Health Insurance Premium Cost Prediction

Yash Gupta

2022-11-05

What is this project all about?

This is a data analysis project for predicting yearly health insurance premium charge based on given features. the features are given below-

- **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**

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

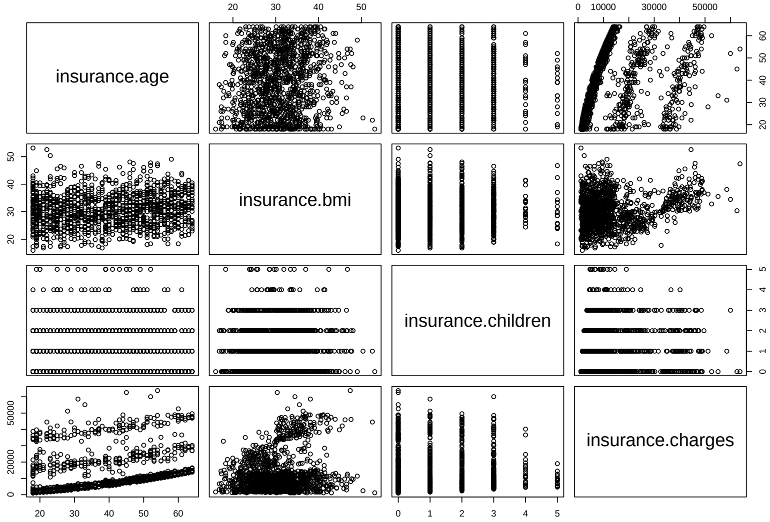
```
##       age            sex                 bmi           children    
##  Min.   :18.00   Length:1338        Min.   :15.96   Min.   :0.000  
##  1st Qu.:27.00   Class :character   1st Qu.:26.30   1st Qu.:0.000  
##  Median :39.00   Mode  :character   Median :30.40   Median :1.000  
##  Mean   :39.21                      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  
```

Exploratory Data Analysis

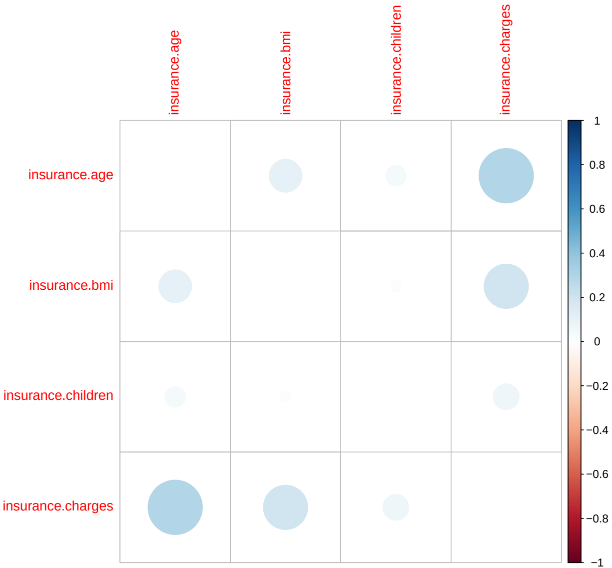
- Correlation

```
##                    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

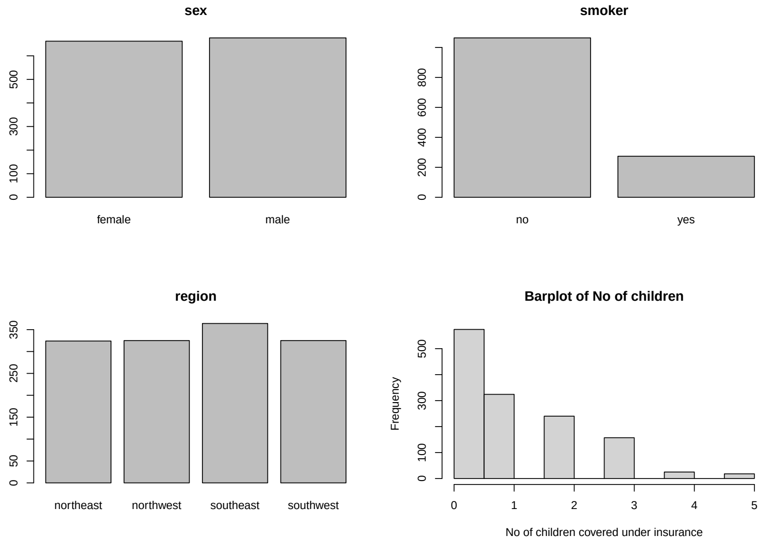


insurance.age
insurance.bmi
insurance.children
insurance.charges
insurance.age
insurance.bmi
insurance.children
insurance.charges
1
0.8
0.6
0.4
0.2
0
−0.2
−0.4
−0.6
−0.8
−1

Explanation

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

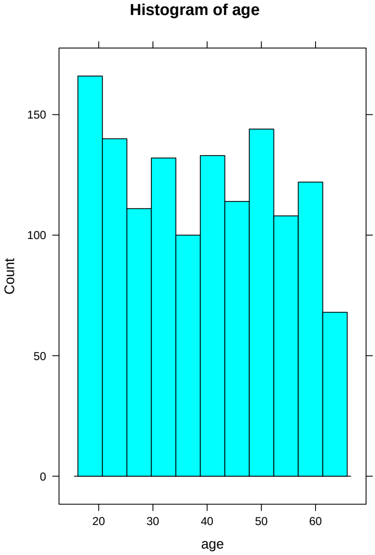
Bar plots of sex, smoker, region, no of children

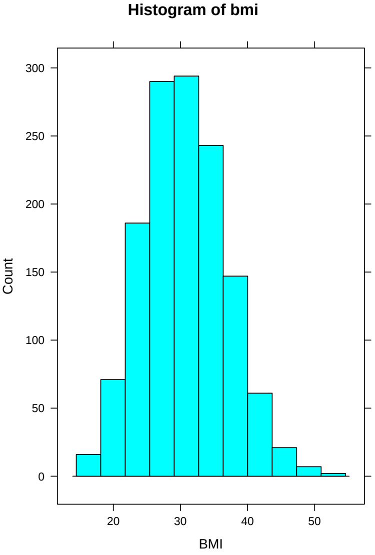


- Smokers are less in numbers than non-smokers - Most of the population seems to have either no or 1-2 children

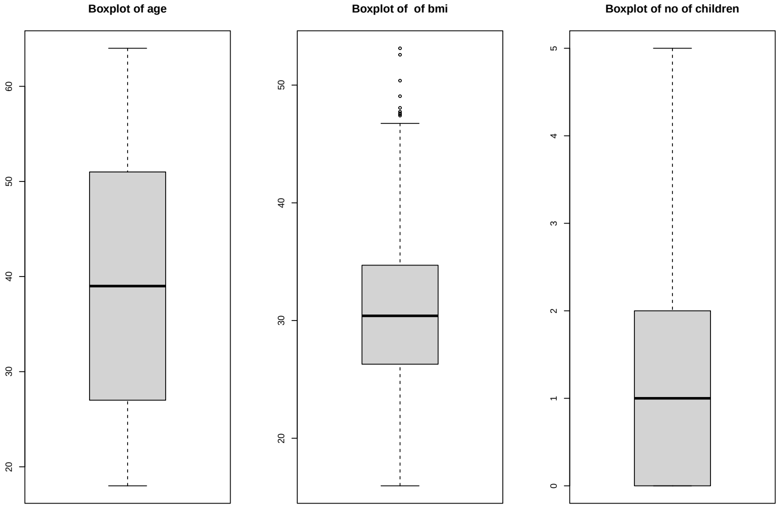
Histogram of age and bmi

```
## Warning: package 'gridExtra' was built under R version 4
```



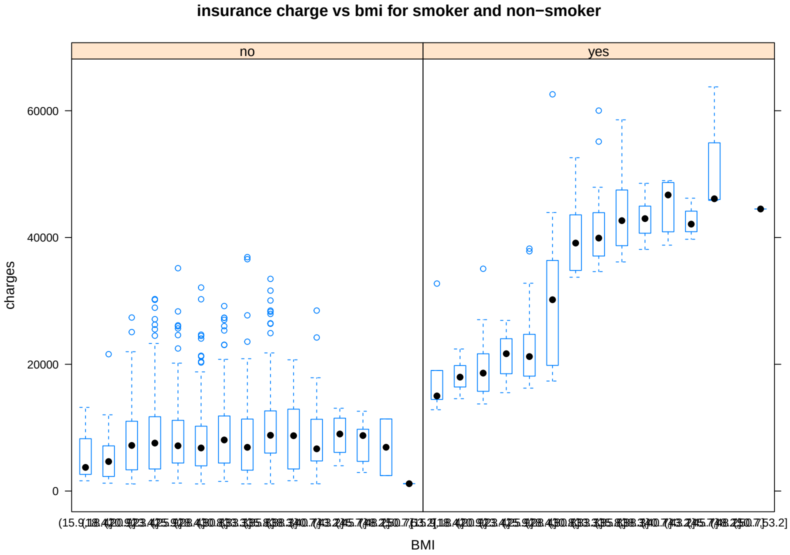


Boxplots of age, bmi and no of children



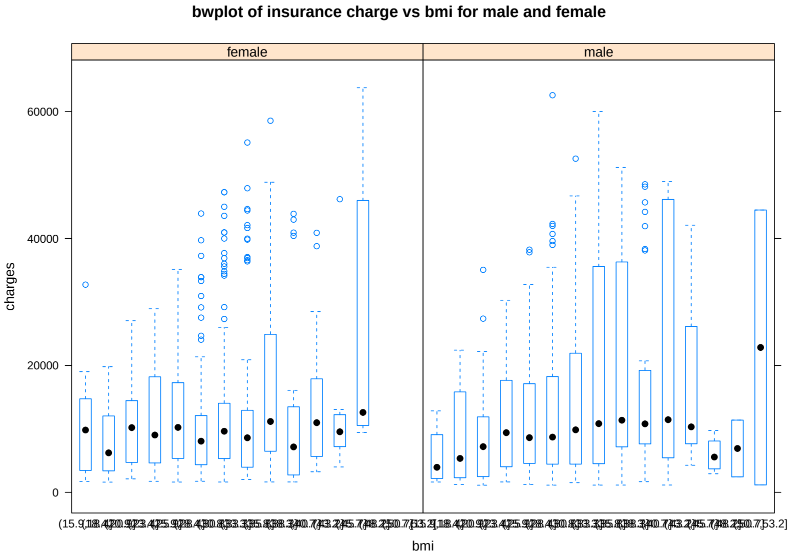
- 50% of the population seems to have obesity problem.

Box whisker plots of charges vs bmi



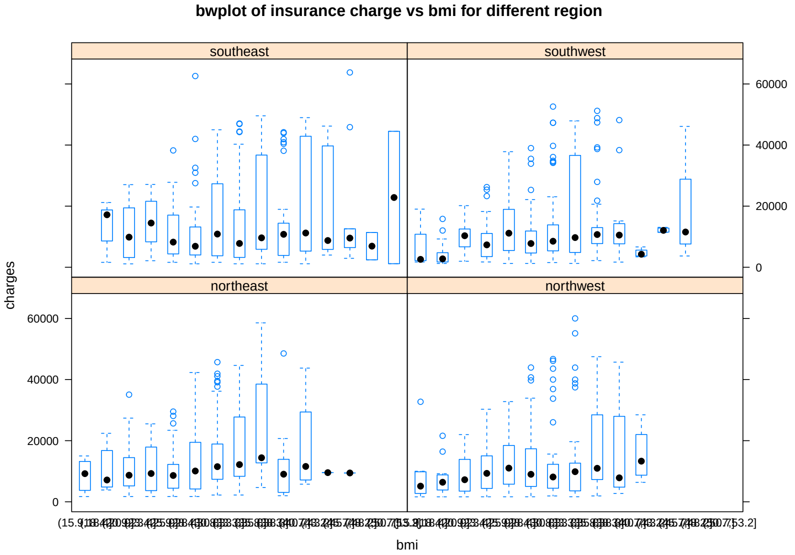
- It seems only smoker's have to bear more insurance cost due to increment in bmi.
- There is a sudden change in charges as bmi category changes to above 30. _ We shall add a dummy variable for different category of BMI and check further the change in the charge is significant or not.

Box whisker plots of charges vs bmi

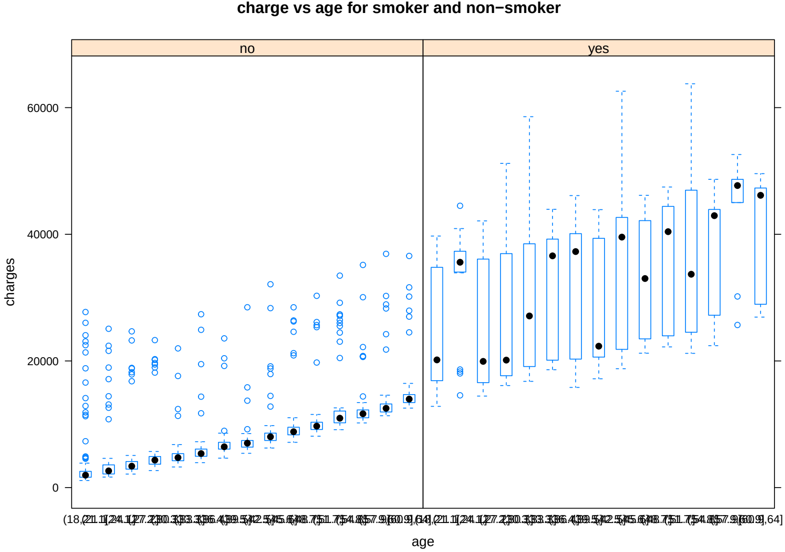


- Plots are more or less same for male and female.

Box whisker plots of charges vs bmi

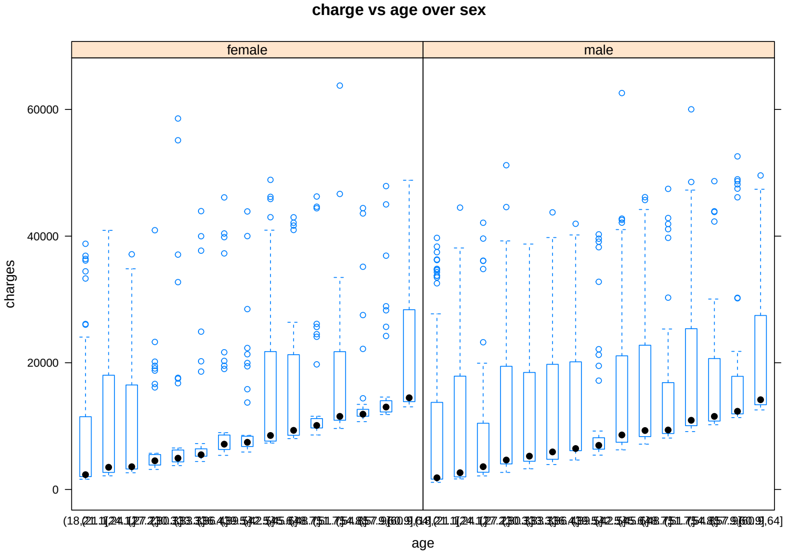


Box whisker plots of charges vs age across smoker category



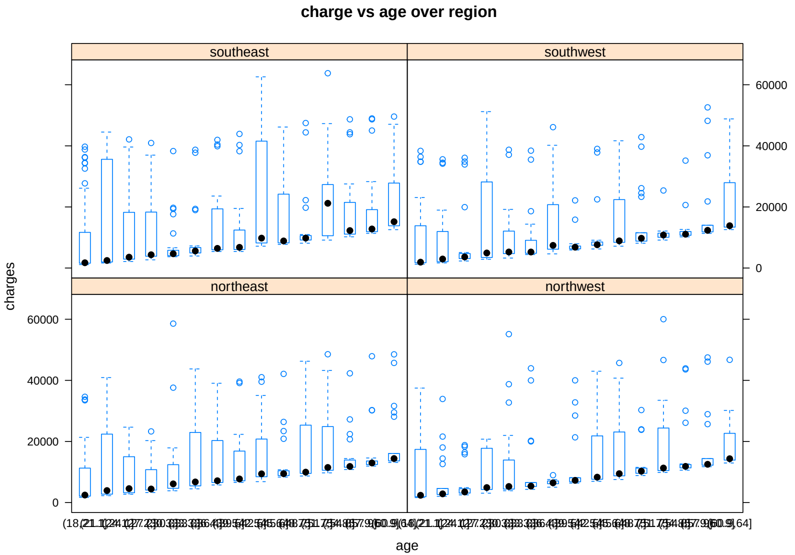
- Non-smoker category has a good number of outliers. - Smoker category has very less outliers but more variability (inter-quantile range)

Box whisker plots of charges vs age across sex



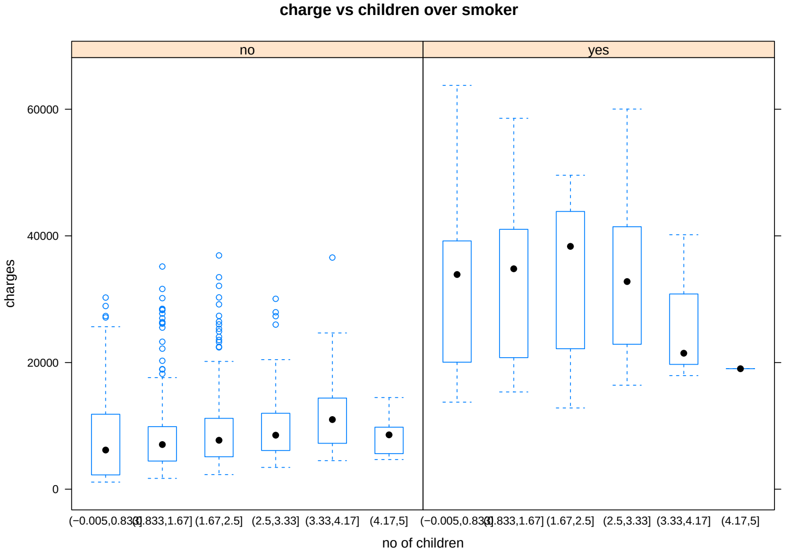
- Both have outliers - But variability in charges for male seems higher than female

Box whisker plots of charges vs age across region



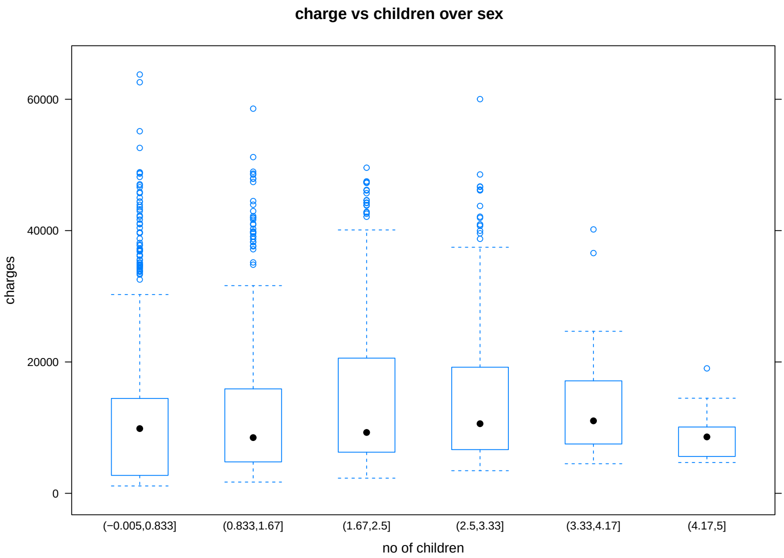
- Nothing new. All region have same pattern

Box whisker plots of charges vs children across smoking

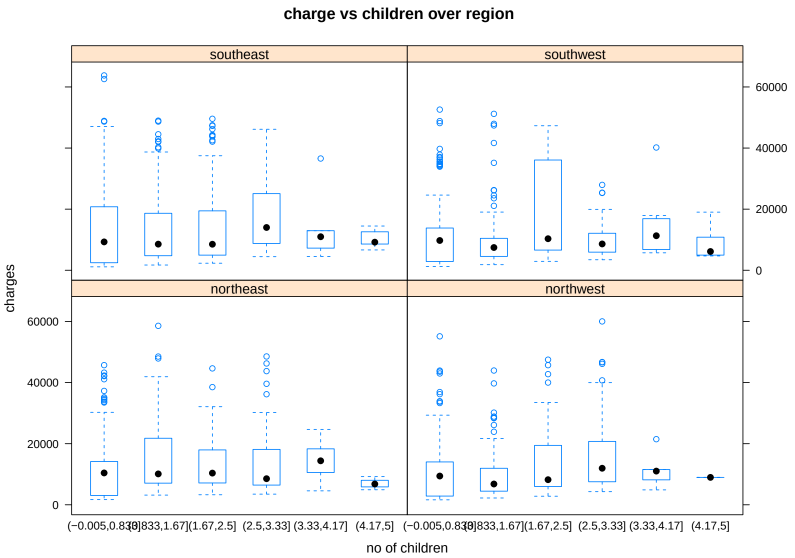


- charges seems to have correlation for smokers. i.e. there is a interaction between smoker and children

Box whisker plots of charges vs children across sex



Box whisker plots of charges vs children across region

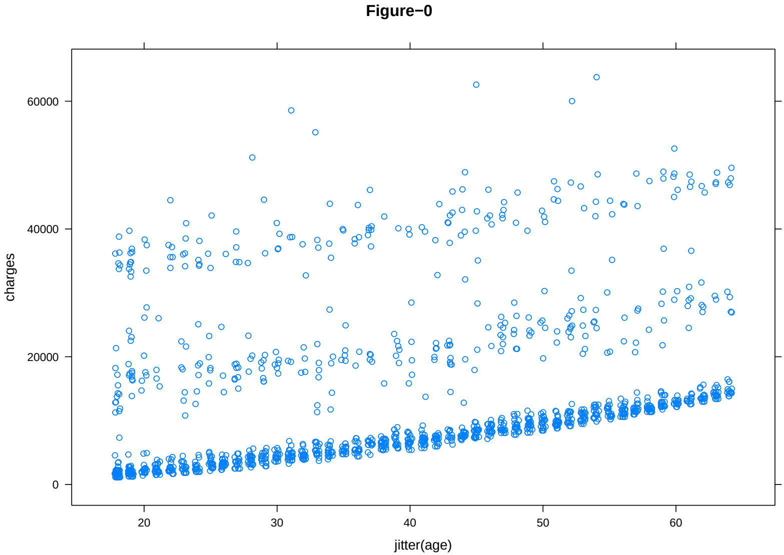


Creating dummy variable for BMI_Category

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

```
##   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
```

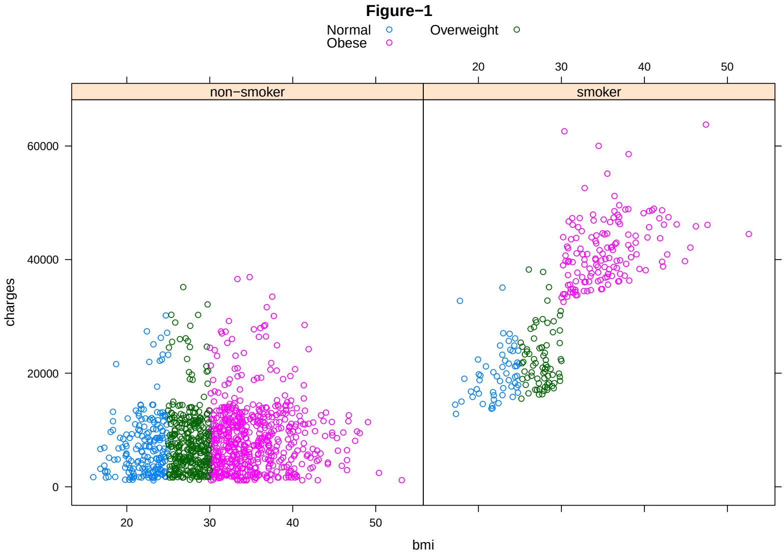
Plot of charge vs age



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)

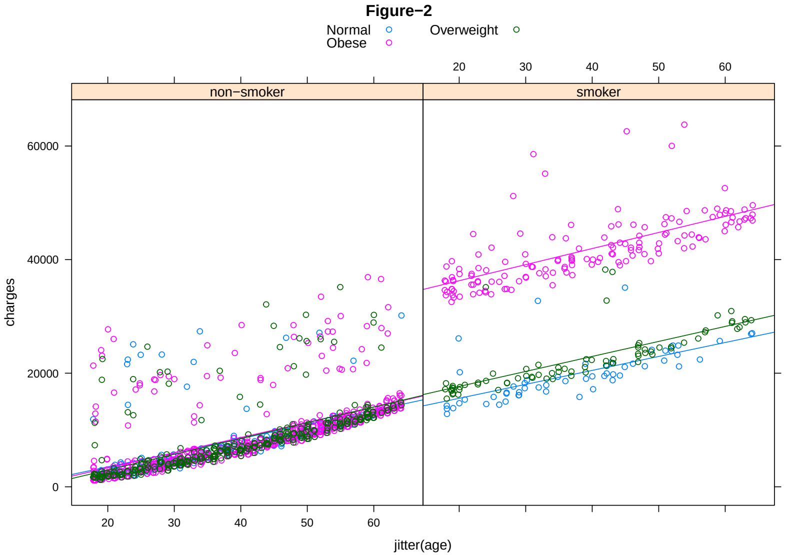
Plot of charge vs BMI across smoking



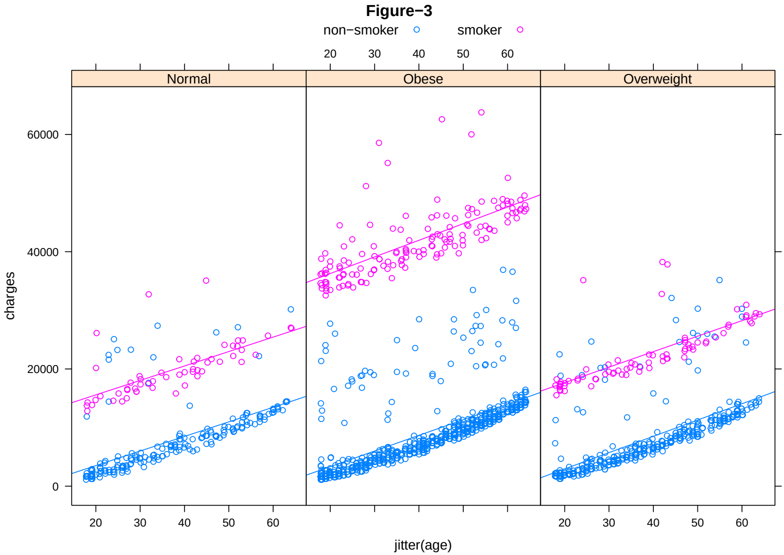
Explanation

- 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.

Plot of charge vs age across smoking



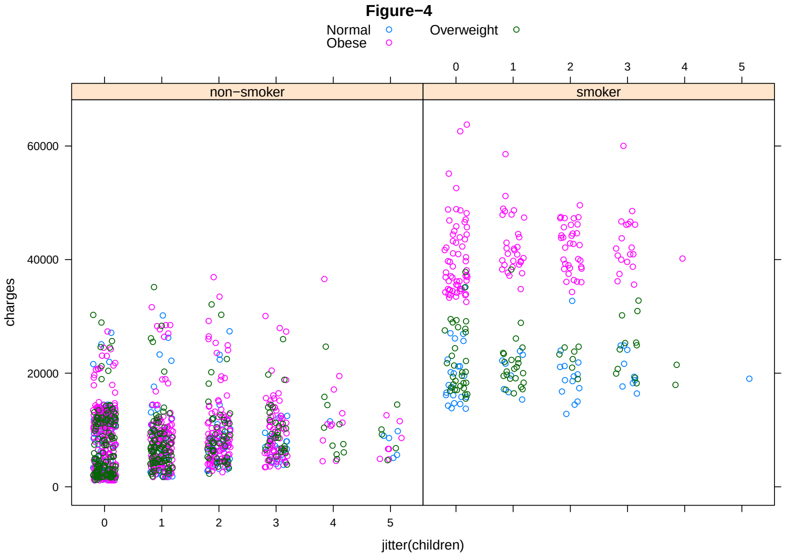
Plot of charge vs age across BMI_Category



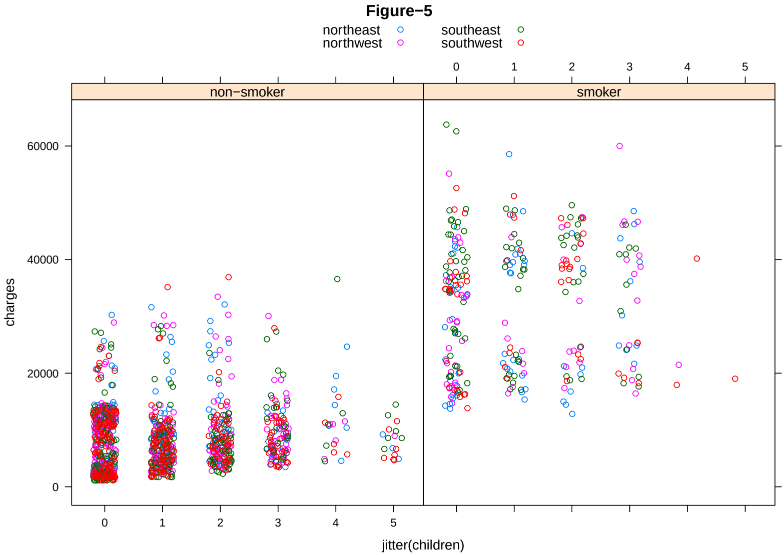
Updated Model

- 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

Plot of charge vs children across smoking



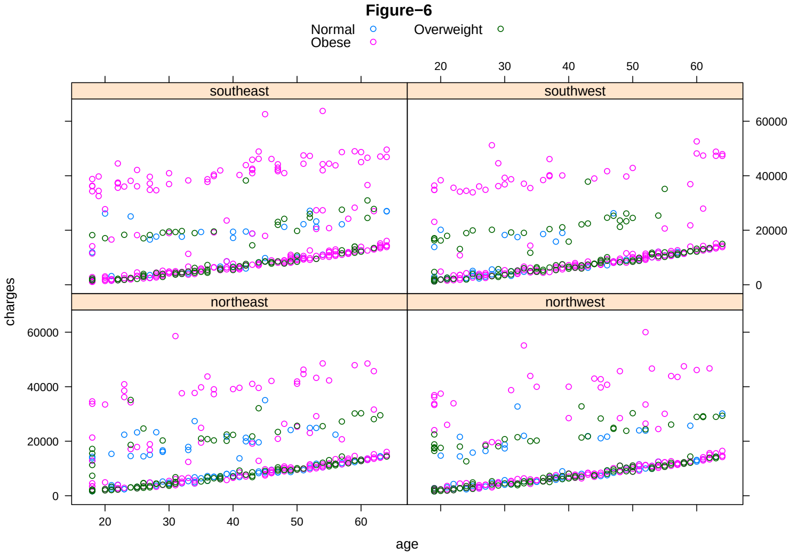
Plot of charge vs children across smoking



Updated model

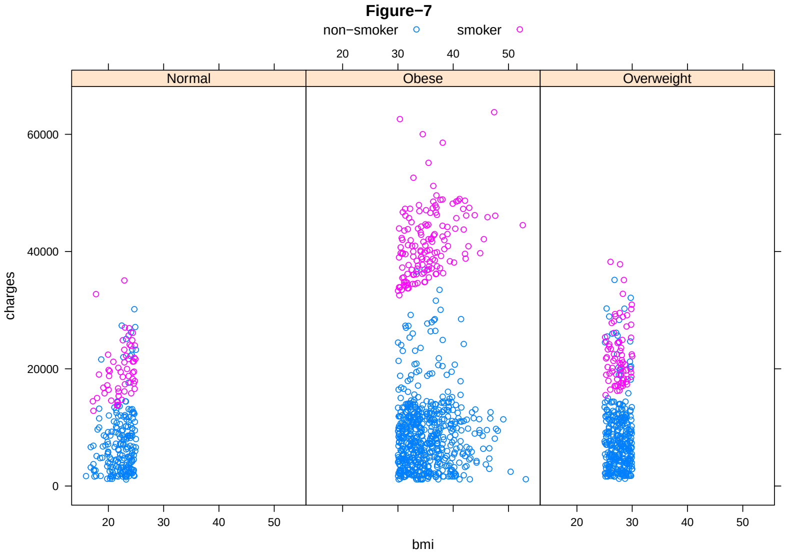
- There is a very slight increment in charge as no of children for non-smoker.
- Assuming the increment is linear wrt no of children, our updated model is... lm(charges ~ 1 + age + BMI_Category*smoker + children)
- Later we will see whether adding children covariate is significant or not..

Plot of charge vs BMI across regions



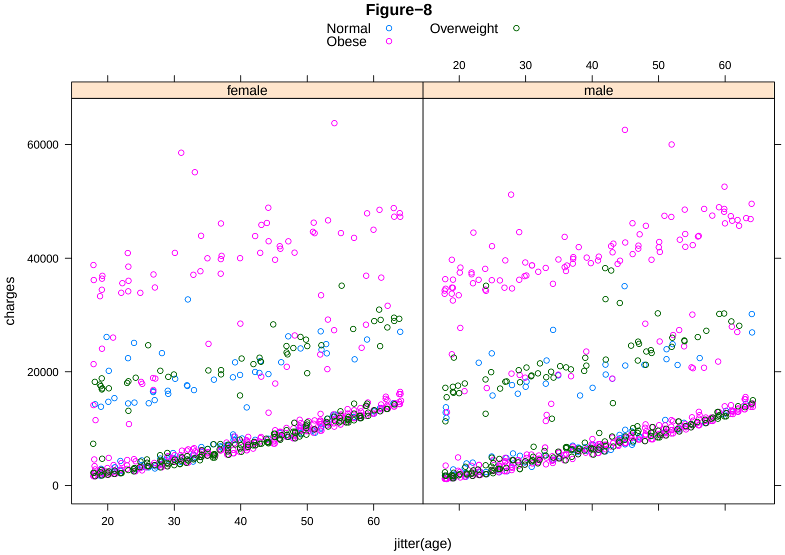
- More or less same pattern for all regions _ So, we won't include region factor in our model

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

Plot of charge vs age across smoking



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

Preliminary Model

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

Dividing the data into training and testing data

```r
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.

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
```

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
```

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.

Comparison with model without children covariate

```r
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
```

```r
sqrt_MSE_fm2
```

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

```r
MAD_fm0
```

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

```r
MAD_fm2
```

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

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$charges- predict(fm3, df_
MAD_fm3= median(abs(df_test$charges- predict(fm3, df_test))
sqrt_MSE_fm3
```

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

```
MAD_fm3
```

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

- much higher MSE and MAD than fm2

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 interactio
```

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

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
```

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
```

Deleting statistically insignificant covariates from the AIC Model

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

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

```
MAD_FM
```

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

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

```r
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
```

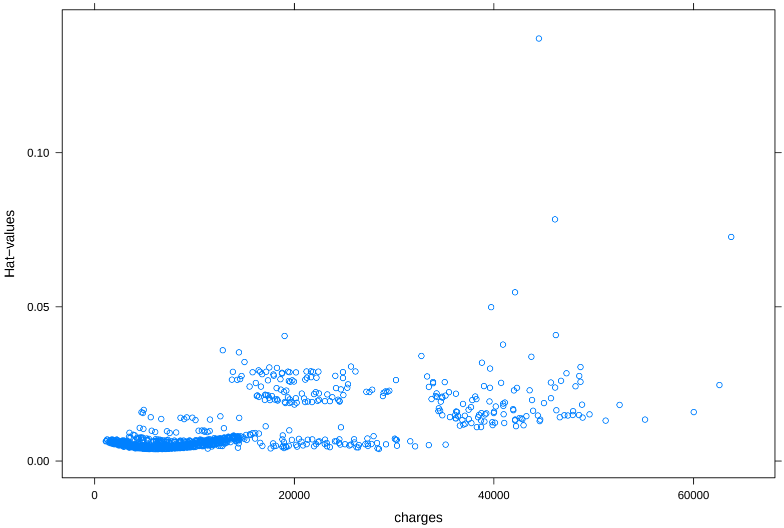
- The P-value of the F test is 0.64»0.05.
- i.e. we failed to reject the null hypothesis.
- So, both models are statistically equivalent but the Final Model has lesser no covariates
- Also LSE_FM $<$ LSE_AIC.

Checking required assumption for multiple linear regression

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

Checking for Outliers

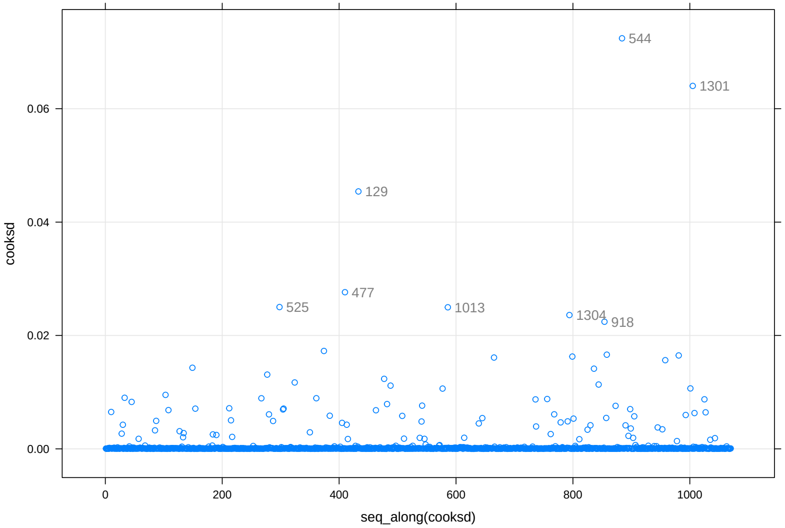
Plot of leverage vs response variable



abs(studententized residual) v/s response variable plot

Cook's Distance 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"))
```



No of outliers (on the basis of Cook distance cutoff)

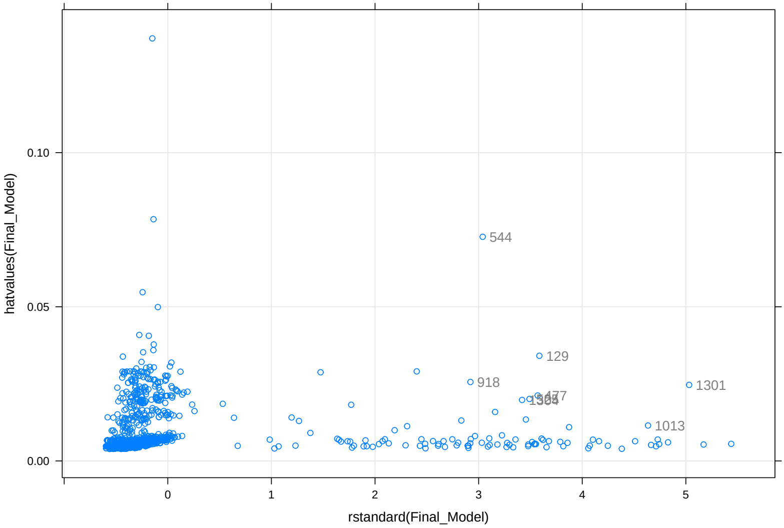
```
outliers=names(which(cooksd > 4/(nrow(df_train)-length(Final_Model$coefficients))))
length(outliers)
```

```
## [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.

Laverage vs Std. Residual Plot

```
xyplot(hatvalues(Final_Model) ~ rstandard(Final_Model), gri
layer(panel.text(x[id], y[id], labels = rownames(df_train)
```



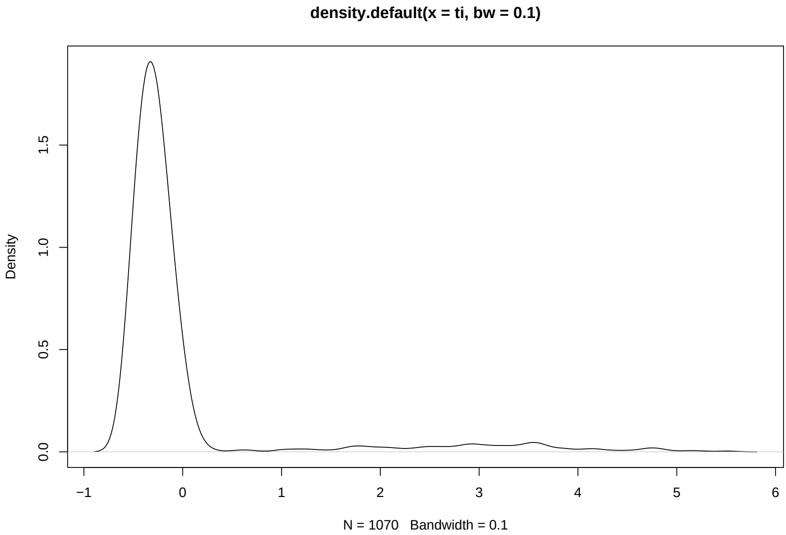
Checking for normality or residuals

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

**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

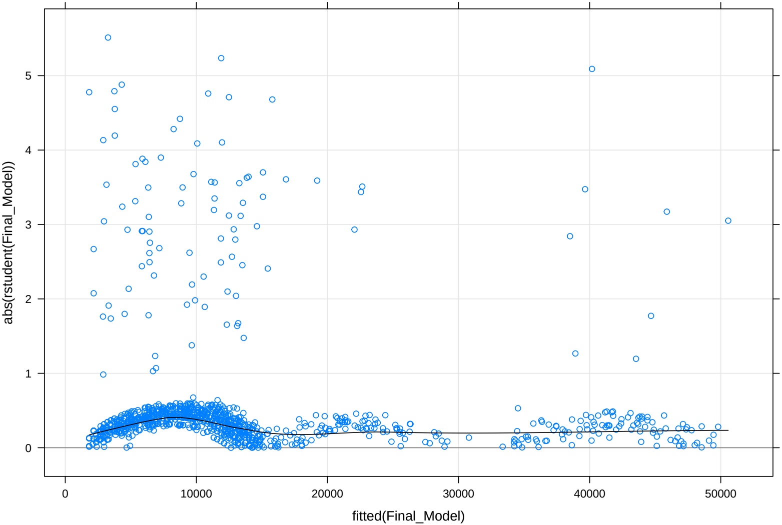
```r
library(car)
```

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

```r
ncvTest(Final_Model)
```

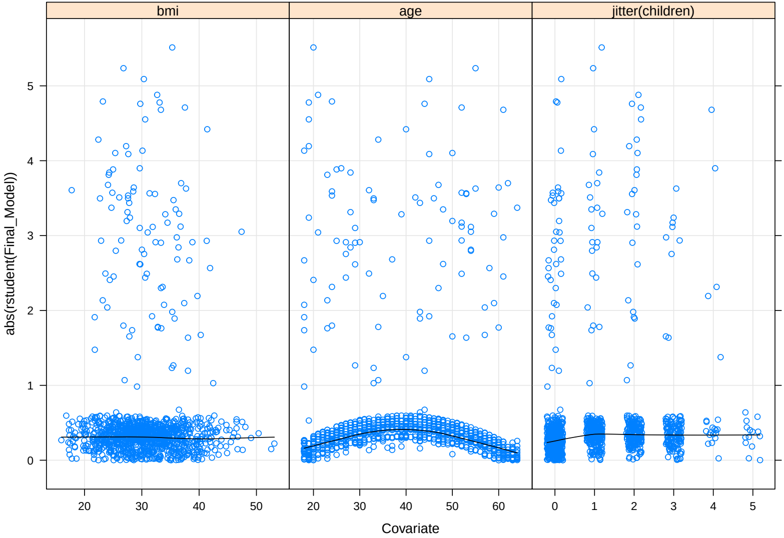
```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 9.868574, Df = 1, p = 0.0016813
```

Plot of absolute studentized residual vs fitted value



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

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

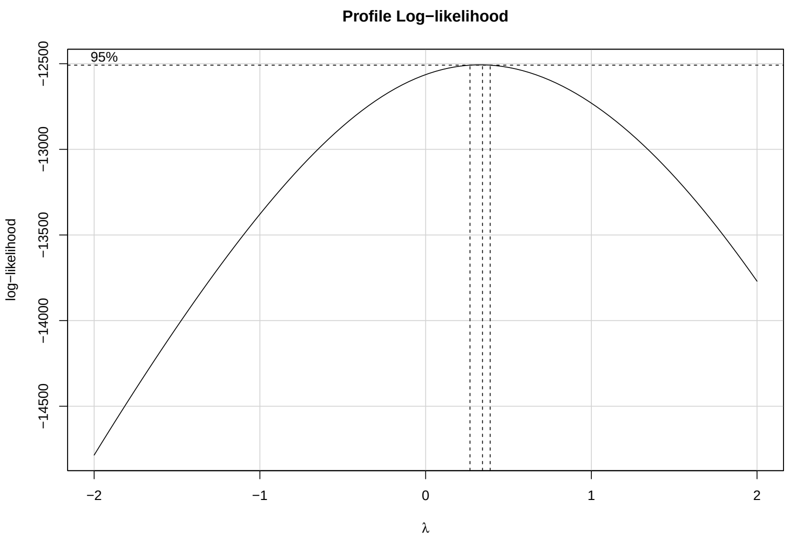
```
fm0 <- lm(formula = charges ~ age + children + BMI_Category
    data = df_train)
ncvTest(fm0) #Test suggest there is variance of residuals
```

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

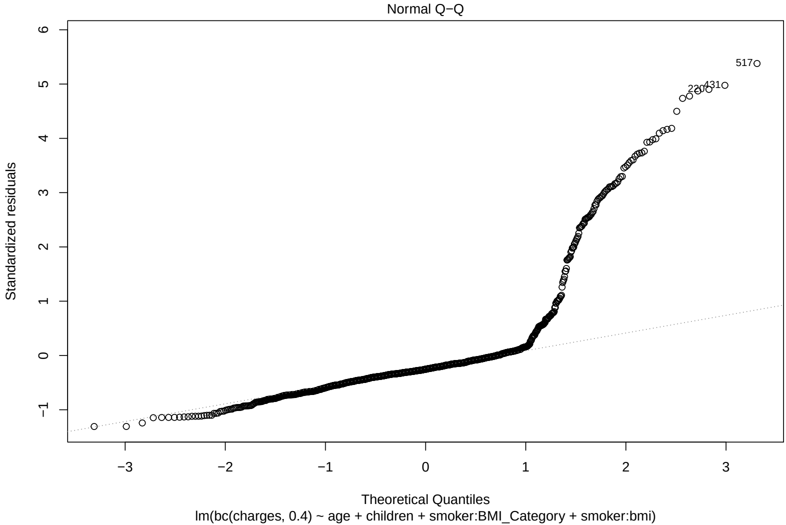
- But problem of non-normality of residuals still exist.
- Now we'll look for possible remedies

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

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+smoke
lad.yhat <- predict(fm.lad,df_test)
MAD_lad= median(abs(df_test$charges- lad.yhat))
MAD_lad
```

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

Huber Loss Function

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

Resistant Regression

Least Trimmed Square(LTS)

```
set.seed(seed = 1001)
fm.lts= ltsreg(charges~ age+children+BMI_Category:smoker+sm
lts.yhat <- predict(fm.lts,df_test)
MAD_lts= median(abs(df_test$charges- lts.yhat))
MAD_lts
```

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

Least Median of Squares (LMS)

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

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

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
```

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

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

Accuracy of the different models

- Say a prediction is good if difference between predicted charge and actual response is less than 1000 dollor
- 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)
}
```

Accuracy of the different models

```r
Accuracy(Final_Model)
```

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

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

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

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

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

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

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

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

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

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

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

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.