Health Insurance Premium Cost Prediction

ISI, NEW DELHI

YASH GUPTA

07-11-2022

Contents

1 Introduction **1**
- 1.1 Abstract . . . 1
- 1.2 Motivation . . . 1
- 1.3 Data Set . . . 1

2 Methodology **3**
- 2.1 Exploratory Data Analysis . . . 3
- 2.2 Model Building . . . 3
  - 2.2.1 Applying Multiple Linear Regression . . . 3
  - 2.2.2 F-statistic, . . . 4
  - 2.2.3 Model Selection Using AIC . . . 4
  - 2.2.4 Deleting unneccessary covariates using T-tset . . . 5
- 2.3 Detecting influentical observatoins and high leverage observations . . . 5
  - 2.3.1 Cook Distance and Laverage . . . 5
  - 2.3.2 standardized residuals . . . 6
  - 2.3.3 Studentized residuals . . . 6
- 2.4 Checking for required assumption for Linear regression . . . 6
  - 2.4.1 Checking for non-normality of error . . . 7
  - 2.4.2 Checking for non-constant error variance . . . 7
  - 2.4.3 Possible remedies for voilation of assumption . . . 7
- 2.5 Box-Cox Transformation . . . 7
- 2.6 . . . 8
- 2.7 Applying Robust Regression . . . 8
  - 2.7.1 Huber Loss Function . . . 8
  - 2.7.2 Bisquare Loss Function . . . 9
  - 2.7.3 Least Absolute Devation . . . 9
  - 2.7.4 Resistant Regresssion . . . 9
  - 2.7.5 Least Medain of Squares (LMS) . . . 9

3 Data Analysis **10**
- 3.0.1 Data . . . 10
- 3.1 **Exploratory Data Analysis** . . . 11
- 3.2 Creating dummy variable for BMI_Category . . . 14
- 3.3 Plot of charge vs age . . . 15

3.4 First step to build prelimnary model . . . 15
3.5 Plot of charge vs BMI across smoking . . . 16
3.6 **Explanation** . . . 16
3.7 Plot of charge vs age across smoking . . . 17
3.8 Plot of charge vs age across BMI_Category . . . 18
3.9 Updated model . . . 18
3.10 Plot of charge vs children across smoking . . . 19
3.11 Plot of charge vs children across smoking . . . 20
3.12 Updated model . . . 20
3.13 Plot of charge vs BMI across regions . . . 21
3.14 Plot of charge vs BMI across BMI_Category . . . 22
3.15 Plot of charge vs age across smoking . . . 23
3.16 Preliminary Model . . . 23
3.17 Dividing the data into training and testing data . . . 23
3.18 Preliminary Model . . . 24
3.19 Full Model . . . 24
3.20 Comaparing both above models on test data using MSE and MAD . . . 25
3.21 Comparison with model without children covariate . . . 25
3.22 Best multiple linear model based on observation . . . 26
3.23 Using AIC to find the best model . . . 26
3.24 BEST MODEL USING AIC . . . 27
3.25 BEST SELECTED MODEL USING AIC . . . 28
3.26 Deleting statistically insignificant covariates from the AIC Model . . . 28
3.27 Comparision between the AIC model and Final Model Using F-test . . . 28
3.28 Checking required assumption for multiple linear regression . . . 29
3.29 Checking for Outliers . . . 29
3.29.1 Cook's Distance Plot . . . 31
3.30 No of outliers (on the basis of Cook distance cutoff) . . . 32
3.31 Laverage vs Std. Residual Plot . . . 32
3.32 Checking for normality or residuals . . . 33
3.32.1 Plot of absolute studentized residual vs fitted value . . . 36
3.33 Plot of residuals for different covariates . . . 37
3.34 BOX-COX transformation . . . 37
3.35 Robust Regression . . . 39
3.36 **Huber Loss Function** . . . 40
3.37 Resistant Regression . . . 40
3.38 Least Median of Squares (LMS) . . . 41
3.39 Comparison between Least Squares, LAD, Huber Loss, BSq Loss, LTS and LMS 41
3.40 Accuracy of the different models . . . 42
3.41 Accuracy of the different models . . . 42

4 Result and conclusion **43**

Chapter 1

Introduction

1.1 Abstract

The Project is related to neccessary project for the course- Regression Techniques ISI, New Delhi. It's based on Regression techniques. The aim of the project is to predict the yearly charges of health insurance premium for the given data set. The data set is publicly available on github provided in datasets for Machine Learning With R by Brett Lantz. An exploratory data analysis(EDA) is carried out in order to have a good idea about data set and build a good model. I have applied some other regression techniques like Akaike information criterion (AIC), Robust Regression by using Bisquare and Huber loss function, Resistant Regression like LTS and LMS.

1.2 Motivation

1.3 Data Set

We have used the USA's medical cost personal dataset from kaggle, having 1338 entries. Features in the dataset that are used for the prediction of insurance cost include: Age, Gender, BMI, Smoking Habit, number of children etc. We used linear regression and also determined the relation between price and these features. We trained the system using a 70-30 split and achieved an accuracy of .

The Dataset Contains Health Related Features Of The Customers.

- **Numeric predictors**
- age: age of beneficiary
- bmi: bmi of beneficiary
- children: no of children covered under health insurance
- **Categorical predictors**

- sex: sex of beneficiary (male or female)
- smoker: yes or no
- region: beneficiary's residential area in USA : northeast, southeast, southwest, northwest.
- **Response variable**
- charges: Individual medical costs billed by health insurance
- **Data size** : 1338 observations (1338 rows X 7 columns)

Chapter 2

Methodology

2.1 Exploratory Data Analysis

To understand the dataset the first step is EDA. I used the following plots to understand the relationship between covariates and response variable.

- Boxplot
- Scatter Plot
- Box Whisker Plot
- Corrplot
- Histogram
- qqmath

2.2 Model Building

After getting idea about the data set and how the response variable is related to covariates, I started to fit multiple linear regression on the data.

2.2.1 Applying Multiple Linear Regression

Multiple linear regression can be defined as extended simple linear regression. It comes under usage when we want to predict a single output depending upon multiple input or we can say that the predicted value of a variable is based upon the value of two or more different variables. The predicted variable or the variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable) and the variables being used in predict of the value of the dependent variable are called the independent variables (predictors).

The procedure for fitting Linear regression model are-

1. There is very strong linear relationship between age and charges.

2. The response were dependent upon the factors like somking and BMI_Category (Normal, Overweight, Obese). To get the effect of that I added another column of BMI_Category in the data set. Such variables are called dummy variables.
3. I avoided adding unnecessary looking covariates (Which has no impact on the insurance charge as per observation thorugh EDA like sex and region) in the preliminary model.
4. I used Anova() function as F-Test to compare the two models (Theory is given below).

Detailed Model Building Procedure is explained in the Data Analysis section.

2.2.2 F-statistic,

- Suppose we are interested in testing H_0 vs H_1, where H_0 is a sub-model of H_1

$$H_0 \subset H_1 \quad (H_m : \boldsymbol{Y} \sim N_n(\boldsymbol{X}_m\beta_m, \sigma^2\boldsymbol{I}))$$

- Let the sum of squared errors for the two models be S_0^2 and S_1^2

$$S_m^2 = \left\|\boldsymbol{Y} - \boldsymbol{X}_m\widehat{\beta}_m\right\|^2, m = 0, 1$$

- Let the number of parameters (length of β) in the two models be p_0 and p_1
- Then the test statistic

$$F = \frac{\frac{S_0^2 - S_1^2}{p_1 - p_0}}{\frac{S_1^2}{n-p_1}} \text{ follows } F_{p_1-p_0, n-p_1} \text{ under } H_0$$

2.2.3 Model Selection Using AIC

Most common approach to select best model is to make some criterion which measure overall quality of a each model and punish both overly simple and overly complex models. one such approach is AIC (Akaike Information Criterion)

Akaike showed that

$$-2E\left(\sum_i \log P_{\hat{\theta}}(y_i)\right) \approx -2E(\text{loglik}) + 2p$$

Here loglik is the maximized log-likelihood for the fitted model

This suggests the Akaike Information Criterion (AIC)

$$\text{AIC} = -2\text{loglik} + 2p$$

For linear models, this is equivalent to (up to an additive constant)

$$\text{AIC} = n \log \text{RSS} + 2p$$

So, Minimum the AIC better the model. USing Step AIC we'll find the best model in terms of AIC value.Then we'll delete the unnecessary coefficients in the model using p-value of t-test given in the model summary.

2.2.4 Deleting unneccessary covariates using T-tset

We know that,

$$\widehat{\beta} j \sim N(\beta_j, \sigma^2 M_{jj})$$

Standard error" of $\widehat{\beta}_j$

$$\widehat{SD}(\widehat{\beta}_j) = \widehat{\sigma}\sqrt{M_{jj}}$$

Now, we can Test the null hypothesis $H_0 : \beta j = 0$

$$t = \frac{\widehat{\beta}_j}{\widehat{\sigma}_{\widehat{\beta}_j}} \sim t_{n-p}.$$

Thus, if the p value is larger than cutoff point then we failed to reject the null hypothesis i.e. the corresponding covariate is not statistically significant.

2.3 Detecting influentical observatoins and high leverage observations

If an observation has a response value that is very different from the predicted value based on a model, then that observation is called an outlier. On the other hand, if an observation has a particularly unusual combination of predictor values (e.g., one predictor has a very different value for that observation compared with all the other data observations), then that observation is said to have high leverage.

It is important to know how to detect outliers and high leverage data points. Once we've identified any outliers and/or high leverage data points, we then need to determine whether or not the points actually have an undue influence on our model.

These are the following measure to detect outliers and influential observations

2.3.1 Cook Distance and Laverage

Think of testing the "hypothesis" that $\beta = \hat{\beta}_{(-i)}$

- Consider the "F-statistic" for this test, recalculated for each i (though not really meaningful)

- This is known as Cook's distance D_i. It can be shown that

$$D_i = \frac{r_i^2}{p} \times \frac{h_i}{1 - h_i}$$

- D_i can be viewed as a combination of discrepancy and leverage.
- Observations with high values of D_i are considered influential

2.3.2 standardized residuals

Standardized residuals are defined as

$$r_i = \frac{e_i}{\hat{\sigma}\sqrt{1 - h_i}}$$

If the vlaue of standardized residual for any observation is unusually higher than other observations then there is higher chances of it being a influential observation

2.3.3 Studentized residuals

Studentized residuals are defined as

$$t_i = \frac{e_i}{\hat{\sigma}_{(-i)}\sqrt{1 - h_i}}$$

If the vlaue of standardized residual for any observation is unusually higher than other observations then there is higher chances of it being a influential observation. This is similar as standardized residual.

2.4 Checking for required assumption for Linear regression

LSE is Best Linear Unbiased Estimator under assumptions of

`* linearity`

`* constant variance`

`* uncorrelated errors`

- Even if LSE is *valid*, it may not be *efficient*, especially with heavy tailed errors (outliers)
- LSE estimates conditional mean $f(x) = E(Y|X = x)$
  - Justified when distribution of $Y|X = x$ is symmetric

  - May not be appropriate measure of central tendency if distribution of $Y|X = x$ is skewed

2.4.1 Checking for non-normality of error

Kolmogorov-Smirnoff Test for non normality of residuals

- – Null hypothesis: $X_1, \ldots, X_n \sim$ i.i.d. F_0 (where F_0 is a completely specified absolutely continuous CDF)
- Empirical CDF

$$\hat{F}_n(x) = \frac{1}{n} \sum_i \mathbf{1}\{X_i \leq x\}$$

- Test statistic:

$$T(X_1, \ldots, X_n) = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F_0(x)|$$

–

- Note that
  - null distribution of T does not depend on F_0 (use $U_i = F_0(X_i) \sim$ i.i.d. $U(0, 1)$ instead)
  - Intuitively, large value of T indicates departure from null, so reject when T is large
    - You can find the p-value using km.test() in R
- Shapiro Wilk is also used for assessing the non-normality of residual plot

2.4.2 Checking for non-constant error variance

We know

$V(Y|X = x)$ depends on $E(Y|X = x)$

$V(Y|X = x)$ depends on x

We can plot residuals against fitted values / covariates to assess the non constant error variance.

I performed the test

2.4.3 Possible remedies for voilation of assumption

2.5 Box-Cox Transformation

The Box-Cox transformation deals with non-normality, non-linearity, and non-constant variance

$$g_\lambda(y) = \begin{cases} \frac{y^\lambda - 1}{\lambda} & \lambda \neq 0 \\ \log y & \lambda = 0 \end{cases}$$

Where λ is choosen using formal inference procedure. No closed form solution for λ is possible.

B

2.6

2.7 Applying Robust Regression

The ordinary least squares estimates for linear regression are optimal when all of the regression assumptions are valid. When some of these assumptions are invalid, least squares regression can perform poorly. Residual diagnostics can help guide you to where the breakdown in assumptions occur, but can be time consuming and sometimes difficult to the untrained eye. Robust regression methods provide an alternative to least squares regression by requiring less restrictive assumptions. These methods attempt to dampen the influence of outlying cases in order to provide a better fit to the majority of the data.

Outliers have a tendency to pull the least squares fit too far in their direction by receiving much more "weight" than they deserve. Typically, you would expect that the weight attached to each observation would be on average 1/n in a data set with n observations. However, outliers may receive considerably more weight, leading to distorted estimates of the regression coefficients. This distortion results in outliers which are difficult to identify since their residuals are much smaller than they would otherwise be (if the distortion wasn't present). As we have seen, scatterplots may be used to assess outliers when a small number of predictors are present. However, the complexity added by additional predictor variables can hide the outliers from view in these scatterplots. Robust regression down-weights the influence of outliers, which makes their residuals larger and easier to identify.

2.7.1 Huber Loss Function

In regression in Huber Loss Function, we find the β for the fitted linear model as

$$\arg\min_\beta \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta)$$

where

$$\rho(x-\theta) = \begin{cases} (x-\theta)^2 & \text{if } |x-\theta| \leq c \\ c(2|x-\theta| - c) & \text{otherwise} \end{cases}$$

2.7.2 Bisquare Loss Function

In regression in Bisquare Loss Function, we find the β for the fitted linear model as

$$\arg\min_{\beta} \sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta)$$

where

$$\rho(x) = \begin{cases} R^2 \left[1 - (1 - (x/R)^2)\right]^3 & |x| \leq R \\ R^2 & \text{otherwise} \end{cases}$$

2.7.3 Least Absolute Devation

In Least Absolute Deviaton (LAD) regression we find the β for the fitted linear model as

$$\arg\min_{\beta} \text{median} \left\{ |y_i - x_i^T \beta|; i = 1, \ldots, n \right\}$$

2.7.4 Resistant Regresssion

Whereas robust regression methods attempt to only dampen the influence of outlying cases, resistant regression methods use estimates that are not influenced by any outliers (this comes from the definition of resistant statistics, which are measures of the data that are not influenced by outliers, such as the median). This is best accomplished by trimming the data, which "trims" extreme values from either end (or both ends) of the range of data values.

2.7.5 Least Medain of Squares (LMS)

In Least Median Squares (LMS) regression we find the β for the fitted linear model as

$$\arg\min_{\beta} \text{median} \left\{ (y_i - x_i^T \beta)^2; i = 1, \ldots, n \right\}$$

2.7.5.1 Least trimmed Squares

In Least Trimmed Squares (LMS) regression we find the β for the fitted linear model as

$$\arg\min_{\beta} \sum_{i=1}^{q} (y_i - x_i^T \beta)_{(i)}^2$$

Chapter 3

Data Analysis

3.0.1 Data

```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
## 2  18   male 33.770        1     no southeast  1725.552
## 3  28   male 33.000        3     no southeast  4449.462
## 4  33   male 22.705        0     no northwest 21984.471
## 5  32   male 28.880        0     no northwest  3866.855
## 6  31 female 25.740        0     no southeast  3756.622
```

Summary of Data

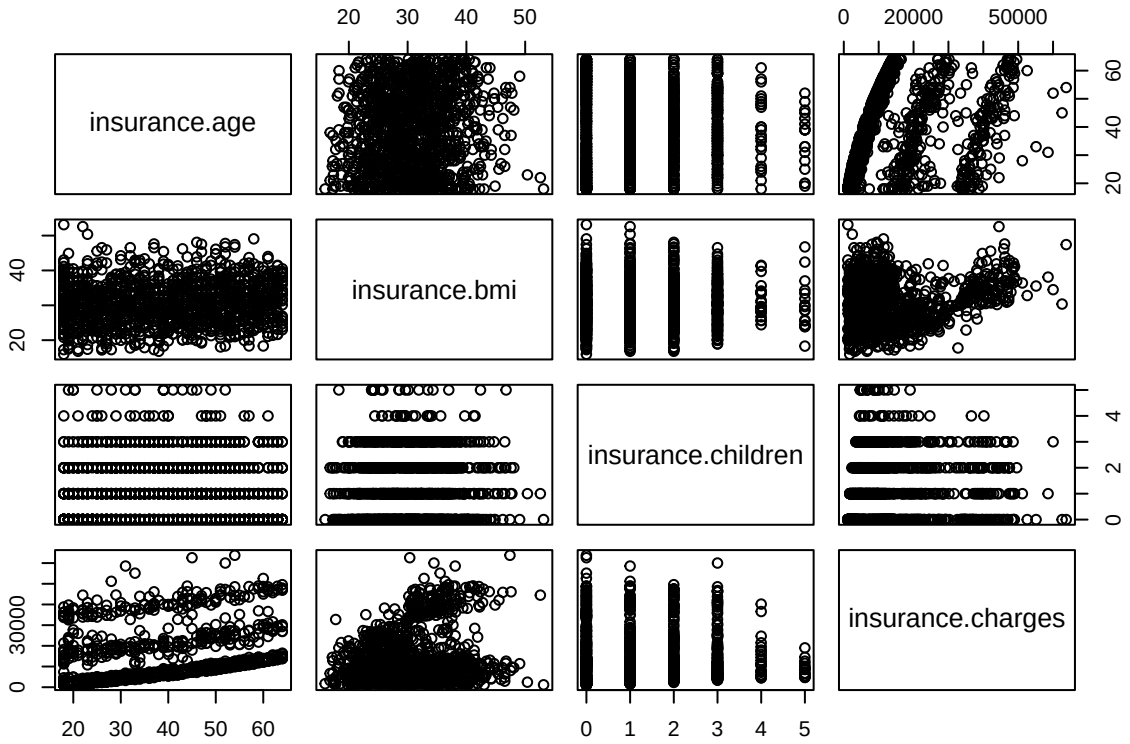
```
##       age            sex                 bmi           children
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000
##  Mean   :39.21     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

3.1 Exploratory Data Analysis

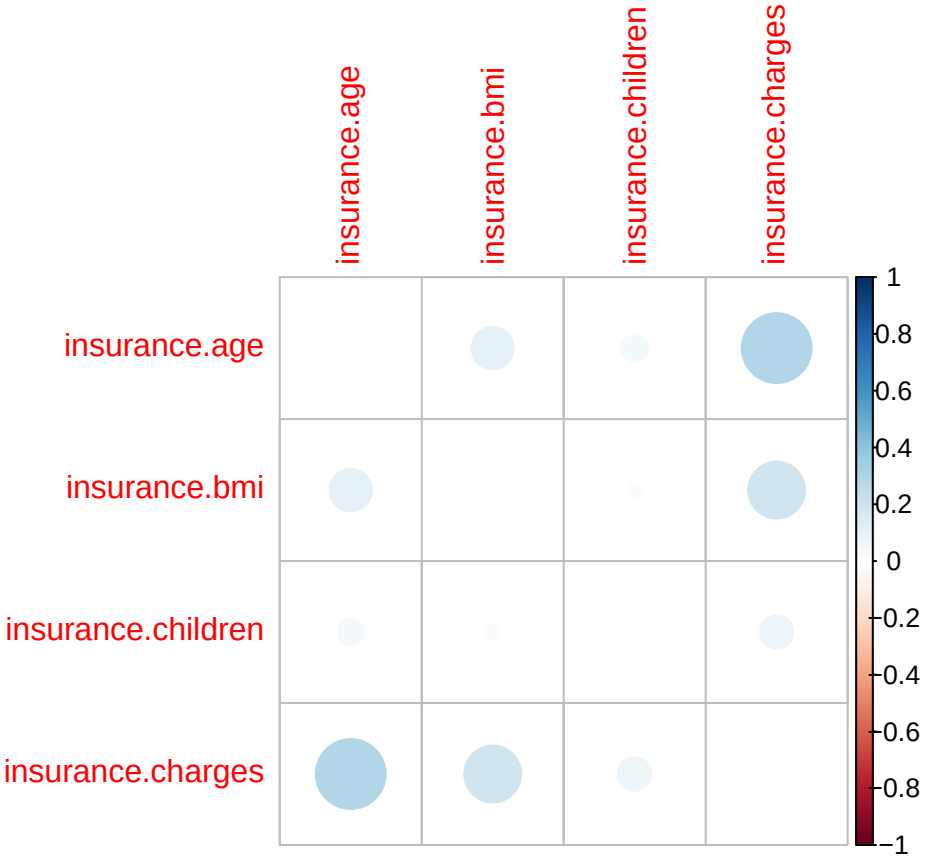
Correlation

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

Plot betweeen numeric variables and response(charges)



Corrplot

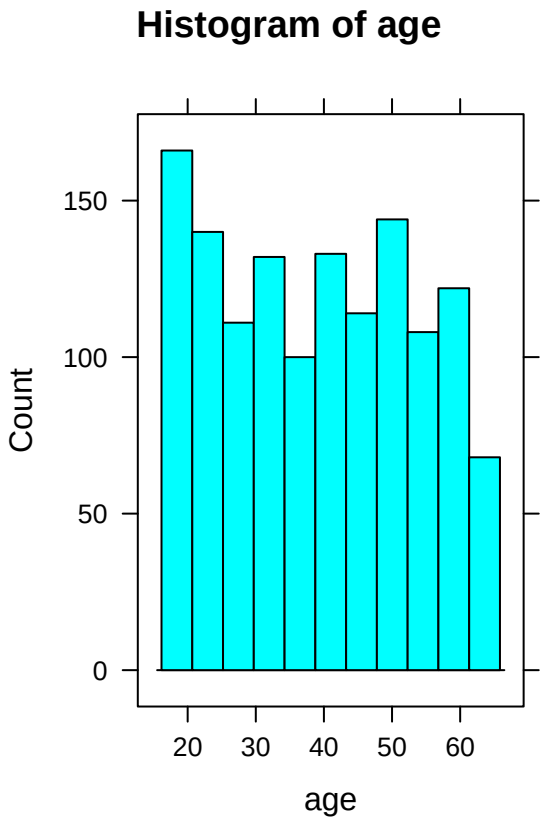


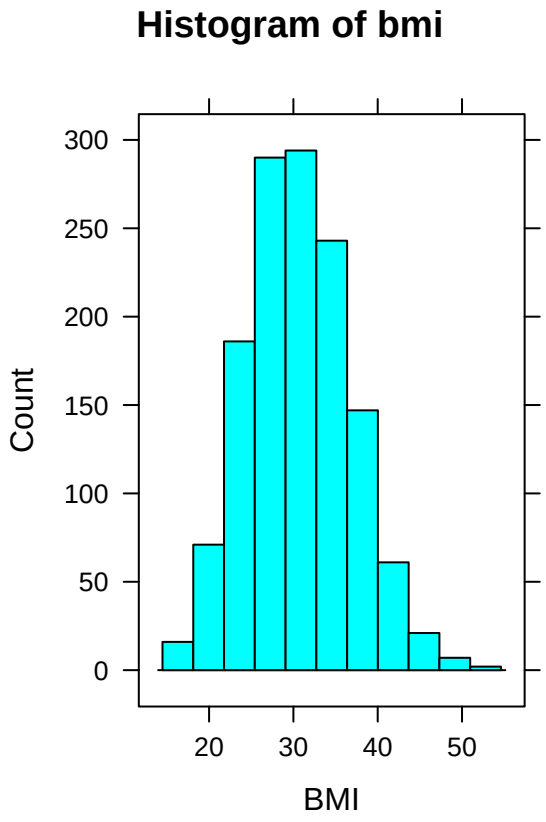
1. There is no strong collinearity between covariates.
2. There is some correlation between charges and age.
3. Also there is correlation between charges and bmi.

Bar plots of sex, smoker, region, no of children

1. Smokers are less in numbers than non-smokers
2. Most of the population seems to have either no or 1-2 children

Histogram of age and bmi



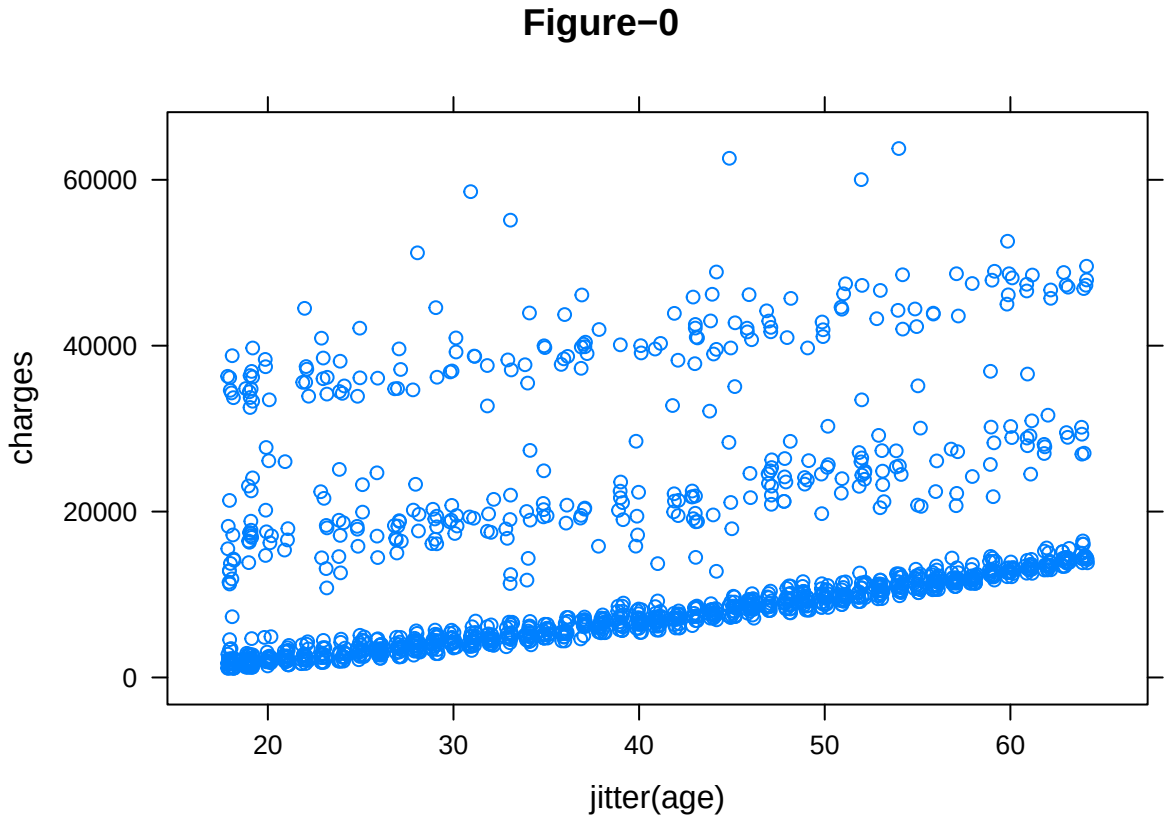


3.2 Creating dummy variable for BMI_Category

- BMI <=25 : Normal
- BMI > 25 but <= 30 : Overweight
- BMI > 30 : Obese

```
##   age    sex    bmi children smoker    region   charges BMI_Category    smoking
## 1  19 female 27.900        0    yes southwest 16884.924   Overweight     smoker
## 2  18   male 33.770        1     no southeast  1725.552        Obese non-smoker
## 3  28   male 33.000        3     no southeast  4449.462        Obese non-smoker
## 4  33   male 22.705        0     no northwest 21984.471       Normal non-smoker
## 5  32   male 28.880        0     no northwest  3866.855   Overweight non-smoker
## 6  31 female 25.740        0     no southeast  3756.622   Overweight non-smoker
```

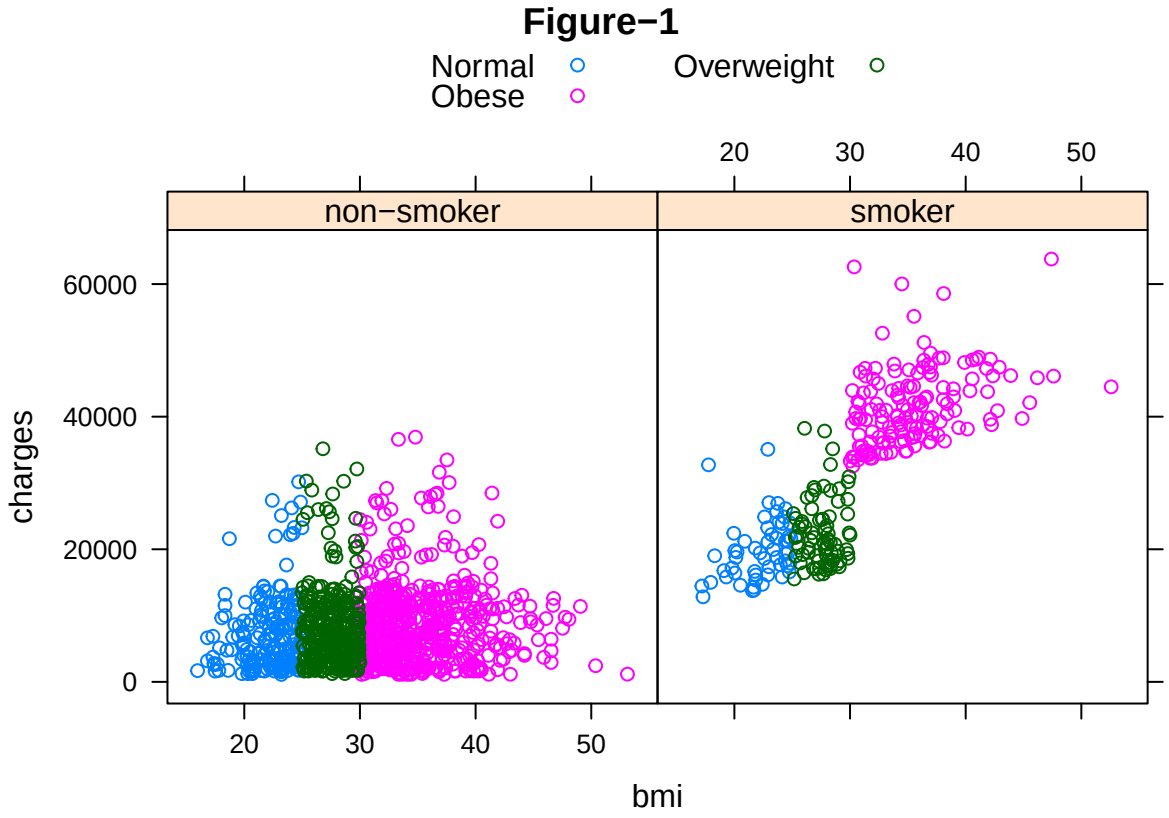
3.3 Plot of charge vs age



3.4 First step to build prelimnary model

- Since there is linear relationship with some factors. For now we can assume model be like
- lm(charges ~ 1 + age + some factors)

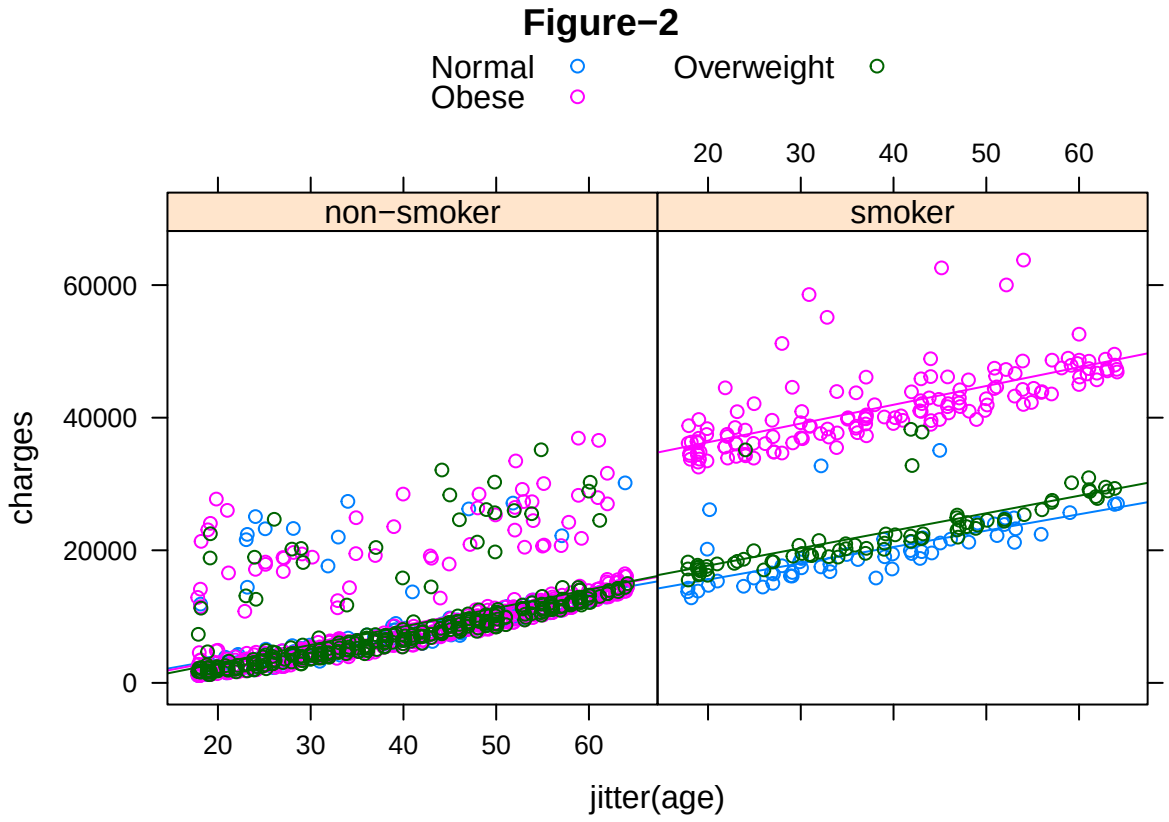
3.5 Plot of charge vs BMI across smoking



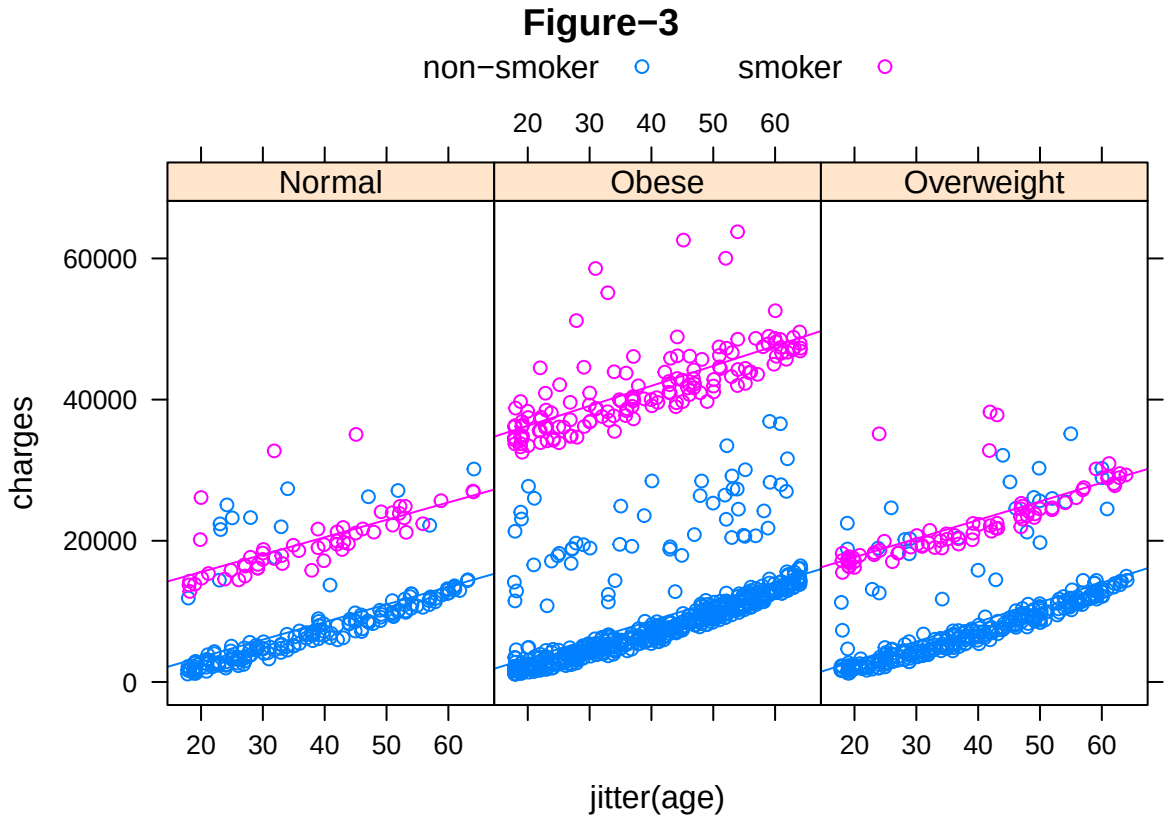
3.6 Explanation

- For non-smoker charges are not increasing due to BMI
- If a person is smoker then there is some fixed penalty and..
- if that person has obesity then the penalty is even more.
- So, charge is independent of BMI for non-smokers but not for smokers.
- So, we must include the interaction term for BMI_category and smoker in the model.

3.7 Plot of charge vs age across smoking



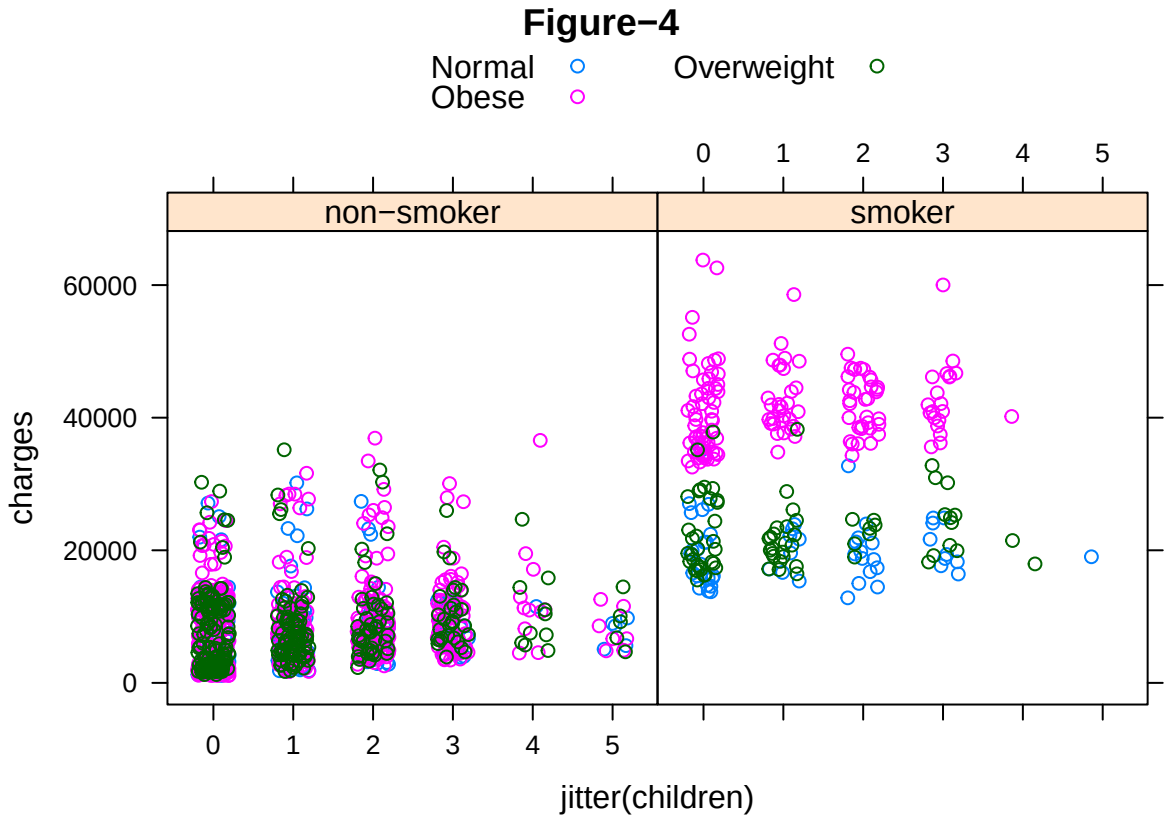
3.8 Plot of charge vs age across BMI_Category



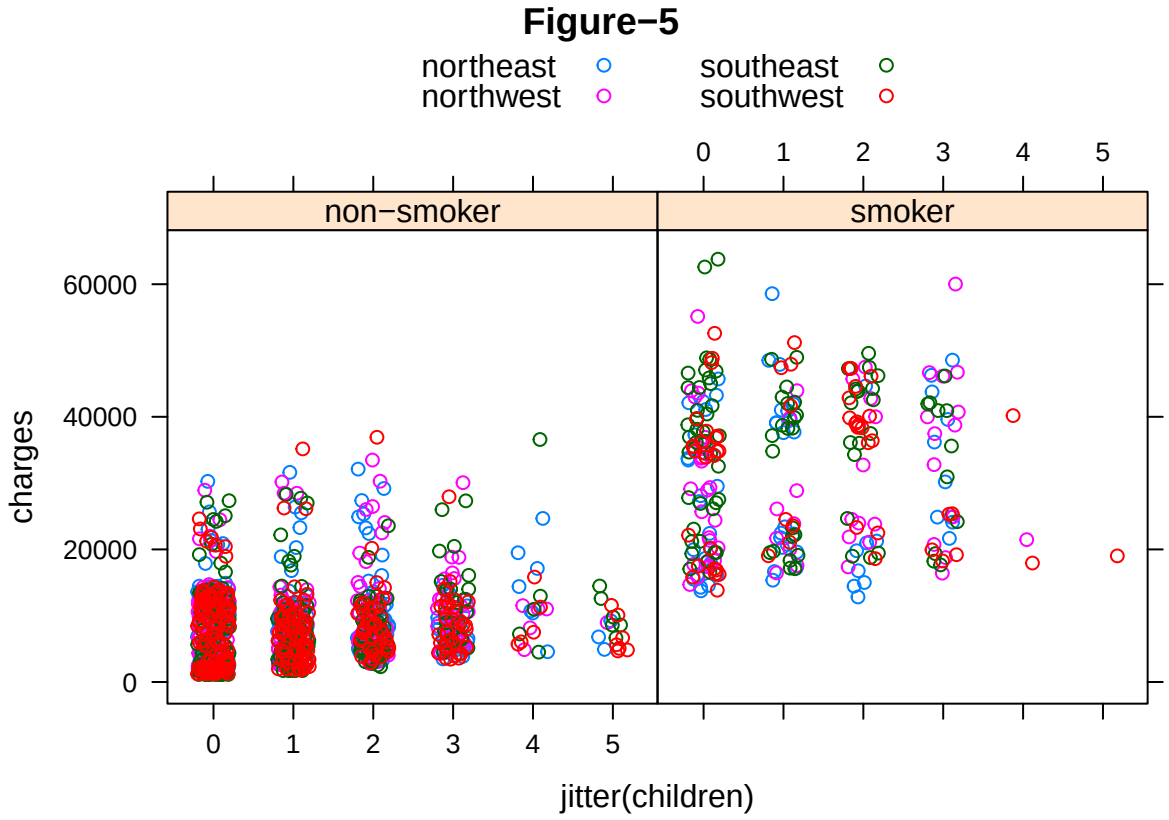
3.9 Updated model

- There is strong linear relationship between age and charges for all categories (BMI and smoking category)
- Now, we are sure smoking and BMI category are responsible factors for high charges.
- Our updated model is...
  lm(charges ~ 1 + age + BMI_Category*smoker*)
  *where BMI_Category*smoker means BMI_Category + smoker + BMI_Category:smoker

3.10 Plot of charge vs children across smoking



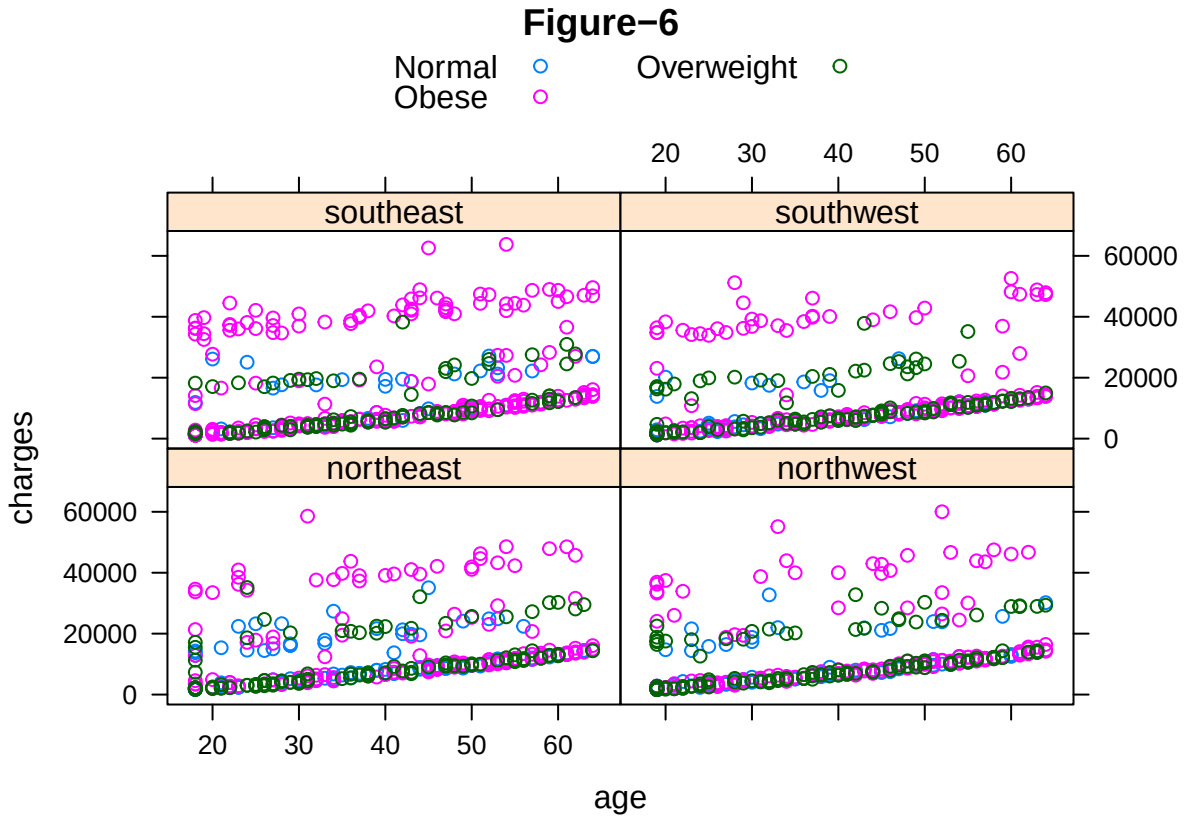
3.11 Plot of charge vs children across smoking



3.12 Updated model

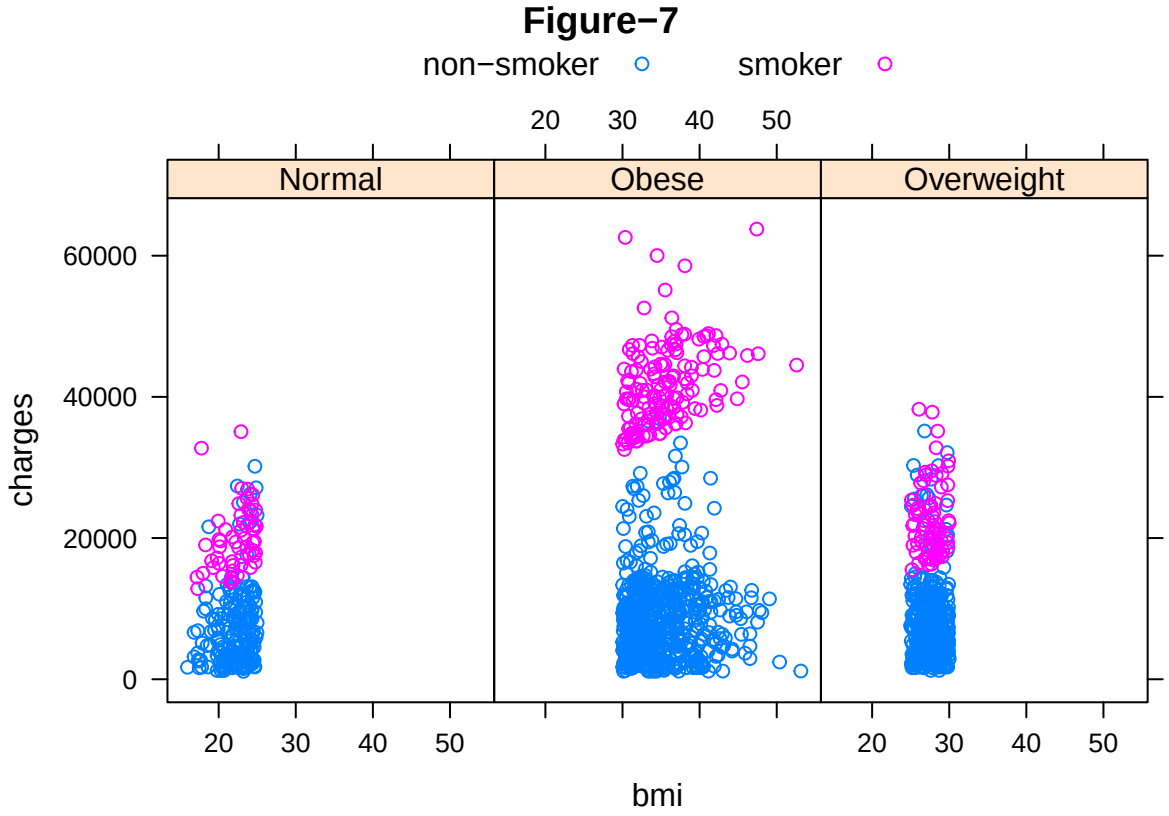
- There is a very slight increment in charge as no of children for non-smoker.
- Assuming the increment is linear wrt no of children, our updated model is... lm(charges ~ 1 + age + BMI_Category*smoker + children)
- Later we will see whether adding children covariate is significant or not..

3.13 Plot of charge vs BMI across regions



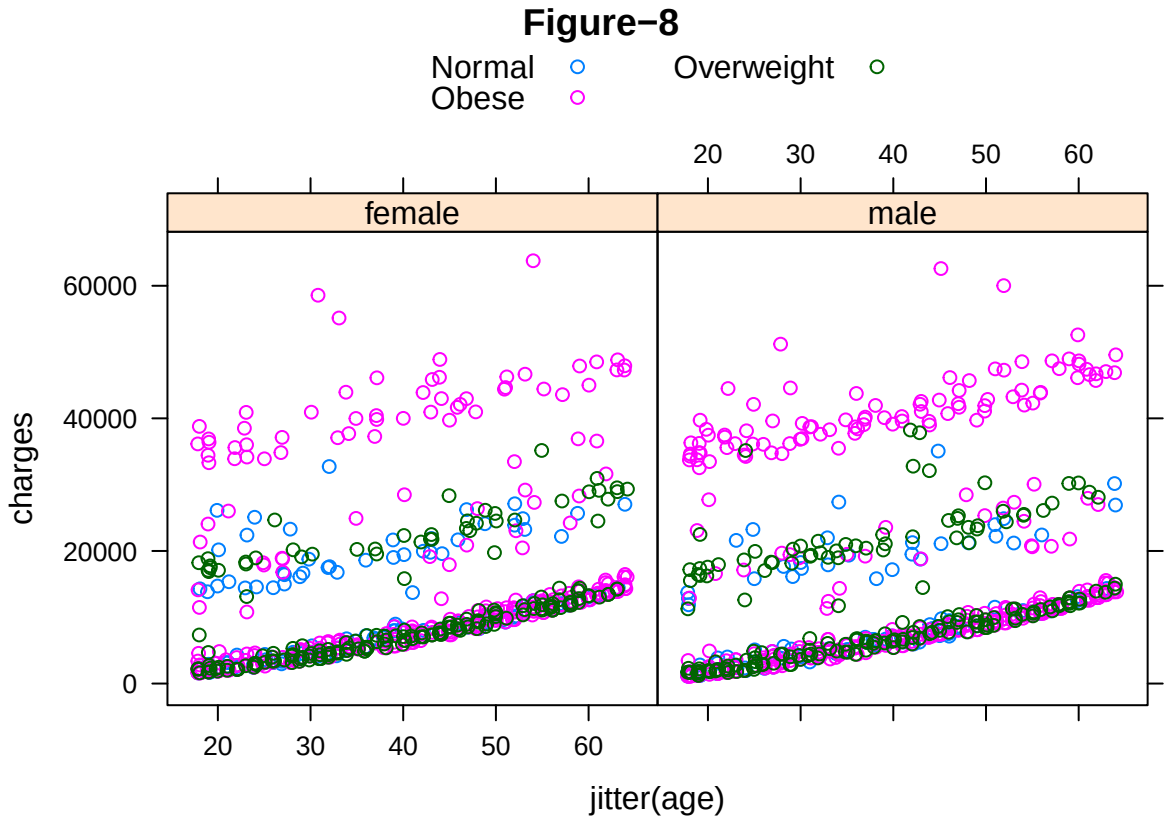
- More or less same pattern for all regions _ So, we won't include region factor in our model

3.14 Plot of charge vs BMI across BMI_Category



- If a smoker has obesity then he has to pay approx. twice penalty that of a normal smoker should pay.
- Obesity factor is already included in our model

3.15 Plot of charge vs age across smoking



- Sex hasn't any effect on the charge, So better to not include it in the model

3.16 Preliminary Model

- After all exploratory data analysis we come to the our preliminary model
- lm(charges ~ 1 + age + BMI_Category*smoker + children)

3.17 Dividing the data into training and testing data

```
set.seed(seed = 1001)
n_train <- round(0.8 * nrow(insurance.df))
train_indices <- sample(1:nrow(insurance.df), n_train)
df_train <- insurance.df[train_indices, ]
df_test <- insurance.df[-train_indices, ]
```

- Divided the data in 80-20% for training and testing.

3.18 Preliminary Model

```
##
## Call:
## lm(formula = charges ~ age + BMI_Category * smoker + children,
##     data = df_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -4501  -1929  -1354   -622  24361
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2427.58     529.74  -4.583 5.14e-06 ***
## age                             264.26      10.01  26.390  < 2e-16 ***
## BMI_CategoryObese               -40.64     436.53  -0.093   0.9259
## BMI_CategoryOverweight         -126.52     475.32  -0.266   0.7902
## smoker                        11862.95     780.51  15.199  < 2e-16 ***
## children                        546.52     115.45   4.734 2.50e-06 ***
## BMI_CategoryObese:smoker      21384.37     919.34  23.261  < 2e-16 ***
## BMI_CategoryOverweight:smoker  2491.82    1013.19   2.459   0.0141 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4582 on 1062 degrees of freedom
## Multiple R-squared:  0.8546, Adjusted R-squared:  0.8537
## F-statistic:   892 on 7 and 1062 DF,  p-value: < 2.2e-16
```

3.19 Full Model

```
##
## Call:
## lm(formula = charges ~ age + sex + bmi + children + smoker +
##     region, data = df_train)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11012  -2827  -1042   1322  30137
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11650.05    1116.69 -10.433  < 2e-16 ***
## age                257.84      13.45  19.177  < 2e-16 ***
## sexmale             10.06     377.93   0.027 0.978775
```

```
## bmi                332.39      32.42  10.252  < 2e-16 ***
## children           534.95     155.06   3.450 0.000583 ***
## smoker           23469.91     468.82  50.061  < 2e-16 ***
## regionnorthwest   -755.53     534.09  -1.415 0.157475
## regionsoutheast  -1068.53     545.00  -1.961 0.050188 .
## regionsouthwest  -1273.40     544.62  -2.338 0.019564 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6144 on 1061 degrees of freedom
## Multiple R-squared:  0.7388, Adjusted R-squared:  0.7369
## F-statistic: 375.2 on 8 and 1061 DF,  p-value: < 2.2e-16
```

3.20 Comaparing both above models on test data using MSE and MAD

```
sqrt_MSE_fm0 = sqrt(mean((df_test$charges- predict(fm0, df_test)) ^ 2))
sqrt_MSE_fm1 = sqrt(mean((df_test$charges- predict(fm1, df_test)) ^ 2))
MAD_fm0 = median(abs(df_test$charges- predict(fm0, df_test)))
MAD_fm1 = median(abs(df_test$charges- predict(fm1, df_test)))
sqrt_MSE_fm0
```

```
## [1] 4035.09
```

```
sqrt_MSE_fm1
```

```
## [1] 5750.597
```

```
MAD_fm0
```

```
## [1] 1664.799
```

```
MAD_fm1
```

```
## [1] 2606.482
```

- sqrt_MSE_fm0 = 4035.1 < 5750.6 = sqrt_MSE_fm1
- MAD_fm0 =1664.8 < 2606.48 = MAD_fm1
- Hence, our preliminary model is better so far.

3.21 Comparison with model without children covariate

```
fm2 <- lm(charges ~ age + BMI_Category * smoker, data = df_train)

sqrt_MSE_fm2 = sqrt(mean((df_test$charges- predict(fm2, df_test)) ^ 2))
```

```
MAD_fm2= median(abs(df_test$charges- predict(fm2, df_test)))

sqrt_MSE_fm0
```

```
## [1] 4035.09
```

```
sqrt_MSE_fm2
```

```
## [1] 4050.444
```

```
MAD_fm0
```

```
## [1] 1664.799
```

```
MAD_fm2
```

```
## [1] 1677.507
```

3.22 Best multiple linear model based on observation

- There is very less difference in sqrt(MSE) or MAD between both models.
- So, it's better to drop one one covariate.
- It also shows that no of children has almost no linear impact on the insurance.
- Our model is get reduced to fm2.
- No further reduction of linear model fm2 is possible as if we drop age also then..

```
fm3 <- lm(charges ~ BMI_Category * smoker, data = df_train)
sqrt_MSE_fm3=  sqrt(mean((df_test$charges- predict(fm3, df_test)) ^ 2))
MAD_fm3= median(abs(df_test$charges- predict(fm3, df_test)))
sqrt_MSE_fm3
```

```
## [1] 5493.814
```

```
MAD_fm3
```

```
## [1] 3618.586
```

- much higher MSE and MAD than fm2

3.23 Using AIC to find the best model

- We will try to improve our model by adding more interaction terms if possible.
- We will include all interaction terms of 2nd order and find best model using AIC Model selection

```
Fitfirst=lm(charges~1,data=df_train)
Fitall=lm(charges~.^2,data=df_train) #2nd order interaction terms
```

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

3.24 BEST MODEL USING AIC

```
summary(AIC_Model)
```

```
##
## Call:
## lm(formula = charges ~ smoker + age + BMI_Category + children +
##     region + bmi + smoker:BMI_Category + smoker:bmi + age:BMI_Category,
##     data = df_train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2879.7 -1822.9 -1297.5  -627.7 24466.1
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -1471.15    1444.01  -1.019 0.308535
## smoker                         1576.18    2476.21   0.637 0.524570
## age                             242.71      24.42   9.941  < 2e-16 ***
## BMI_CategoryObese              -950.93    1280.43  -0.743 0.457847
## BMI_CategoryOverweight        -1141.68    1276.12  -0.895 0.371181
## children                        570.32     114.15   4.996 6.84e-07 ***
## regionnorthwest                -534.54     394.95  -1.353 0.176201
## regionsoutheast                -777.49     403.28  -1.928 0.054136 .
## regionsouthwest               -1375.24     402.83  -3.414 0.000665 ***
## bmi                              16.08      49.51   0.325 0.745371
## smoker:BMI_CategoryObese      15417.96    1670.59   9.229  < 2e-16 ***
## smoker:BMI_CategoryOverweight   174.93    1155.02   0.151 0.879647
## smoker:bmi                      457.98     105.98   4.321 1.70e-05 ***
## age:BMI_CategoryObese            23.27      27.75   0.838 0.401976
## age:BMI_CategoryOverweight       28.00      30.90   0.906 0.365207
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4519 on 1055 degrees of freedom
## Multiple R-squared:  0.8595, Adjusted R-squared:  0.8577
## F-statistic: 461.1 on 14 and 1055 DF,  p-value: < 2.2e-16
```

3.25 BEST SELECTED MODEL USING AIC

- StepAIC_model= lm(formula = charges ~ smoker + age + BMI_Category + children + region + bmi + smoker:BMI_Category + smoker:bmi + age:BMI_Category, data = df_train)

```
sqrt_MSE_fm_AIC=  sqrt(mean((df_test$charges- predict(AIC_Model, df_test)) ^ 2))
MAD_fm_AIC= median(abs(df_test$charges- predict(AIC_Model, df_test)))
sqrt_MSE_fm_AIC
```

```
## [1] 3966.251
```

```
MAD_fm_AIC
```

```
## [1] 1436.912
```

```
sqrt_MSE_fm2
```

```
## [1] 4050.444
```

```
MAD_fm2
```

```
## [1] 1677.507
```

3.26 Deleting statistically insignificant covariates from the AIC Model

```
Final_Model = lm(formula = charges ~ smoker + age  + children +
    region  + smoker:BMI_Category + smoker:bmi,
    data = df_train)
sqrt_MSE_FM=  sqrt(mean((df_test$charges- predict(Final_Model, df_test)) ^ 2))
MAD_FM= median(abs(df_test$charges- predict(Final_Model, df_test)))
sqrt_MSE_FM
```

```
## [1] 3974.228
```

```
MAD_FM
```

```
## [1] 1447.452
```

3.27 Comparision between the AIC model and Final Model Using F-test

```
anova(Final_Model,AIC_Model)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: charges ~ smoker + age + children + region + smoker:BMI_Category +
##     smoker:bmi
## Model 2: charges ~ smoker + age + BMI_Category + children + region + bmi +
##     smoker:BMI_Category + smoker:bmi + age:BMI_Category
##   Res.Df        RSS Df Sum of Sq      F Pr(>F)
## 1   1060 2.1568e+10
## 2   1055 2.1545e+10  5  23193637 0.2271 0.9508
```

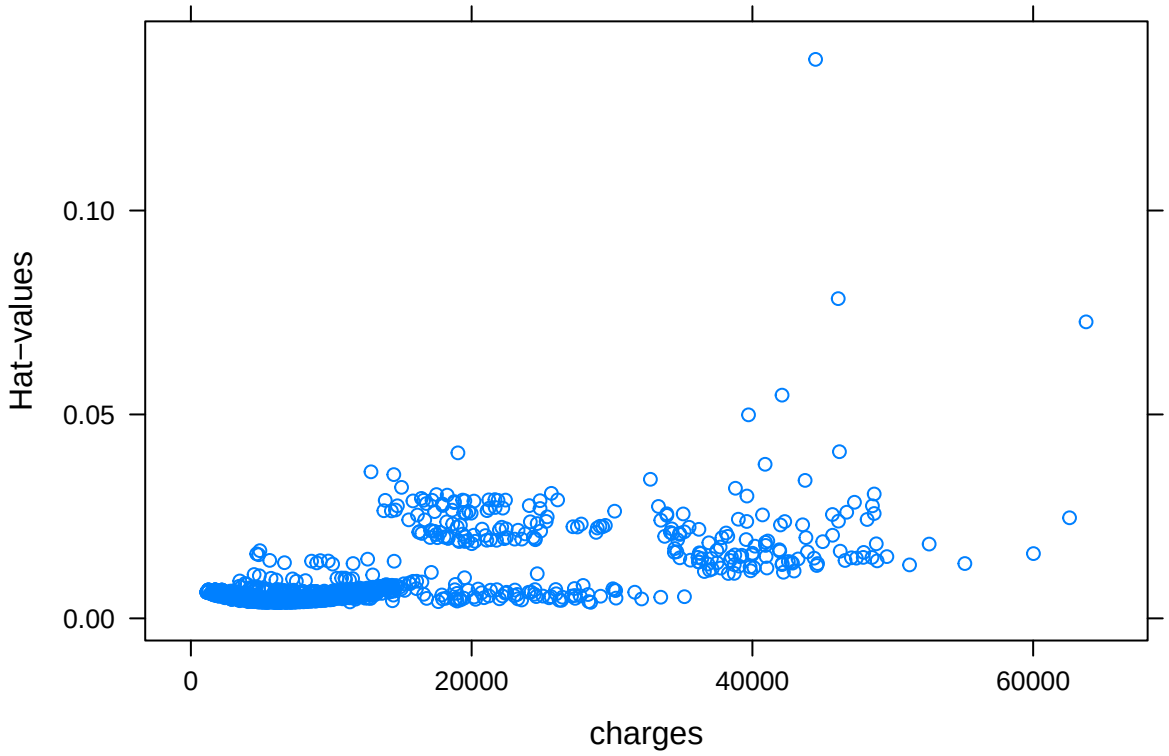
- The P-value of the F test is 0.64»0.05.
- i.e. we failed to reject the null hypothesis.
- So, both models are statistically equivalent but the Final Model has lesser no covariates
- Also LSE_FM < LSE_AIC.

3.28 Checking required assumption for multiple linear regression

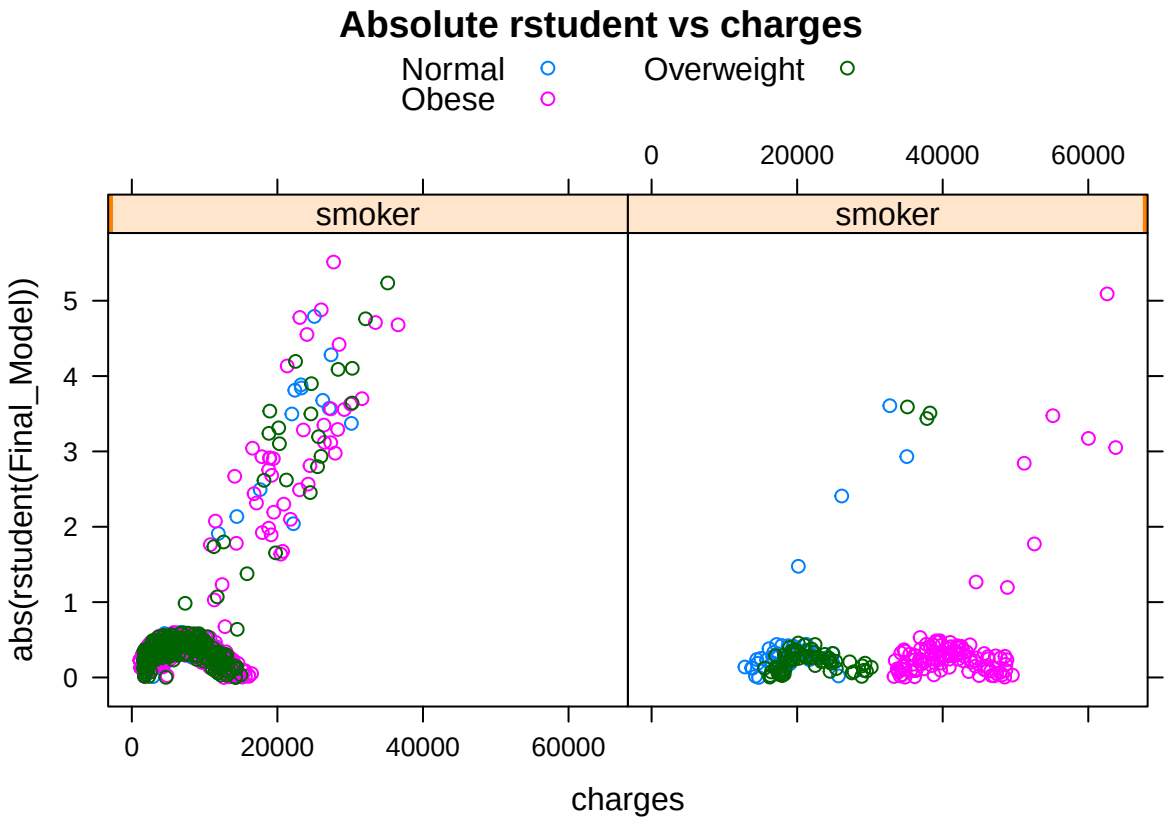
- Checking for Outliers
- Normality of error (residuals)
- Non-constant variance of error (residuals)
- Possible remedies

3.29 Checking for Outliers

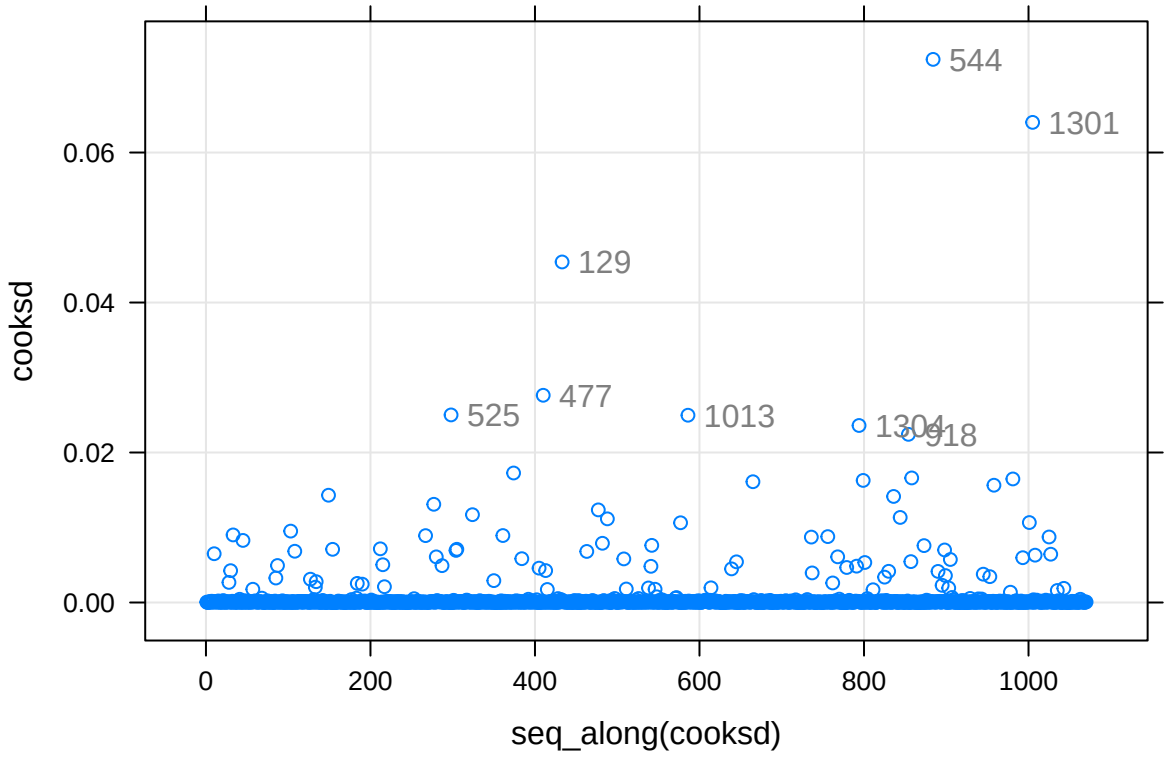
Plot of leverage vs response variable



abs(studententized residual) v/s response variable plot




```
dfb <- dfbetas(Final_Model); cooksd <- cooks.distance(Final_Model)
id <- cooksd > 0.018
xyplot(cooksd ~ seq_along(cooksd), grid = TRUE) +
layer(panel.text(x[id], y[id], labels = rownames(df_train)[id], pos = 4, col = "grey50")
```



3.30 No of outliers (on the basis of Cook distance cut-off)

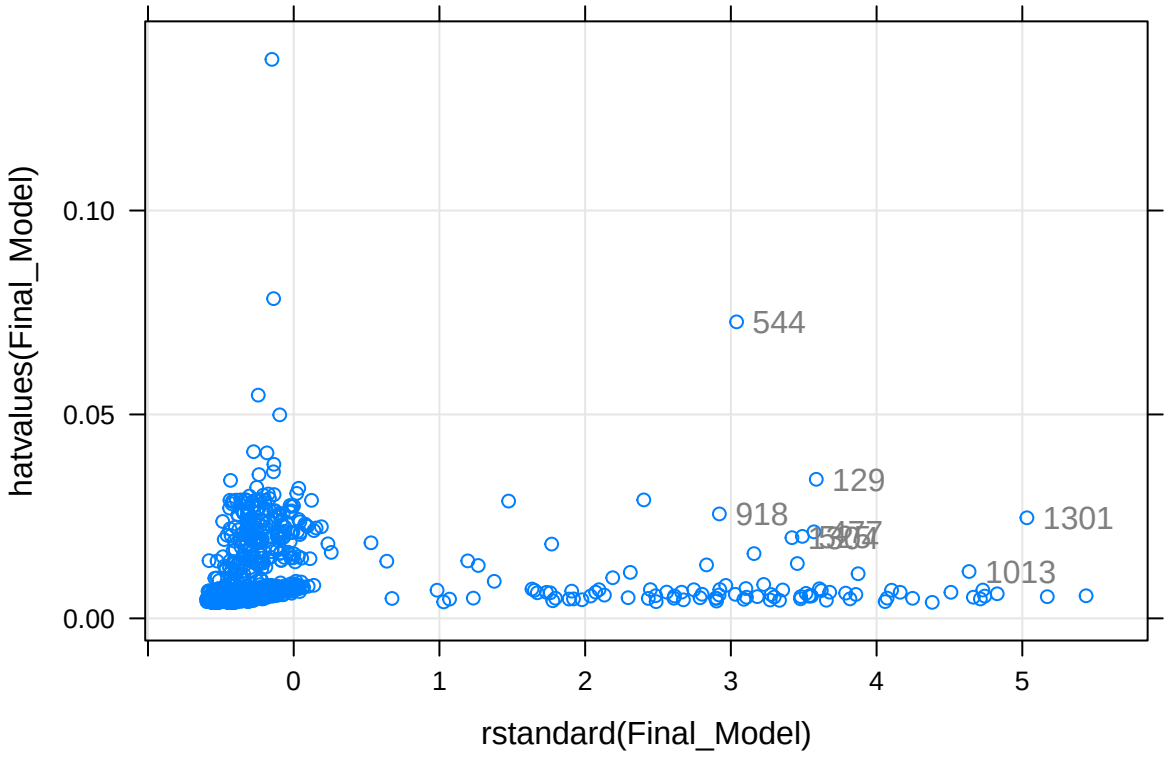
```
outliers=names(which(cooksd > 4/(nrow(df_train)-length(Final_Model$coefficients))))
length(outliers)
```

```
## [1] 67
```

- Obviously we can not exclude all the outliers
- The outliers are nothing but the insurance charges which could not be explined by any of the given covariates in the data.

3.31 Laverage vs Std. Residual Plot

```
xyplot(hatvalues(Final_Model) ~ rstandard(Final_Model), grid = TRUE)+
layer(panel.text(x[id], y[id], labels = rownames(df_train)[id], pos = 4, col = "grey50"))
```



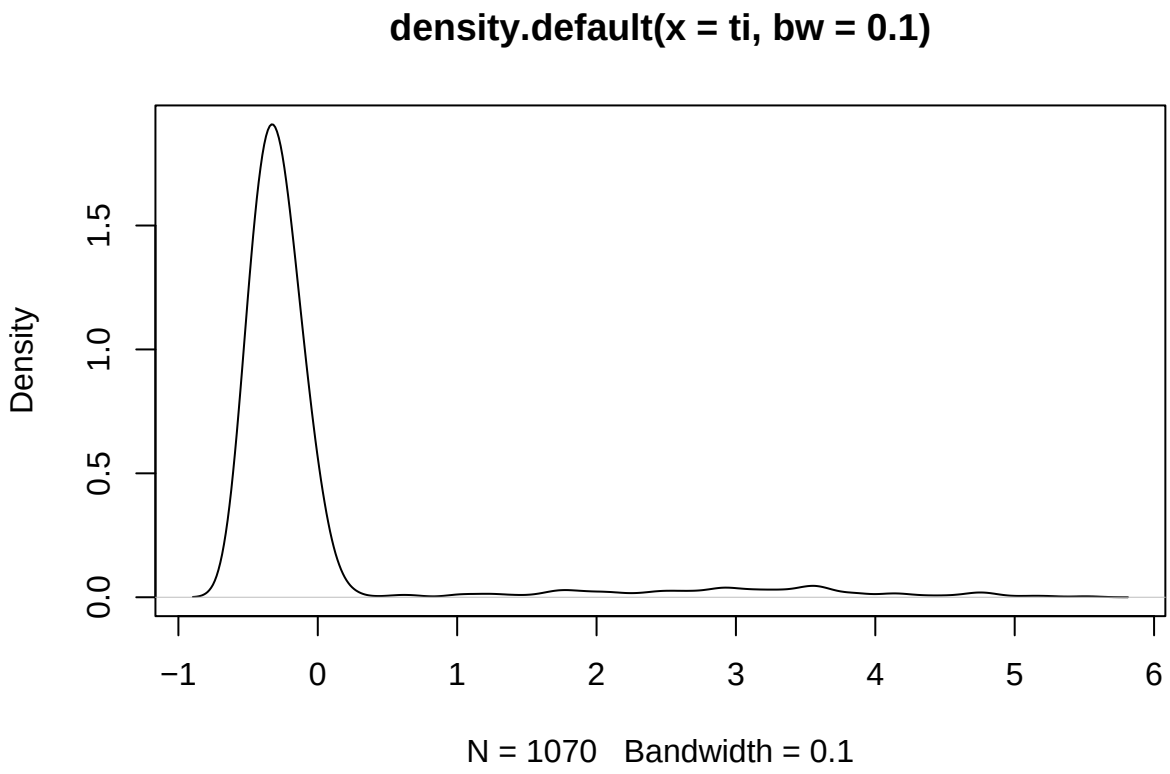
3.32 Checking for normality or residuals

qqplot for studentized residuals over t distribution (df=n-p-1)

- There are 99 such observations whose ti > 0.3

Density plot studentized residual

```
plot(density(ti, bw = 0.1))
```



- One sided heavy tailed
- Not good for the linear regression model

Shaipro-Wilk and KS Test for Non-Normality

```
e.FM<- rstudent(Final_Model)
shapiro.test(e.FM) #failed
```

```
##
##  Shapiro-Wilk normality test
##
## data:  e.FM
## W = 0.47298, p-value < 2.2e-16
```

```
ks.test(e.FM, pnorm) #failed
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  e.FM
## D = 0.3756, p-value < 2.2e-16
```

```
## alternative hypothesis: two-sided
```

- Both test suggest that There is no normality in the density plot of residuals

```
n <- with(Final_Model, rank + df.residual)
names.id = as.numeric(names(which(ti > 0.3)))
length(names.id)
```

```
## [1] 99
```

```
insurance_new = insurance.df[-names.id,]

fm_new <- lm(formula = charges ~ age + children + smoker:BMI_Category +
    smoker:bmi, data = insurance_new)
```

- Checking for non-constant variance

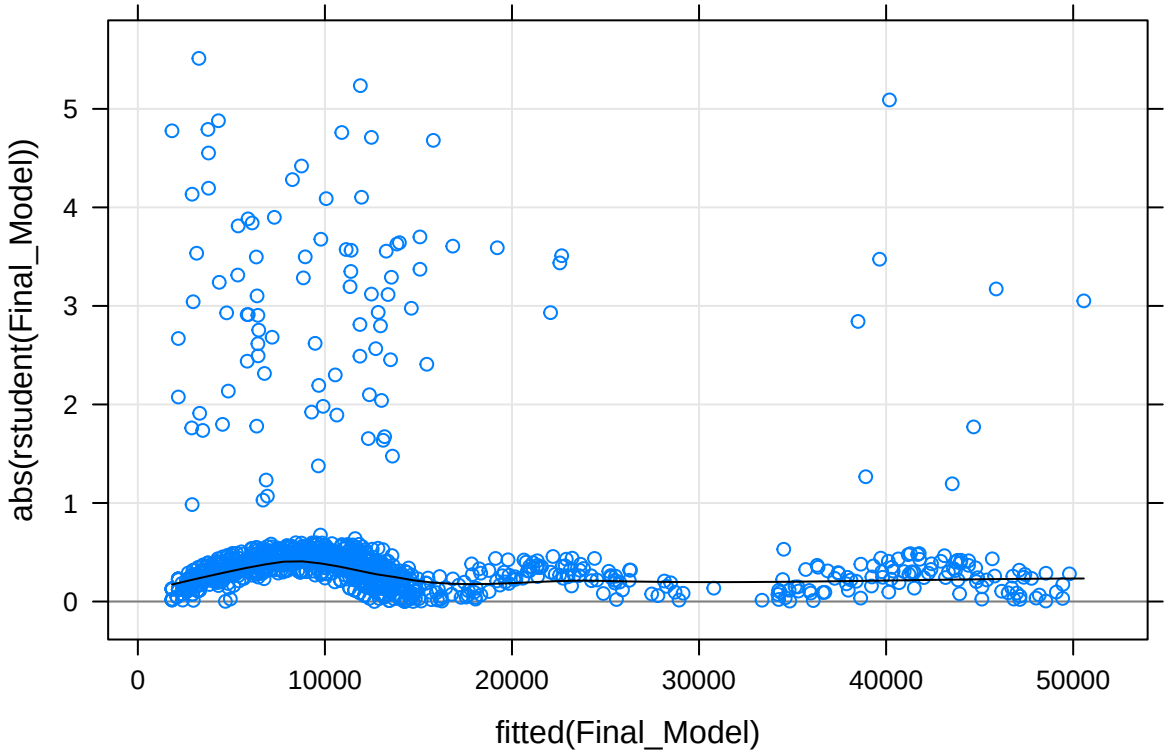
```
library(car)
```

```
## Loading required package: carData
```

```
ncvTest(Final_Model)
```

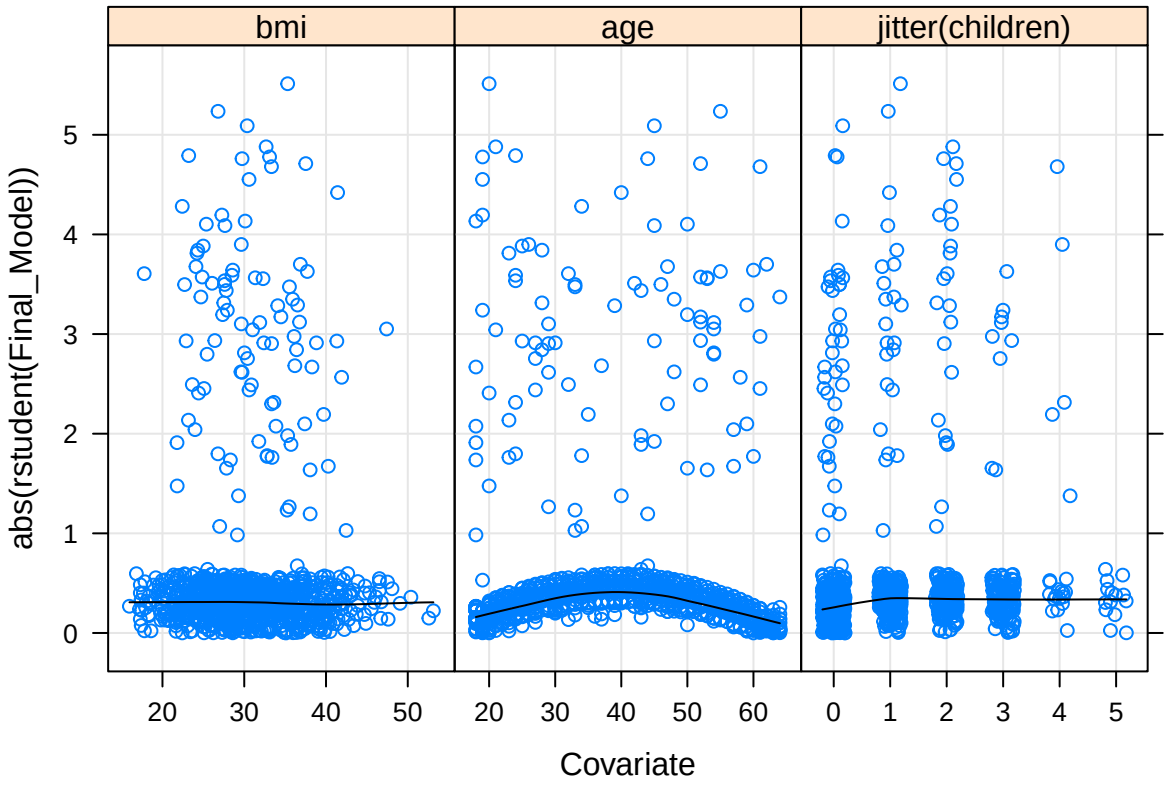
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 9.868574, Df = 1, p = 0.0016813
```

3.32.1 Plot of absolute studentized residual vs fitted value



-It can be seen the cause of non-constant variance is big no of outliers

3.33 Plot of residuals for different covariates



- residual plot is more or less constant if we ignore the outliers
- If we consider model very first model fm0 then we can get rid of the non constant variance

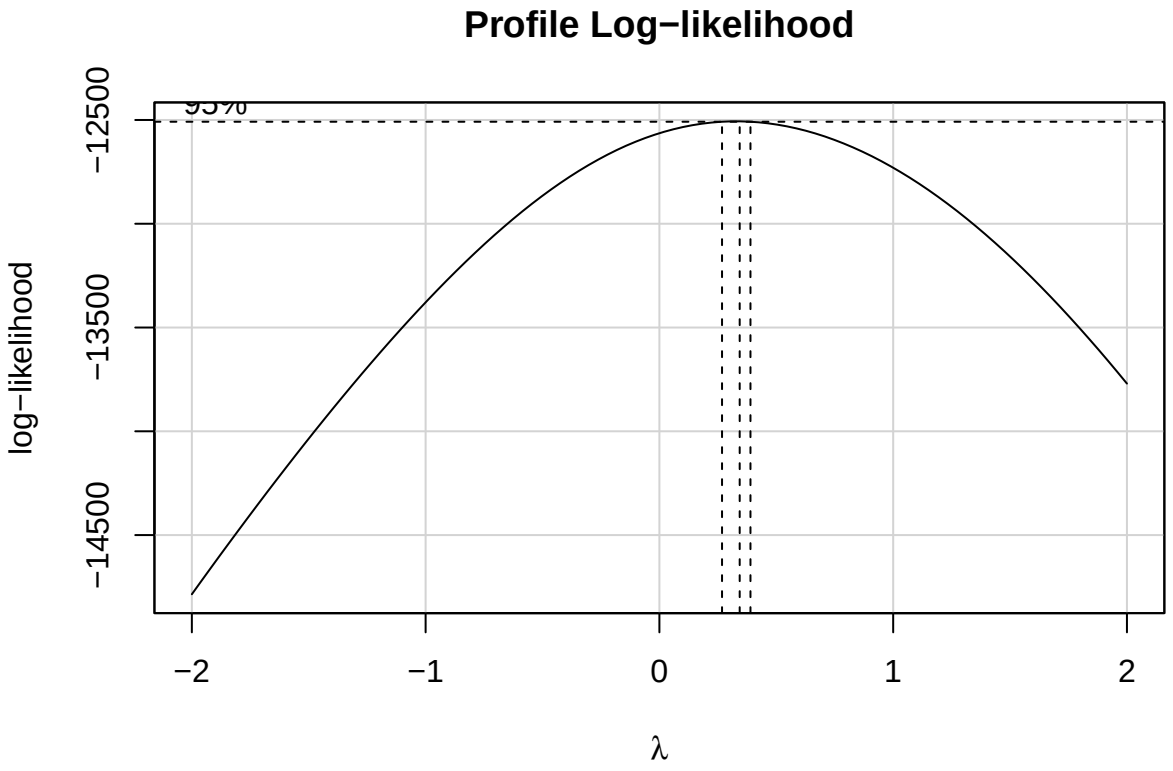
```
fm0 <- lm(formula = charges ~ age + children + BMI_Category + smoker*bmi,
    data = df_train)
ncvTest(fm0) #Test suggest there is variance of residuals is constant
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.466175, Df = 1, p = 0.22595
```

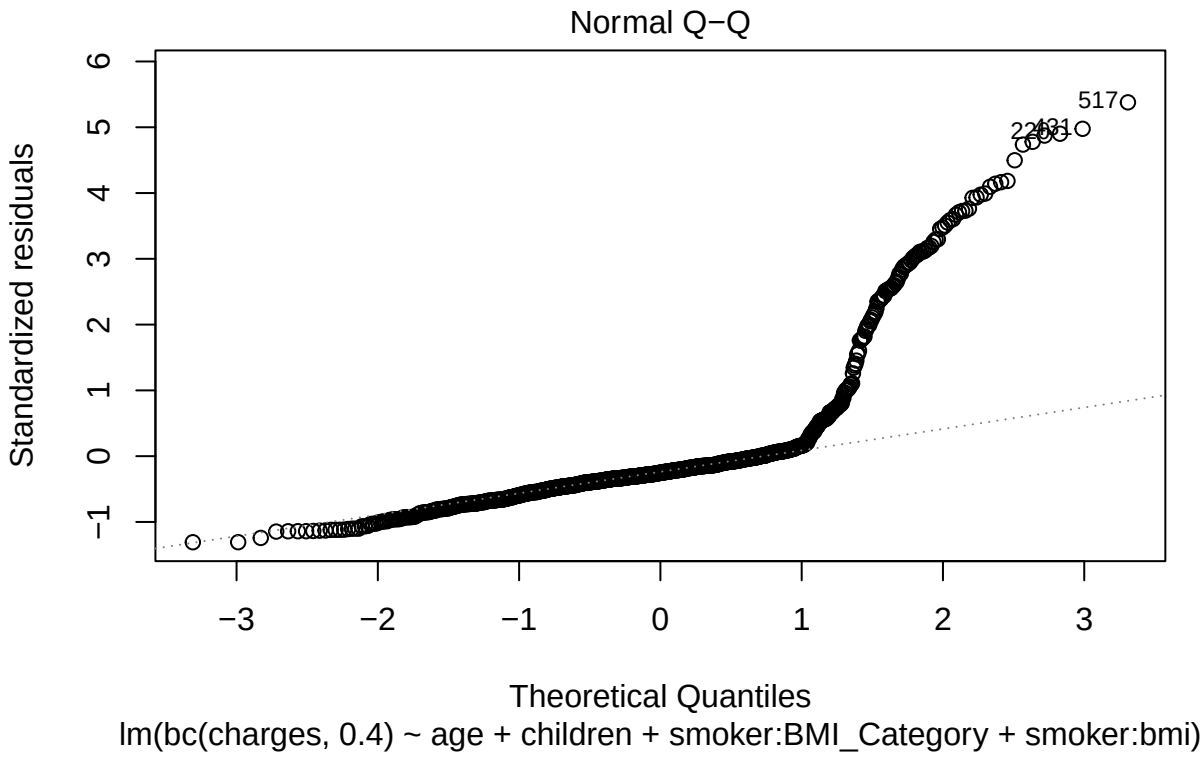
- But problem of non-normality of residuals still exist.
- Now we'll look for possible remedies

3.34 BOX-COX transformation

```
boxCox(Final_Model)
```




```
bc <- function(x, lambda) { if (lambda == 0) log(x) else (x^lambda - 1) / lambda }
fm_box <- lm(bc(charges,0.4)~age + children + smoker:BMI_Category +
    smoker:bmi,data=df_train)
plot(fm_box,which=2)
```



- No improvement at all. - Now we'll go for Robust Regression

3.35 Robust Regression

LAD REGRESSION

```
library(quantreg)
```

```
## Loading required package: SparseM
```

```
##
## Attaching package: 'SparseM'
```

```
## The following object is masked from 'package:base':
##
##     backsolve
```

```
fm.lad<- rq(charges~ age+children+BMI_Category:smoker+smoker:bmi, data=df_train)
lad.yhat <- predict(fm.lad,df_test)
MAD_lad= median(abs(df_test$charges- lad.yhat))
MAD_lad
```

```
## [1] 556.333
```

3.36 Huber Loss Function

```
fm.huber<- rlm(charges~ age+children+BMI_Category:smoker+smoker:bmi, data=df_train, psi=
huber.yhat <- predict(fm.huber,df_test)

MAD_huber= median(abs(df_test$charges- predict(fm.huber, df_test)))
MAD_huber
```

```
## [1] 562.4914
```

```
MAD_FM
```

```
## [1] 1447.452
```

- MAD for Huber loss functio is very less as comparison to Least Square loss function

Bisquare Loss Function

```
fm.bsq<- rlm(charges~ age+children+BMI_Category:smoker+smoker:bmi, data=df_train, psi=ps
bsq.yhat <- predict(fm.bsq,df_test)

MAD_bsq= median(abs(df_test$charges- bsq.yhat))
MAD_bsq
```

```
## [1] 516.7187
```

```
MAD_huber
```

```
## [1] 562.4914
```

```
MAD_FM
```

```
## [1] 1447.452
```

- MAD(bsq) < MAD(huber) < MAD(Least Square)
- Applying Robust regression is a good idea for this type of data where the data has many outliers

3.37 Resistant Regression

Least Trimmed Square(LTS)

```
set.seed(seed = 1001)
fm.lts= ltsreg(charges~ age+children+BMI_Category:smoker+smoker:bmi,data=df_train)
lts.yhat <- predict(fm.lts,df_test)
```

```
MAD_lts= median(abs(df_test$charges- lts.yhat))
MAD_lts
```

```
## [1] 508.0683
```

3.38 Least Median of Squares (LMS)

```
set.seed(seed = 1001)
fm.lms= lmsreg(charges~ age+children+BMI_Category:smoker+smoker:bmi,data=df_train)
lms.yhat <- predict(fm.lms,df_test)
MAD_lms= median(abs(df_test$charges- lms.yhat))
MAD_lms
```

```
## [1] 505.4326
```

3.39 Comparison between Least Squares, LAD, Huber Loss, BSq Loss, LTS and LMS

```
MAD_FM
```

```
## [1] 1447.452
```

```
MAD_lad
```

```
## [1] 556.333
```

```
MAD_huber
```

```
## [1] 562.4914
```

```
MAD_bsq
```

```
## [1] 516.7187
```

```
MAD_lts
```

```
## [1] 508.0683
```

```
MAD_lms
```

```
## [1] 505.4326
```

1. Performance of Robust Regressoin is better than Least square regression when no of outliers in the data is high.

2. Among performed Robust Regressions Bisquare and Resistant regressions are good options.

3.40 Accuracy of the different models

1. Say a prediction is good if difference between predicted charge and actual response is less than 1000 dollor.

2. Define accuracy as proportion of good prediction on test data.

```
test_size = nrow(df_test)
Accuracy = function (fm){
difference =abs(predict(fm,df_test)-df_test$charges)
total = sum(ifelse(difference <= 1000,1,0))
return(total/test_size)
}
```

3.41 Accuracy of the different models

```
Accuracy(Final_Model)
```

```
## [1] 0.2761194
```

```
Accuracy(fm.lad)
```

```
## [1] 0.7985075
```

```
Accuracy(fm.huber)
```

```
## [1] 0.7985075
```

```
Accuracy(fm.bsq)
```

```
## [1] 0.8097015
```

```
Accuracy(fm.lms)
```

```
## [1] 0.7276119
```

```
Accuracy(fm.lts)
```

```
## [1] 0.7276119
```

Chapter 4

Result and conclusion

- There are many influential observation. This may be because of the fact that some information about the beneficiary is not given in data.
- Due to high no of influential observation LSE regression is not a good regression for prediction of insurance charges.
- Robust regression is very good option for prediction as it reduces the effect of outliers very much.
- Bisquare loss function and resistant regression are best options for robust regression under Robust regression if there are too many outliers.
- Health insurance charge is dependet upon age, smoking category, BMI_Category and no of children.