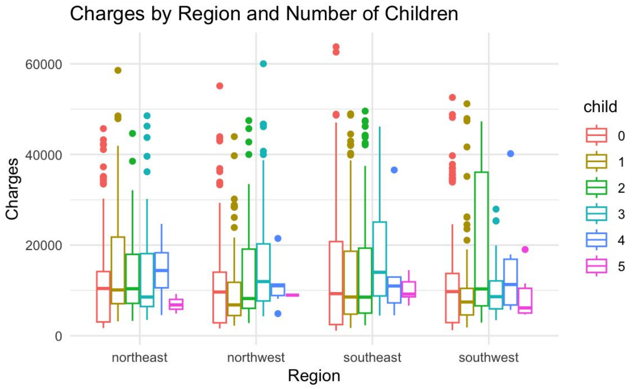


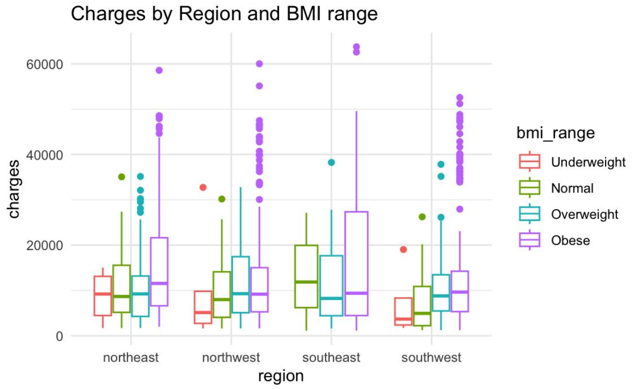
- ❑ Right skewed distribution
- ❑ Majority of charges are under $20,000
- ❑ Range from $1.12k to $63.8k on the high end

- ❑ Southeast region has highest insurance charges
- ❑ Other regions have about the same

Do the higher charges in the Southeast region affect other variables?

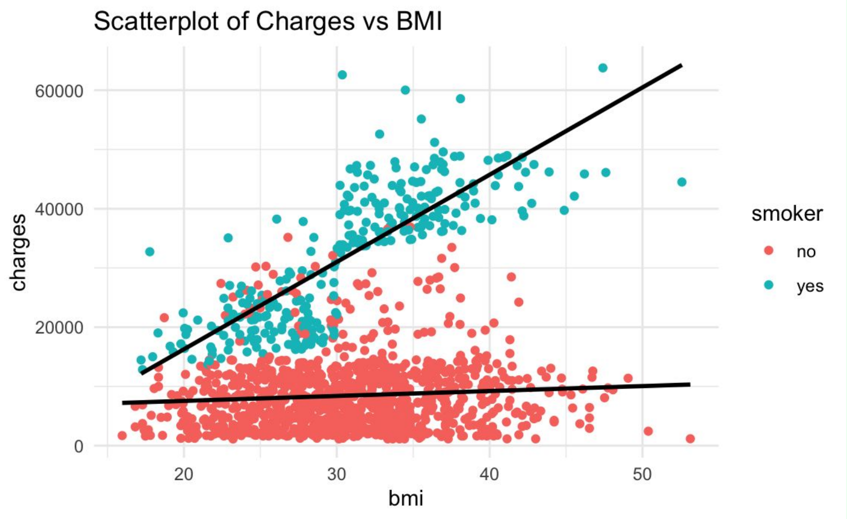


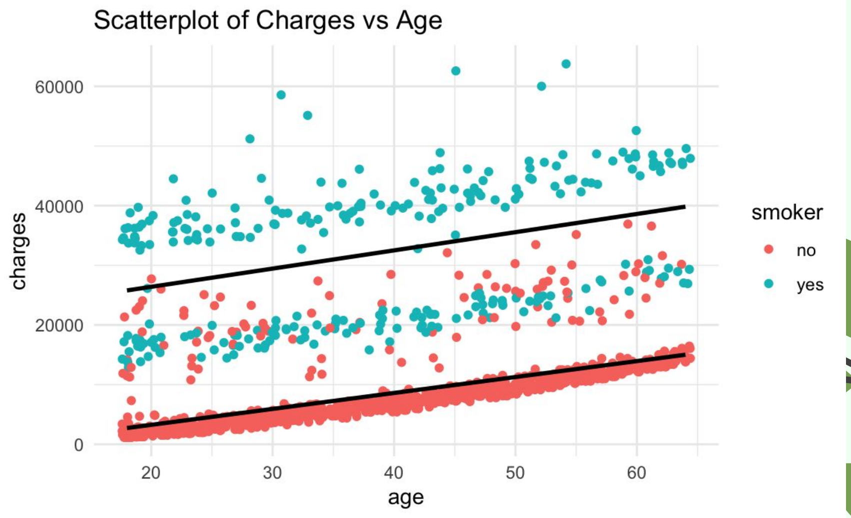




- ❑ Insurance charges are significantly higher for smokers considering they only make 20% of sample population
- ❑ IQR of charges for all counts of children do not have significant variation
- ❑ BMI's that show obesity have the largest ranges of charges generally on the higher end
- ❑ Southeast region does have higher insurance charges in their reasonable range
- ❑ Overall, Southeast region does not show effect from smoking status, number of children, or BMI

- Smokers generally have significantly higher charges than non smokers
- BMI and smokers together show steep linear trend
- Weak but evident linear trend between age and charges





02

Fitting Our Model

Models

```
summary(full_model)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-11305.1  -2850.3   -979.9   1395.0  29992.8

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11936.56     988.23 -12.079  < 2e-16 ***
age                256.76      11.91  21.555  < 2e-16 ***
sexmale           -129.48     333.20  -0.389 0.697630
bmi                339.25      28.61  11.857  < 2e-16 ***
children           474.82     137.90   3.443 0.000593 ***
smokeryes        23847.33     413.35  57.693  < 2e-16 ***
regionnorthwest   -349.23     476.82  -0.732 0.464053
regionsoutheast  -1035.27     478.87  -2.162 0.030804 *
regionsouthwest   -960.08     478.11  -2.008 0.044836 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6064 on 1328 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7492
F-statistic:   500 on 8 and 1328 DF,  p-value: < 2.2e-16
```

```
summary(reduced_model)
```

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

Residuals:
     Min       1Q   Median       3Q      Max
-11366.5  -2841.4   -976.9   1364.0  29936.4

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     -11987.42     979.21 -12.242  < 2e-16 ***
age                256.88      11.91  21.577  < 2e-16 ***
bmi                338.73      28.57  11.856  < 2e-16 ***
children           473.86     137.83   3.438 0.000604 ***
smokeryes        23835.21     412.04  57.847  < 2e-16 ***
regionnorthwest   -348.25     476.66  -0.731 0.465152
regionsoutheast  -1034.63     478.71  -2.161 0.030852 *
regionsouthwest   -959.42     477.95  -2.007 0.044914 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6062 on 1329 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7494
F-statistic: 571.8 on 7 and 1329 DF,  p-value: < 2.2e-16
```

Anova

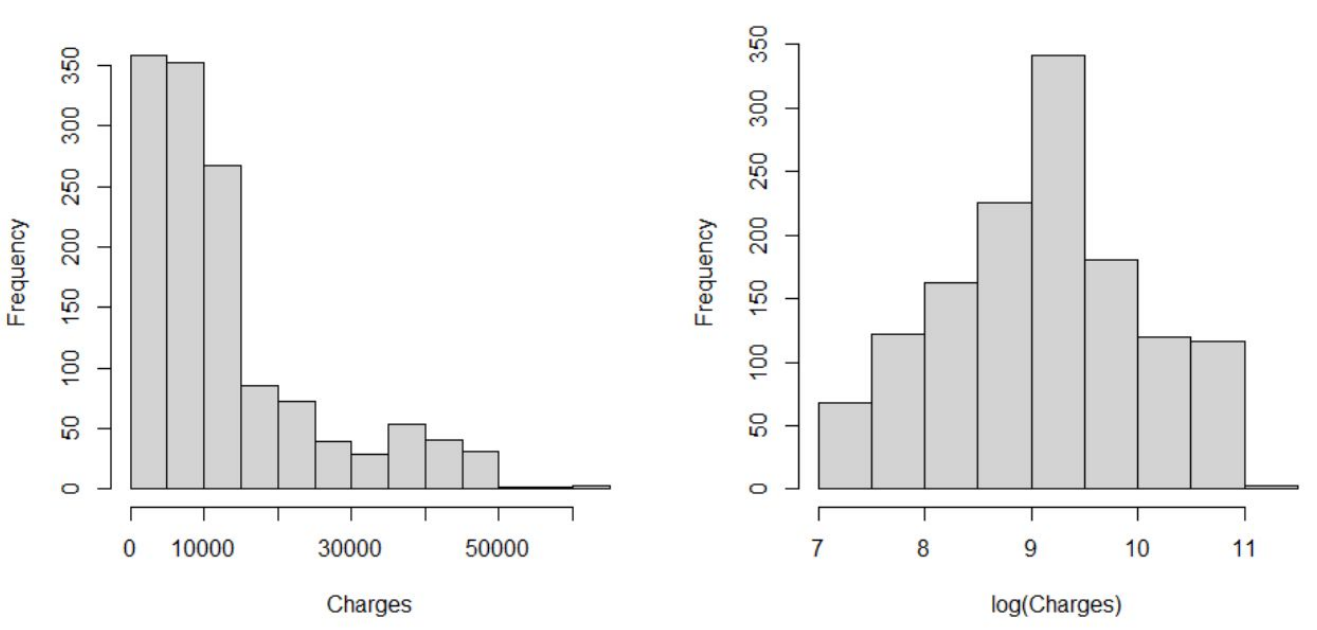
```
Analysis of Variance Table

Model 1: charges ~ age + bmi + children + smoker + region
Model 2: charges ~ age + sex + bmi + children + smoker + region
  Res.Df        RSS Df Sum of Sq     F Pr(>F)
1   1329 4.8844e+10
2   1328 4.8838e+10  1   5553651 0.151 0.6976
```

- Since the p value is > 0.05, then that leads us to be able to assume that we can reduce our model to not utilize the “sex” variable

03

Residual and Influential Analysis

Transforming Our Data

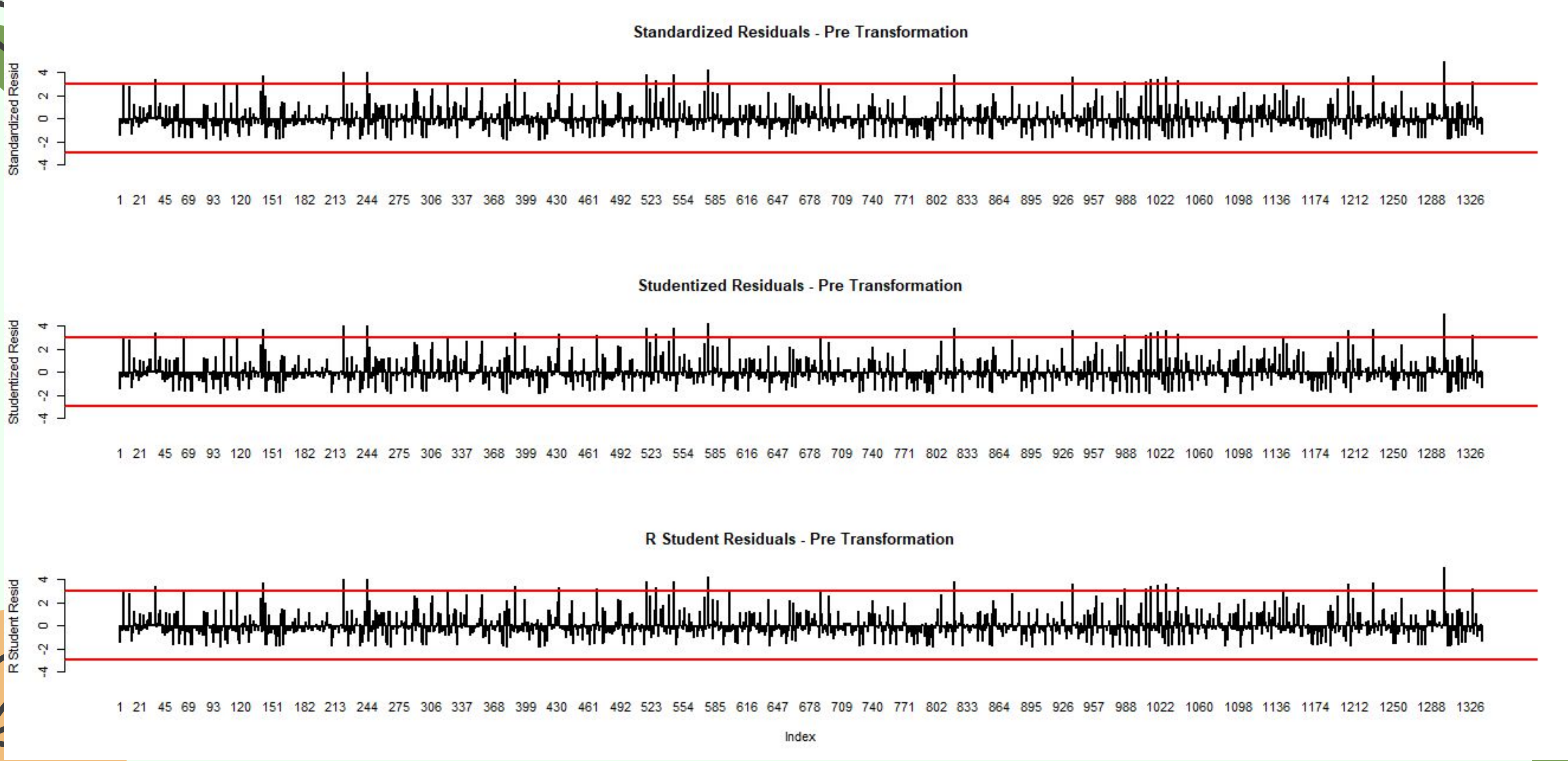



```
> AIC(fit, fit_log)
        df        AIC
fit     10 27096.2154
fit_log 10  -594.9611
> BIC(fit, fit_log)
        df        BIC
fit     10 27148.1973
fit_log 10  -542.9792
```

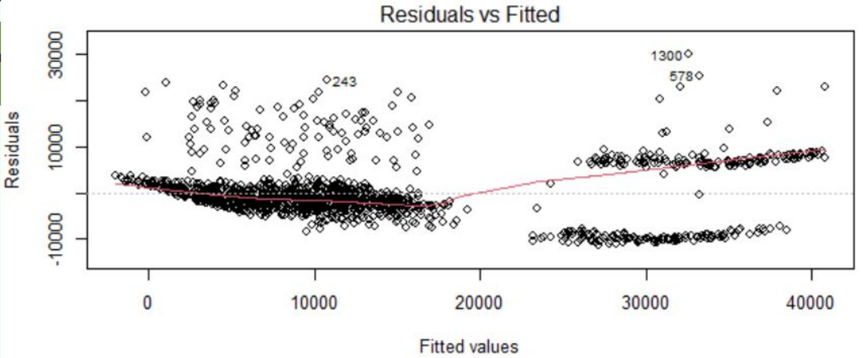
Applying the log function to our charges gave us an approximately normal distribution

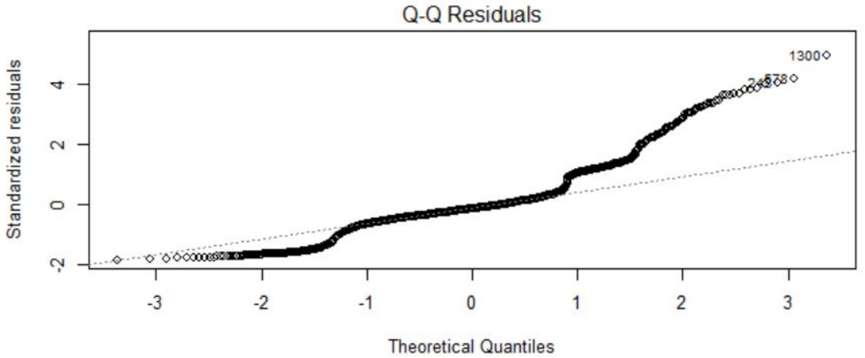
Low AIC and BIC values

Pre-Transformation Residuals

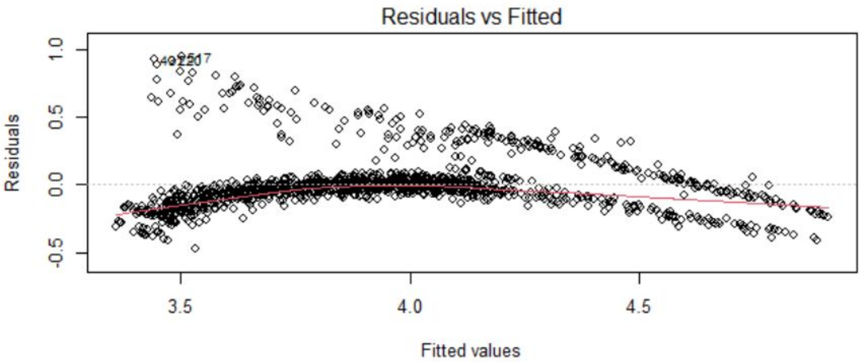


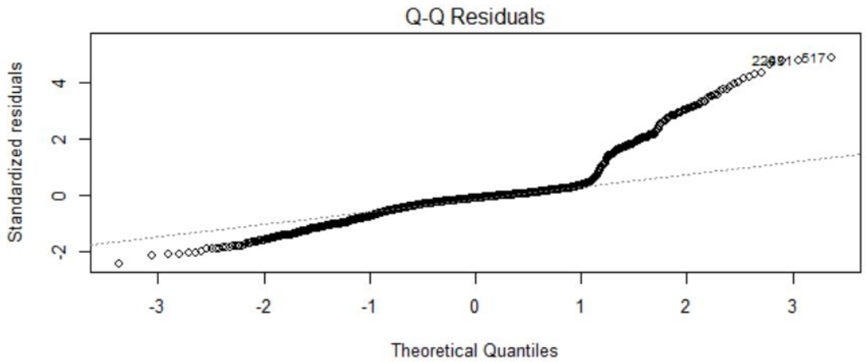
Pre-Transformation



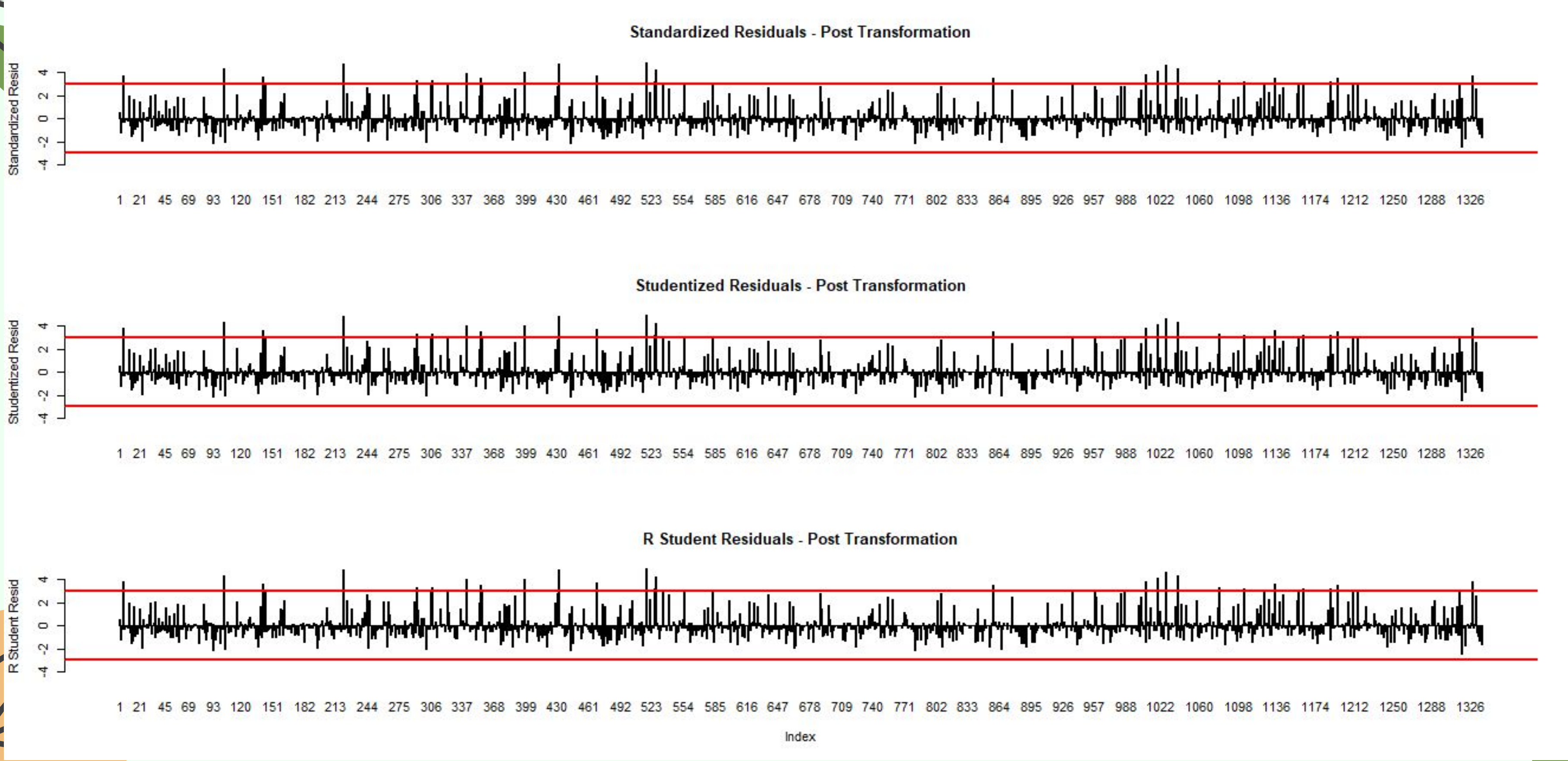


Post-Transformation

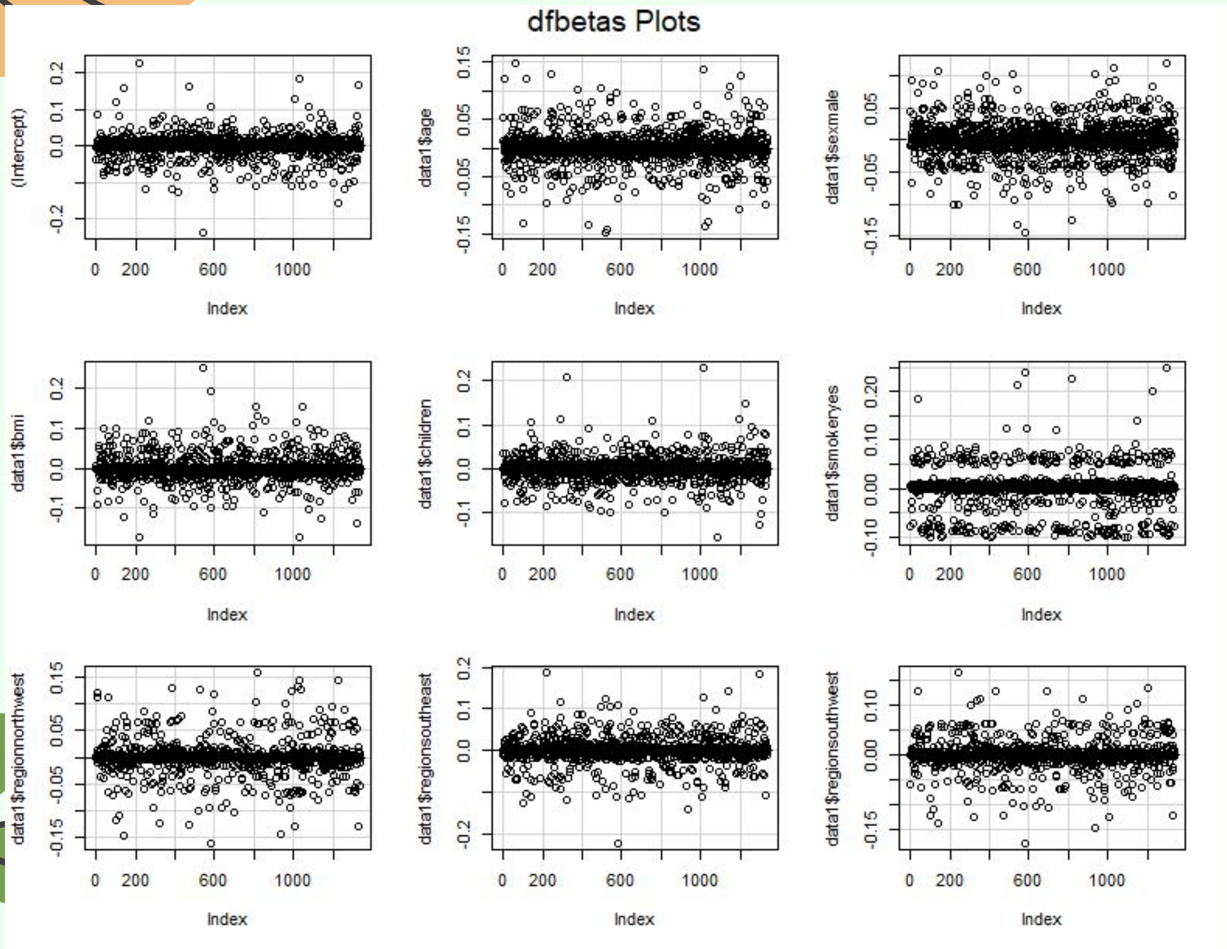




Post-Transformation Residuals



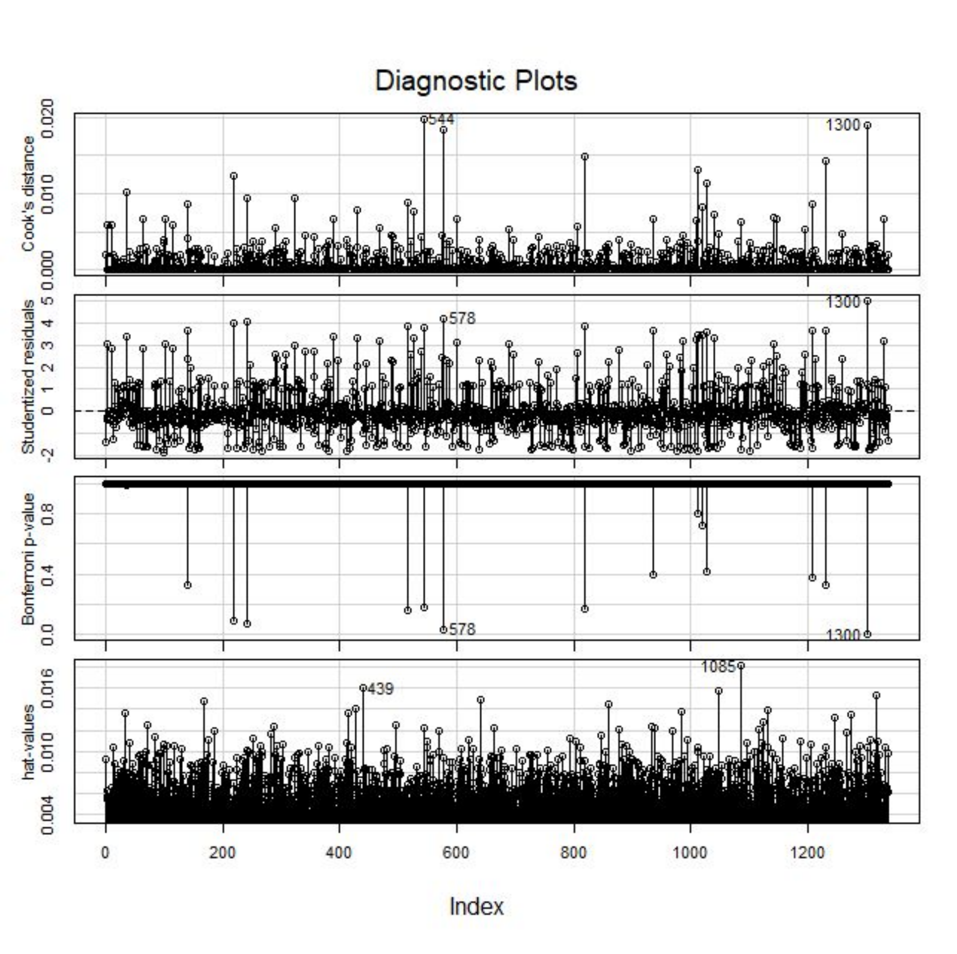
Influential Analysis



A pattern that repeats in most of our graphs is the concentration of observations. This tells us that the majority of points are of comparable influences.

This is the case for most graphs except for the one corresponding to the "smoker" variable. This discrepancy suggests a significant distinction in influence based on whether the user is categorized as a smoker or non-smoker.

Influential Analysis



Most influential observations:

544 - 54 y/o Female, 47.41BMI, No children, Smoker, Southeast, $63,770.42

1300 - 45 y/o Male, 30.36BMI, No children, Smoker, Southeast, $62,592.87

Potential Leverage Points:

439 - 52 y/o Female, 46.75BMI, Five children, Nonsmoker, Southeast, $12,592.53

1085 - 39 y/o Female, 18.3BMI, Five children, Smoker, Southwest, $19,023.26

Our VIFs are small so there is most likely no multicollinearity in our data

```
> vif(fit)
                    GVIF Df GVIF^(1/(2*Df))
data1$age       1.016794  1        1.008362
data1$sex       1.008944  1        1.004462
data1$bmi       1.106742  1        1.052018
data1$children  1.004017  1        1.002006
data1$smoker    1.012100  1        1.006032
data1$region    1.099037  3        1.015864
```

Conclusion

Pre-Transformation

Pre-Transformation, our numerical-based model was:
Charges = -12098.82 + 257.77(**age**) + 321.87(**bmi**) + 472.98(**children**) + 23810.40(**smoker)**

Post-Transformation

Post-Transformation, our numerical-based model was:
log(Charges) = 6.99 + 0.03(**age**) + 0.01(**bmi**) + 0.1(**children**) + 1.54(**smoker**)

Furthermore, we noticed that sex was not a significant predictor for medical insurance costs

Sidenote: we were not able to include the categorical regions in here as R splits that variable up into various ones

Thank you!