Project Final Draft

Madeline Pope, Zachary Sturgeon, David Oniani

November 12, 2019

Abstract

We analyze the dataset describing personal medical costs for 1,339 individuals (posted on Kaggle by Miri Choi). It consists of 6 predictor variables and one response variable. The dataset is diverse featuring both numerical and categorical variables. We will conduct a general analysis of the dataset utilizing a number of statistical approaches. That said, we are particularly interested in exploring the relationships between medical costs and the rest of the numerical variables. Throughout the report, we use a number of statistical methods including linear regression, stepwise regression, and cross validation.

Contents

Data and Preparation

The dataset was obtained from https://www.kaggle.com and features 7 predictor variables. The description of the variables is provided below.

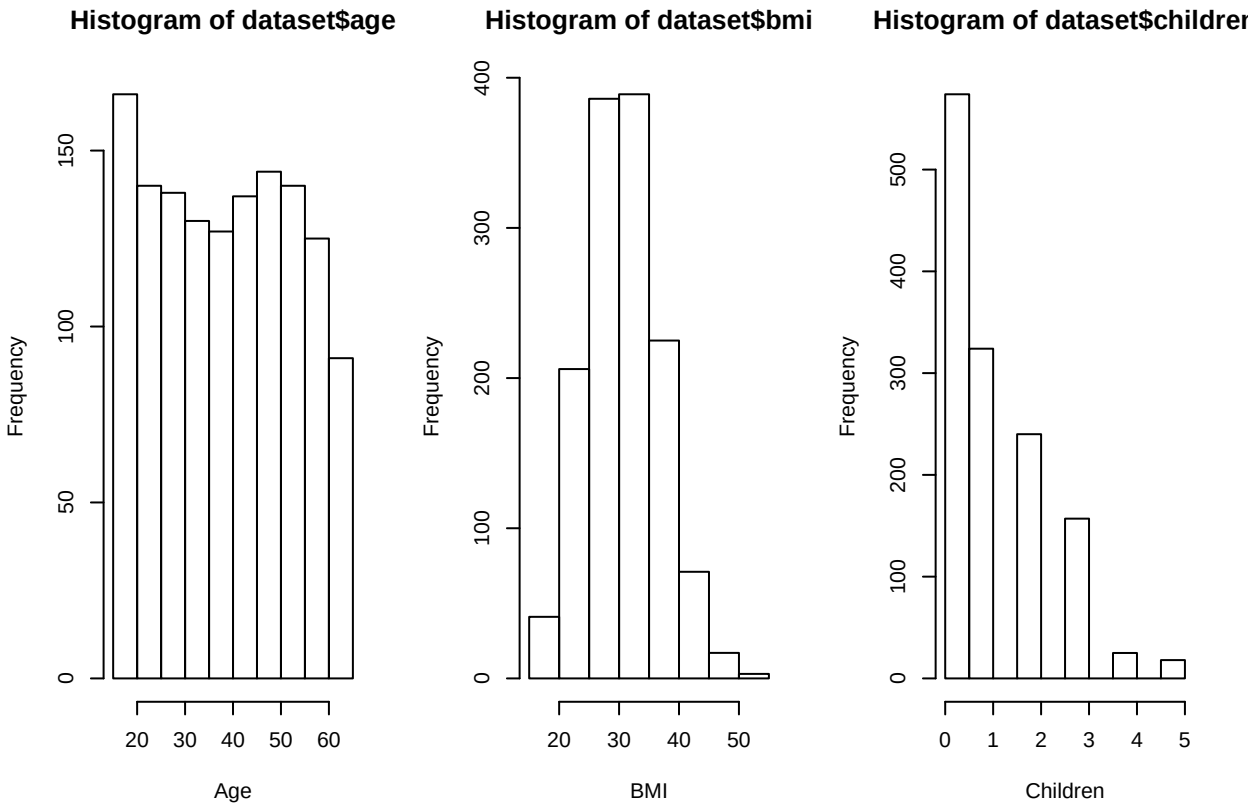
- **age**: age of primary beneficiary.
- **sex**: insurance contractor gender, female, male.
- **bmi**: body mass index, providing an understanding of body weights that are relatively high or low relative to height, objective index of body weight (kg / m^2) using the ratio of height to weight, ideally 18.5 to 24.9.
- **children**: number of children covered by health insurance / Number of dependents.
- **smoker**: whether person smokes or not.
- **region**: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.
- **charges**: individual medical costs billed by health insurance.

```
# Read the CSV data
dataset <- read.csv("./insurance.csv", header=TRUE, sep=",")
```

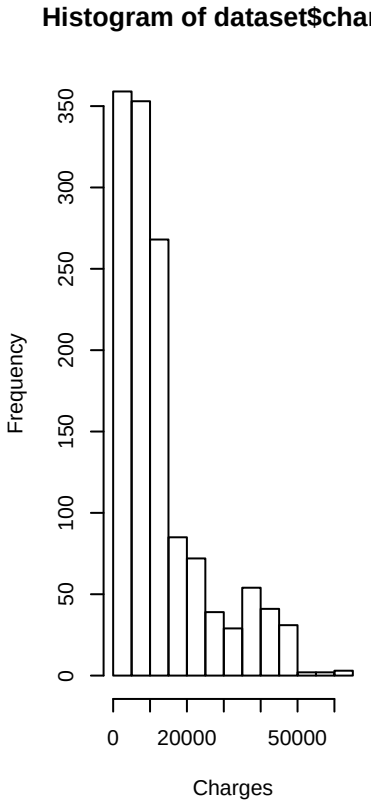
Analyzing Distributions

Analyzing Numerical Variables

```
# Four plots side-by-side
par(mfrow=c(1,3))

# Histograms of all numerical variables
hist(dataset$age, xlab="Age")
hist(dataset$bmi, xlab="BMI")
hist(dataset$children, xlab="Children")
```




```
hist(dataset$charges, xlab="Charges")
```



For numerical variables (age, bmi, children, and charges), we used a traditional histogram to look at the

distributions.

The distribution for age shows a right-skewed trend which is reasonable as, on average, young people are more likely to partake in the survey. However, it would not be unreasonably to try a log or a square transformation on this column.

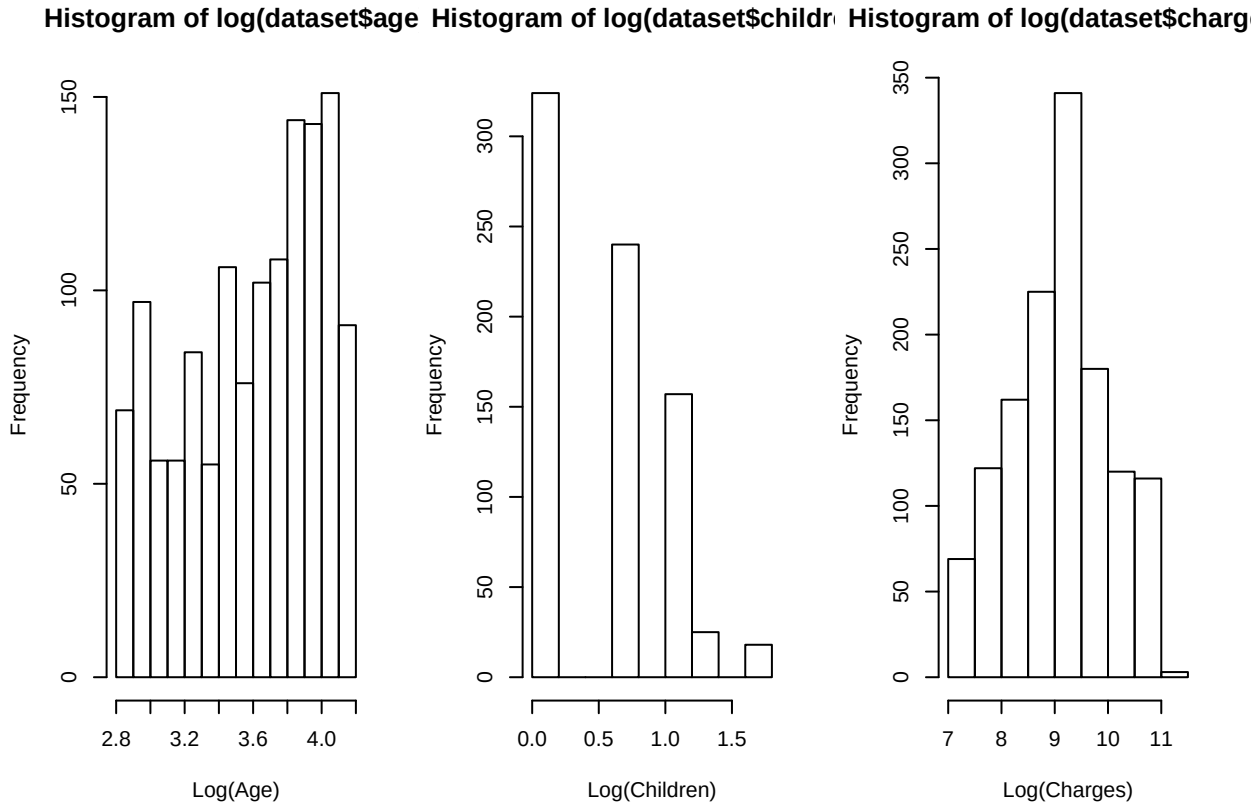
BMI (body-mass-index) shows a normal distribution which is expected as the sample size is over 1000 and, historically, BMI tends to follow a bell curve.

The distribution for children is noticeably right-skewed which is to be expected. Log or square root transformation may be applied.

The distribution for medical charges is also right-skewed. This can be linked with the fact that most people, on average, do not spend a lot of money for the medical purposes (especially if billed by insurance). Just like in the first case, attempting both logarithmic and square root transformations might lead to better results.

Log transformations

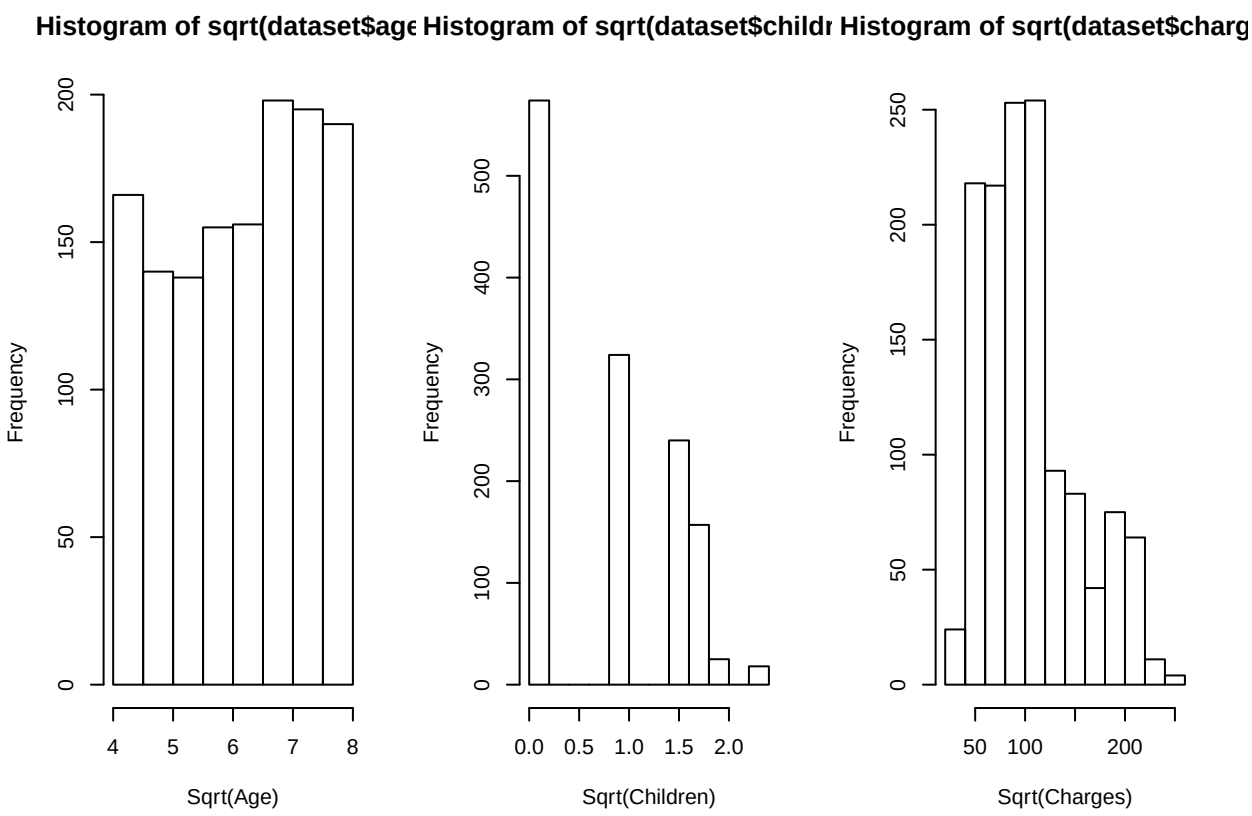
```
# Four plots side-by-side
par(mfrow=c(1,3))

# Histograms of all numerical variables
hist(log(dataset$age), xlab="Log(Age)")
hist(log(dataset$children), xlab="Log(Children)")
hist(log(dataset$charges), xlab="Log(Charges)")
```



Applying a log-transformation on the numerical variables with a right-skewed distribution yields better results. The distribution for charges was normalized; the distribution for age now became left-skewed but the overall shape is promising; distribution for the number of children is still right-skewed.

Square-root transformations

```
# Four plots side-by-side
par(mfrow=c(1,3))

# Histograms of all numerical variables
hist(sqrt(dataset$age), xlab="Sqrt(Age)")
hist(sqrt(dataset$children), xlab="Sqrt(Children)")
hist(sqrt(dataset$charges), xlab="Sqrt(Charges)")
```



The square-root transformation shows the similar results for the distribution of number of children. The distribution for charges got a bit worse, but the age distribution is better. Therefore, moving forward, we will use a log transform on charges and a square root transform on the age. The distribution for the number of children did not get any significant improvements post-transformations.

Analyzing Categorical Variables

```
# Four plots side-by-side
par(mfrow=c(1,4))

# Tables of all categorical variables
tableSex = table(dataset$sex)
tableSmoker = table(dataset$smoker)
tableRegion = table(dataset$region)

# Histograms of all categorical variables
barplot(tableSex)
```

```r
barplot(tableSmoker)
barplot(tableRegion)

# Percentages
tableSex["female"] / sum(tableSex) * 100
```

```
##   female
## 49.47683
```

```r
tableSex["male"] / sum(tableSex) * 100
```

```
##     male
## 50.52317
```

```r
tableSmoker["yes"] / sum(tableSex) * 100
```

```
##      yes
## 20.47833
```

```r
tableSmoker["no"] / sum(tableSex) * 100
```

```
##       no
## 79.52167
```

```r
tableRegion["northeast"] / sum(tableSex) * 100
```

```
## northeast
##  24.21525
```

```r
tableRegion["northwest"] / sum(tableSex) * 100
```

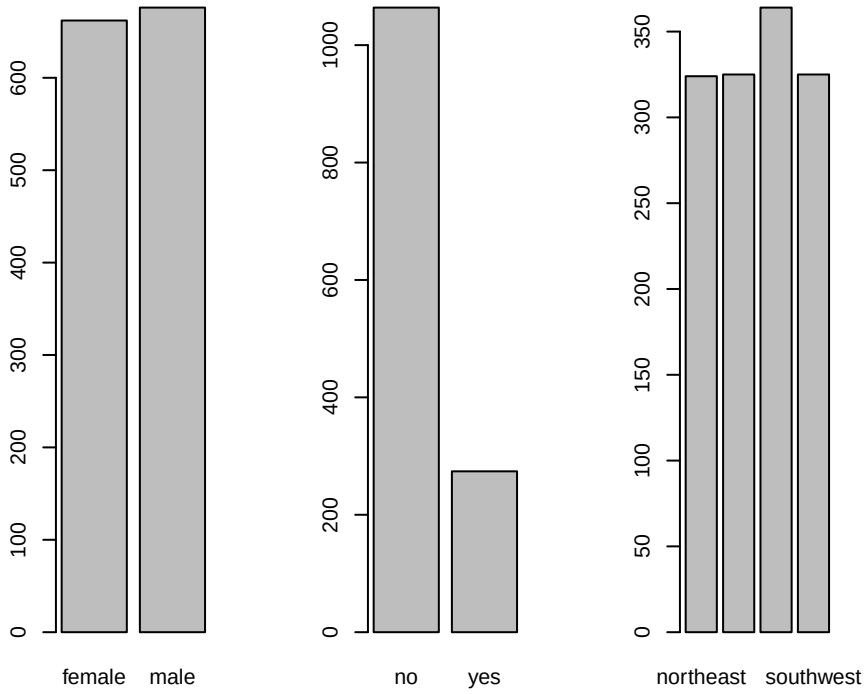
```
## northwest
##  24.28999
```

```r
tableRegion["southeast"] / sum(tableSex) * 100
```

```
## southeast
##  27.20478
```

```r
tableRegion["southwest"] / sum(tableSex) * 100
```

```
## southwest
##  24.28999
```



Barplot was used for the categorical variables. We used `barplot` and `table` functions from the standard library.

It seems like there were slightly more males than females (49.47% females and 50.523% males).

Only 20.478% of people who took the survey were smokers with overwhelming 79.52% being non-smokers.

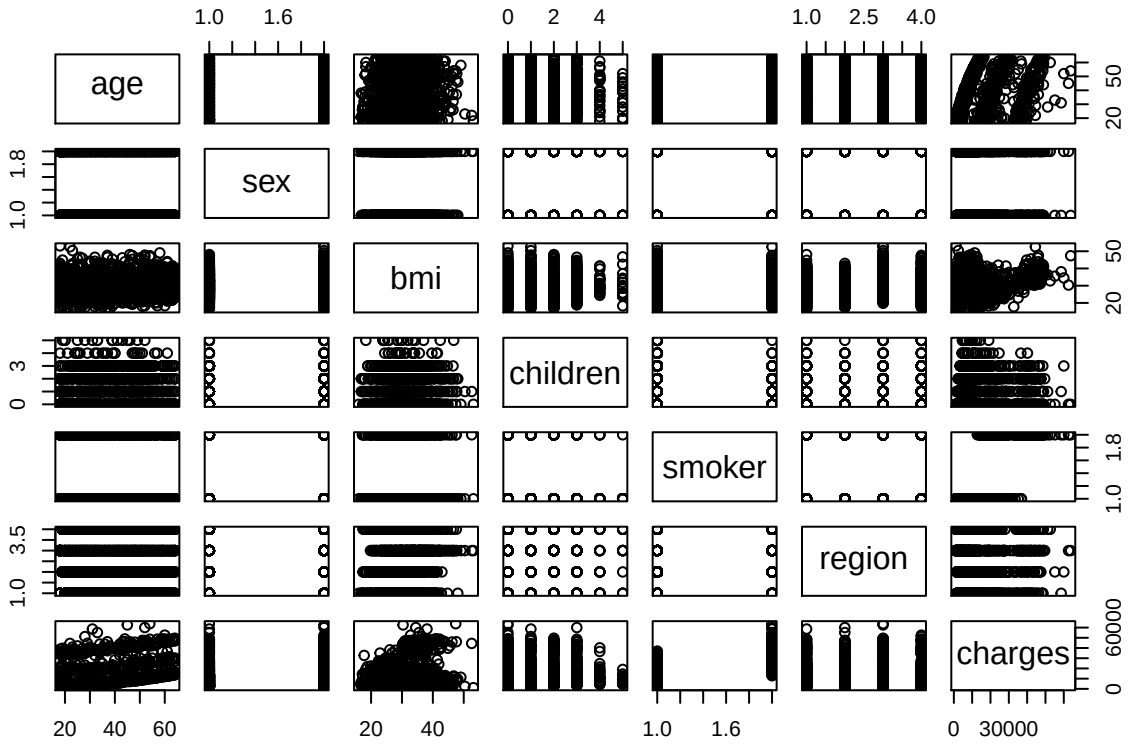
24.21% of the participants were from the northeast region, 24.29% from the northwest region, 27.20% from the southeast region, and 24.29% from the southwest region.

Being categorical variables, we did not perform any transformations.

Linear Models

Simple Linear Relationships

```
plot(dataset)
```



We would like to check if there is any strongly correlated variables. For this purpose, we used the `plot` function and analyzed the generated scatterplot matrix.

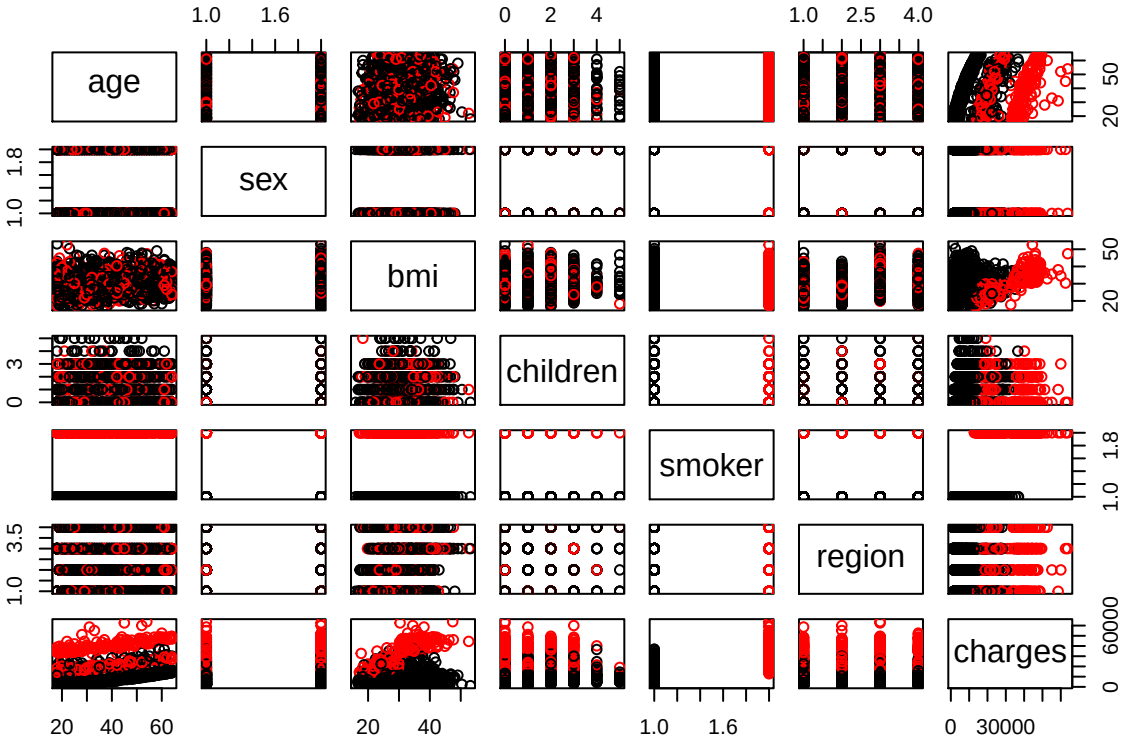
Since the dataset includes 3 categorical variables, a vast majority of scatterplots are "bar-clustered" (vertically or horizontally).

Albeit we have considered the number of children to be a numerical variable, the scatterplot shows that it is of a categorical nature. This is an important conclusion performing linear regression analysis using this variable would be error-prone. In fact, plotting charges VS children would further reinforce our observation.

Variables charges and age are likely to be strongly correlated. At the first glance, there seems to be no correlation between age and bmi. The relationship between charges and bmi is rather unclear and requires more investigation.

At this point, our categorical variables are sex, smoker, region, **and children**.

```
with (dataset, plot(dataset, col=as.factor(smoker)))
```



After exploring relationships, we decided to color the plots by the smoker status. The coloring made it clear that of the categorical variable (smoker) is significant.

From the scatterplot matrix, it is clear that smokers, on average, spend a lot more on medical expenses that non-smokers.

Fitting the Full Linear Model

```
allFit0 <- lm(charges ~ sqrt(age) + sex + bmi + children + smoker + region, data=dataset)
summary(allFit0)
```

```
##
## Call:
## lm(formula = charges ~ sqrt(age) + sex + bmi + children + smoker +
##     region, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -11674  -2891  -1017   1556  29716
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -20813.44    1219.58 -17.066   <2e-16 ***
## sqrt(age)         3075.79     145.78  21.099   <2e-16 ***
## sexmale           -131.83     334.90  -0.394   0.6939
## bmi                342.41      28.76  11.907   <2e-16 ***
## children           405.95     138.79   2.925   0.0035 **
## smokeryes        23837.59     415.56  57.362   <2e-16 ***
## regionnorthwest   -351.16     479.06  -0.733   0.4637
## regionsoutheast  -1044.47     481.49  -2.169   0.0302 *
```

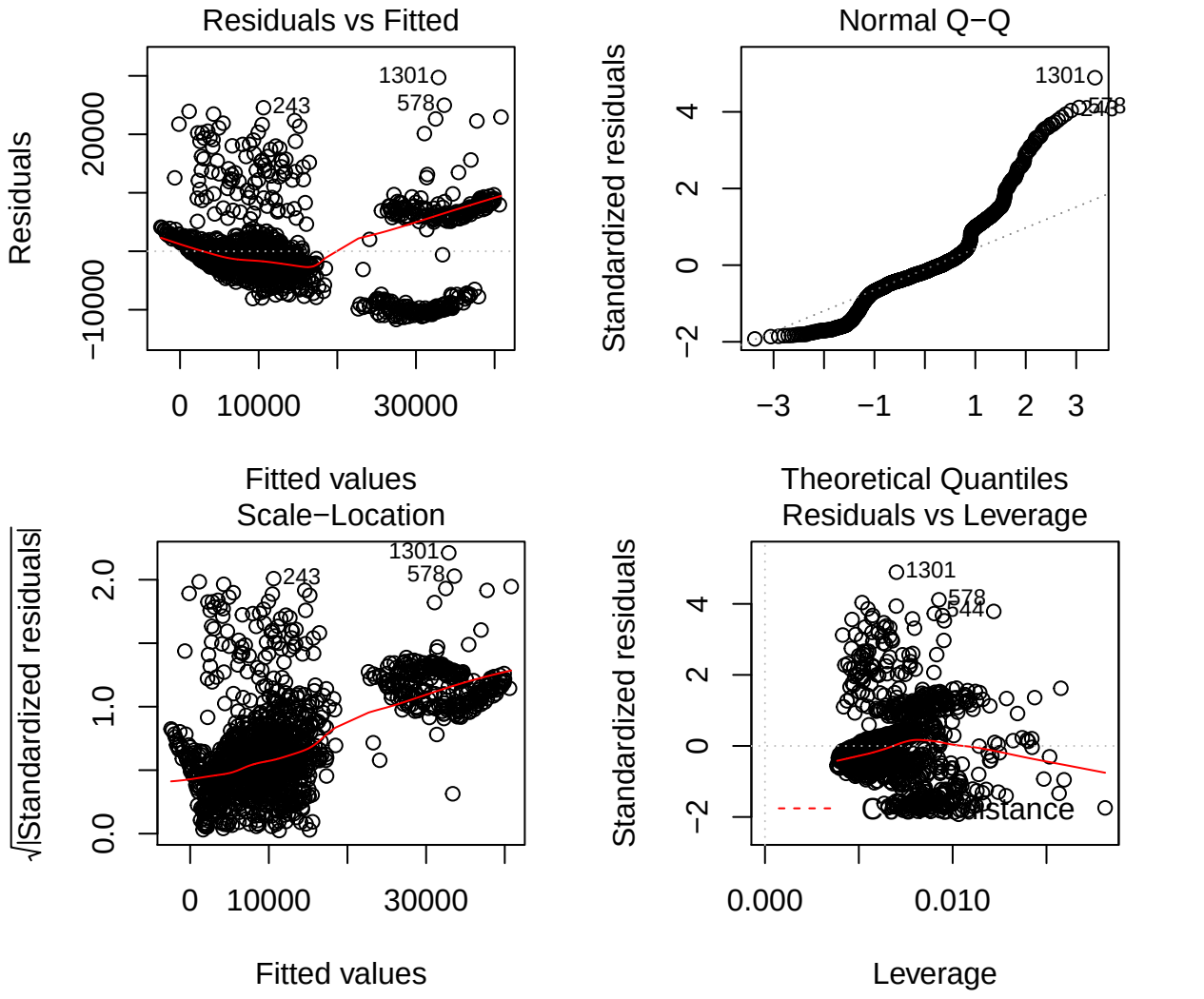
```
## regionsouthwest   -966.36     480.73  -2.010   0.0446 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6098 on 1329 degrees of freedom
## Multiple R-squared:  0.748,  Adjusted R-squared:  0.7465
## F-statistic: 493.1 on 8 and 1329 DF,  p-value: < 2.2e-16
```

For fitting the full linear model, we use the previous observations. Note that the age is square-root transformed (the decision is justified by the histogram).

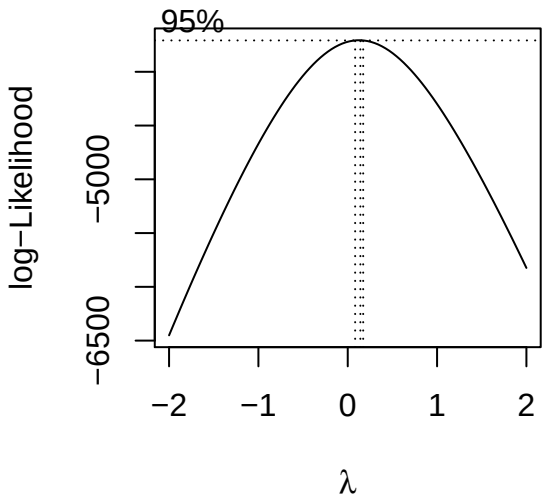
The adjusted R-squared value is 0.7465 which tells us that 74.65% of the variation in charges is explained by the model.

Residual standard error is 6098 meaning that, on average, predictions of the model are 6098 dollars away from the real value.

```
par(mfrow=c(1,2))
plot(allFit0)
```




```
MASS::boxcox(charges ~ sqrt(age) + sex + bmi + children + smoker + region, data=dataset)
```



The first plot (Residuals VS Fitted Values) shows evidence of curvature. There is a strong evidence of heteroscedasticity.

The second plot (Standardized Residuals VS Theoretical Quantiles) shows weak evidence that both sets of residuals are coming from normal distributions. This is the case since the points are, to some degree, aligned across line, but again, this visual check is not a strong evidence. Besides, most points are not aligned across the desired (dotted) line.

The third plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows a set of lines which look like a curved line. This suggests that the residuals are not spread equally along the ranges of predictors and that the variance is not constant.

Residuals VS Leverage plot shows no significant outliers.

Before fully interpreting the model, we performed the Box-Cox analysis which further reinforced our previous observation (from a histogram) that log-transforming charges is suitable for this case. This is due to the fact that the optimal value for λ (lambda) is far from 1.

Manual Backward Elimination

```
allFit1 <- lm(log(charges) ~ sqrt(age) + sex + bmi + children + smoker + region, data=dataset)
summary(allFit1)
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + sex + bmi + children +
##     smoker + region, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01382 -0.20734 -0.06487  0.05970  2.21077
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.788769   0.088497  65.412  < 2e-16 ***
## sqrt(age)        0.422666   0.010578  39.956  < 2e-16 ***
## sexmale         -0.074988   0.024301  -3.086 0.002072 **
## bmi              0.013616   0.002087   6.525 9.62e-11 ***
## children         0.091960   0.010071   9.131  < 2e-16 ***
```

```
## smokeryes        1.553346   0.030155  51.512  < 2e-16 ***
## regionnorthwest -0.063431   0.034763  -1.825 0.068275 .
## regionsoutheast -0.157451   0.034939  -4.506 7.17e-06 ***
## regionsouthwest -0.129619   0.034884  -3.716 0.000211 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4425 on 1329 degrees of freedom
## Multiple R-squared:  0.7698, Adjusted R-squared:  0.7685
## F-statistic: 555.7 on 8 and 1329 DF,  p-value: < 2.2e-16
```

The adjusted R-squared value is 0.7685 which tells us that 76.85% of the variation in charges is explained by the model. This is an improved over the previous model (`allFit0`) with the adjusted R-squared value of 0.7465 (0.22 or 2.2% improvement).

Residual standard error is 0.4425 meaning that, on average, predictions of the model are 0.4425 log(dollars) away from the real value. We cannot compare this result with that of the previous model (`allFit0`) as the units do not match.

Once again, before doing a full interpretation of the model, we proceed by first backward eliminating the least significant estimates.

Both sex or region can be considered as the least significant estimates. Due to this reason, we will attempt to remove both individually and compare the results.

```
allFit2 <- lm(log(charges) ~ sqrt(age) + bmi + children + smoker + region, data=dataset)
summary(allFit2)
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + bmi + children + smoker +
##     region, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04467 -0.20677 -0.05805  0.06544  2.17490
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.756776   0.088169  65.292  < 2e-16 ***
## sqrt(age)        0.423493   0.010609  39.919  < 2e-16 ***
## bmi              0.013314   0.002091   6.367 2.64e-10 ***
## children         0.091406   0.010102   9.048  < 2e-16 ***
## smokeryes        1.546359   0.030166  51.262  < 2e-16 ***
## regionnorthwest -0.062983   0.034874  -1.806 0.071138 .
## regionsoutheast -0.157073   0.035050  -4.481 8.05e-06 ***
## regionsouthwest -0.129234   0.034995  -3.693 0.000231 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4439 on 1330 degrees of freedom
## Multiple R-squared:  0.7682, Adjusted R-squared:  0.767
## F-statistic: 629.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

`allFit2` is a model with sex regressor variable removed.

Adjusted R-squared value decreased to 0.7670 (from 0.7685 in the previous model that included sex).

Residual standard error value is 0.4439 log(dollars) which is a slight increase comparing to the previous model with sex inclusive (0.4439 log(dollars) VS 0.4425 log(dollars)).

Overall, the new model, excluding sex variable, is similar to the previous model, yet performs slightly worse. Therefore, we proceed by leaving sex and removing region regressor variable.

```
allFit3 <- lm(log(charges) ~ sqrt(age) + sex + bmi + children + smoker, data=dataset)
summary(allFit3)
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + sex + bmi + children +
##     smoker, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.02413 -0.21103 -0.06328  0.06678  2.15618
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.765780   0.087026  66.253  < 2e-16 ***
## sqrt(age)    0.424275   0.010653  39.827  < 2e-16 ***
## sexmale     -0.074580   0.024493  -3.045  0.00237 **
## bmi          0.011143   0.002014   5.534 3.77e-08 ***
## children     0.091794   0.010142   9.051  < 2e-16 ***
## smokeryes    1.549248   0.030308  51.117  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.446 on 1332 degrees of freedom
## Multiple R-squared:  0.7657, Adjusted R-squared:  0.7648
## F-statistic: 870.5 on 5 and 1332 DF,  p-value: < 2.2e-16
```

`allFit3` is a model without region predictor variable.

Adjusted R squared value is smaller than in `allFit1` (0.7657 VS 0.7685).

Residual standard error is slightly bigger than in `allFit1` (0.446 log(dollars) VS 0.4425 log(dollars)).

These observations suggest not removing the region predictor variable. In fact, we have tried removing all predictor variables individually and the most optimal choice is leaving the initial model without any alterations.

We will now interpret the full model.

Interpretation

```
summary(allFit1)
```

```
##
## Call:
## lm(formula = log(charges) ~ sqrt(age) + sex + bmi + children +
##     smoker + region, data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.01382 -0.20734 -0.06487  0.05970  2.21077
```

```
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.788769   0.088497  65.412  < 2e-16 ***
## sqrt(age)         0.422666   0.010578  39.956  < 2e-16 ***
## sexmale          -0.074988   0.024301  -3.086 0.002072 **
## bmi               0.013616   0.002087   6.525 9.62e-11 ***
## children          0.091960   0.010071   9.131  < 2e-16 ***
## smokeryes         1.553346   0.030155  51.512  < 2e-16 ***
## regionnorthwest -0.063431   0.034763  -1.825 0.068275 .
## regionsoutheast -0.157451   0.034939  -4.506 7.17e-06 ***
## regionsouthwest -0.129619   0.034884  -3.716 0.000211 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4425 on 1329 degrees of freedom
## Multiple R-squared:  0.7698, Adjusted R-squared:  0.7685
## F-statistic: 555.7 on 8 and 1329 DF,  p-value: < 2.2e-16
```

```
confint(allFit1)
```

```
##                         2.5 %       97.5 %
## (Intercept)       5.615158737  5.962378750
## sqrt(age)         0.401914374  0.443418523
## sexmale          -0.122661217 -0.027314431
## bmi               0.009522581  0.017709536
## children          0.072202316  0.111716834
## smokeryes         1.494189772  1.612502754
## regionnorthwest -0.131626404  0.004765323
## regionsoutheast -0.225992246 -0.088910180
## regionsouthwest -0.198051897 -0.061186178
```

The summary section shows that all estimates are significant and therefore, we proceed by interpreting them all.

The mean response changes between 0.402 and 0.443 log(dollars) per sqrt(years) (`sqrt(age)`), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between -0.123 and -0.027 log(dollars) per as to whether sex is equal to male (`sexmale`), for any change in the category, holding all other predictors fixed.

The mean response changes between 0.010 and 0.018 log(dollars) per bmi unit (`bmi`), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between 0.072 and 0.112 log(dollars) per child (`children`), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between 1.494 and 1.613 log(dollars) as to whether smoker is equal to yes (`smokeryes`), for any 1-unit increase in the predictor with 95% confidence, holding all other predictors fixed.

The mean response changes between -0.132 and 0.005 log(dollars) per as to whether region is equal to regionnorthwest (`regionnorthwest`), for any change in the regionnortheast category, holding all other predictors fixed.

The mean response changes between -0.226 and -0.089 log(dollars) per as to whether region is equal to regionsoutheast (`regionsoutheast`), for any change in the regionnortheast category, holding all other predictors fixed.

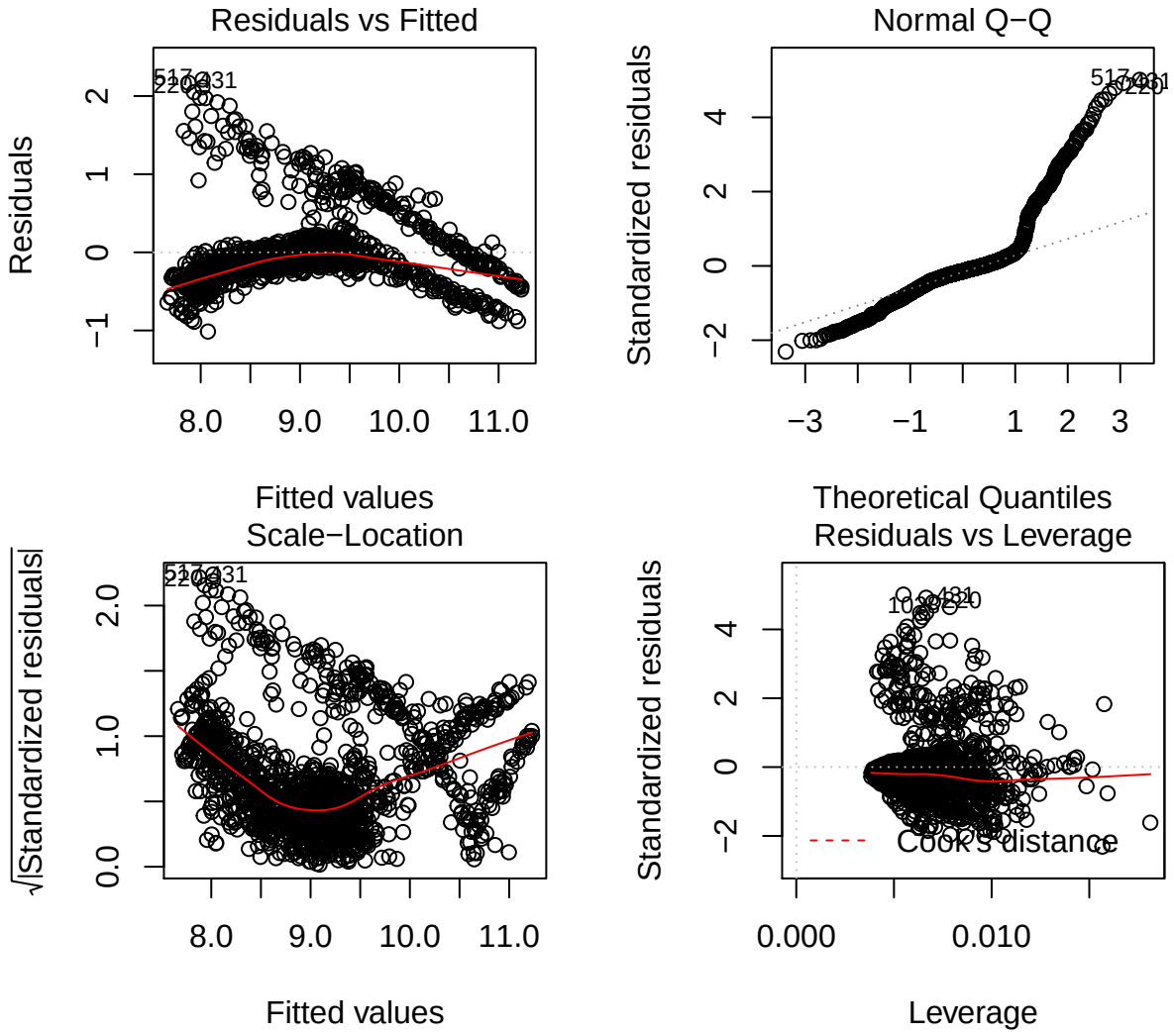
The mean response changes between -0.198 and -0.061 log(dollars) per as to whether region is equal to

regionsouthwest (`regionsouthwest`), for any change in the regionnortheast category, holding all other predictors fixed.

The Adjusted R-squared value is 0.7685 which tells us that 76.85% of the variation in charges is explained by this model.

Residual standard error is 0.4425 meaning that on average, predictions of the model are 0.4425 `log`(dollars) away from the real value.

```
par(mfrow=c(1,2))
plot(allFit1)
```



The first plot (Residuals VS Fitted Values) shows a curvature and that the spread is not constant (that said, this is an improvement over the initial model `allFit0`). There is evidence of heteroscedasticity.

The second plot (Standardized Residuals VS Theoretical Quantiles) gives a strong evidence against a claim that both sets of residuals are coming from normal distributions. Besides, points are not aligned across the desired (dotted) line.

The third plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows almost a curved line. This suggests that the residuals are not spread equally along the ranges of predictors and that the variance is not constant. That said, this is an improvement over the initial model `allFit0`.

Residuals VS Leverage plot shows no significant outliers.

Model Improvements

Step-wise Regression

```
allFitAll <- lm(log(charges) ~ sqrt(age) +
                               as.factor(sex) +
                               bmi +
                               as.factor(children) +
                               as.factor(smoker) +
                               as.factor(region),
                               data=dataset)

allFitAllAIC = step(allFitAll, direction = "both")
```

```
## Start:  AIC=-2175.37
## log(charges) ~ sqrt(age) + as.factor(sex) + bmi + as.factor(children) +
##     as.factor(smoker) + as.factor(region)
##
##                        Df Sum of Sq    RSS      AIC
## <none>                              258.18 -2175.37
## - as.factor(sex)        1      1.88 260.06 -2167.67
## - as.factor(region)     3      4.75 262.94 -2156.96
## - bmi                   1      8.14 266.32 -2135.82
## - as.factor(children)   5     18.33 276.51 -2093.62
## - sqrt(age)             1    311.58 569.76 -1118.26
## - as.factor(smoker)     1    517.24 775.42  -705.91
```

In order to verify our observations and the final model, we have also performed step-wise regression analysis which yielded similar results and the same final linear regression model.

Interaction Effects

It seems like interaction effects may benefit our model.

```
# Let's throw in some intercation effects and do step wise regression
# For these purposes, we first need to center the continuous variables

centeredSqrtAge = sqrt(dataset$age) - mean(sqrt(dataset$age))
centeredBMI = dataset$bmi - mean(dataset$bmi)

allFitAll <- lm(log(charges) ~ (centeredSqrtAge +
                                as.factor(sex) +
                                centeredBMI +
                                as.factor(children) +
                                as.factor(smoker) +
                                as.factor(region))^2,
                                data=dataset)

allFitAllAIC = step(allFitAll, direction = "both")
```

```
## Start:  AIC=-2612.82
## log(charges) ~ (centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region))^2
##
```

```
##                                         Df Sum of Sq    RSS     AIC
## - as.factor(children):as.factor(region) 15     1.705 173.71 -2629.6
## - centeredBMI:as.factor(children)        5     0.143 172.14 -2621.7
## - as.factor(sex):as.factor(children)     5     0.211 172.21 -2621.2
## - as.factor(sex):as.factor(region)       3     0.073 172.07 -2618.2
## - as.factor(sex):centeredBMI             1     0.025 172.03 -2614.6
## - as.factor(smoker):as.factor(region)    3     0.653 172.65 -2613.8
## - centeredSqrtAge:centeredBMI            1     0.221 172.22 -2613.1
## <none>                                             172.00 -2612.8
## - as.factor(sex):as.factor(smoker)       1     0.439 172.44 -2611.4
## - centeredBMI:as.factor(region)          3     1.469 173.47 -2607.4
## - centeredSqrtAge:as.factor(sex)         1     1.368 173.37 -2604.2
## - centeredSqrtAge:as.factor(region)      3     2.955 174.96 -2596.0
## - as.factor(children):as.factor(smoker)  5     4.308 176.31 -2589.7
## - centeredSqrtAge:as.factor(children)    5     6.665 178.67 -2571.9
## - centeredBMI:as.factor(smoker)          1    18.544 190.54 -2477.8
## - centeredSqrtAge:as.factor(smoker)      1    42.019 214.02 -2322.4
##
## Step:  AIC=-2629.62
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(children) + as.factor(sex):as.factor(smoker) +
##     as.factor(sex):as.factor(region) + centeredBMI:as.factor(children) +
##     centeredBMI:as.factor(smoker) + centeredBMI:as.factor(region) +
##     as.factor(children):as.factor(smoker) + as.factor(smoker):as.factor(region)
##
##                                         Df Sum of Sq    RSS     AIC
## - as.factor(sex):as.factor(children)     5     0.202 173.91 -2638.1
## - centeredBMI:as.factor(children)        5     0.450 174.16 -2636.2
## - as.factor(sex):as.factor(region)       3     0.087 173.79 -2634.9
## - as.factor(sex):centeredBMI             1     0.031 173.74 -2631.4
## - centeredSqrtAge:centeredBMI            1     0.190 173.90 -2630.2
## - as.factor(smoker):as.factor(region)    3     0.756 174.46 -2629.8
## <none>                                             173.71 -2629.6
## - as.factor(sex):as.factor(smoker)       1     0.455 174.16 -2628.1
## - centeredBMI:as.factor(region)          3     1.603 175.31 -2623.3
## - centeredSqrtAge:as.factor(sex)         1     1.513 175.22 -2620.0
## - centeredSqrtAge:as.factor(region)      3     2.837 176.54 -2613.9
## + as.factor(children):as.factor(region) 15     1.705 172.00 -2612.8
## - as.factor(children):as.factor(smoker)  5     4.649 178.35 -2604.3
## - centeredSqrtAge:as.factor(children)    5     6.920 180.63 -2587.3
## - centeredBMI:as.factor(smoker)          1    18.765 192.47 -2494.4
## - centeredSqrtAge:as.factor(smoker)      1    43.602 217.31 -2332.0
##
## Step:  AIC=-2638.07
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(smoker) + as.factor(sex):as.factor(region) +
```

```
##     centeredBMI:as.factor(children) + centeredBMI:as.factor(smoker) +
##     centeredBMI:as.factor(region) + as.factor(children):as.factor(smoker) +
##     as.factor(smoker):as.factor(region)
##
##                                         Df Sum of Sq    RSS     AIC
## - centeredBMI:as.factor(children)        5     0.432 174.34 -2644.7
## - as.factor(sex):as.factor(region)       3     0.094 174.00 -2643.3
## - as.factor(sex):centeredBMI             1     0.032 173.94 -2639.8
## - centeredSqrtAge:centeredBMI            1     0.195 174.10 -2638.6
## - as.factor(smoker):as.factor(region)    3     0.753 174.66 -2638.3
## <none>                                               173.91 -2638.1
## - as.factor(sex):as.factor(smoker)       1     0.481 174.39 -2636.4
## - centeredBMI:as.factor(region)          3     1.562 175.47 -2632.1
## + as.factor(sex):as.factor(children)     5     0.202 173.71 -2629.6
## - centeredSqrtAge:as.factor(sex)         1     1.512 175.42 -2628.5
## - centeredSqrtAge:as.factor(region)      3     2.878 176.78 -2622.1
## + as.factor(children):as.factor(region) 15     1.696 172.21 -2621.2
## - as.factor(children):as.factor(smoker)  5     4.760 178.67 -2611.9
## - centeredSqrtAge:as.factor(children)    5     7.039 180.95 -2595.0
## - centeredBMI:as.factor(smoker)          1    18.861 192.77 -2502.3
## - centeredSqrtAge:as.factor(smoker)      1    43.578 217.49 -2340.9
##
## Step:  AIC=-2644.74
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(smoker) + as.factor(sex):as.factor(region) +
##     centeredBMI:as.factor(smoker) + centeredBMI:as.factor(region) +
##     as.factor(children):as.factor(smoker) + as.factor(smoker):as.factor(region)
##
##                                         Df Sum of Sq    RSS     AIC
## - as.factor(sex):as.factor(region)       3     0.062 174.40 -2650.3
## - as.factor(sex):centeredBMI             1     0.041 174.38 -2646.4
## - centeredSqrtAge:centeredBMI            1     0.189 174.53 -2645.3
## - as.factor(smoker):as.factor(region)    3     0.761 175.10 -2644.9
## <none>                                               174.34 -2644.7
## - as.factor(sex):as.factor(smoker)       1     0.476 174.82 -2643.1
## - centeredBMI:as.factor(region)          3     1.571 175.91 -2638.7
## + centeredBMI:as.factor(children)        5     0.432 173.91 -2638.1
## + as.factor(sex):as.factor(children)     5     0.184 174.16 -2636.2
## - centeredSqrtAge:as.factor(sex)         1     1.485 175.82 -2635.4
## + as.factor(children):as.factor(region) 15     1.990 172.35 -2630.1
## - centeredSqrtAge:as.factor(region)      3     2.960 177.30 -2628.2
## - as.factor(children):as.factor(smoker)  5     4.844 179.18 -2618.1
## - centeredSqrtAge:as.factor(children)    5     7.216 181.56 -2600.5
## - centeredBMI:as.factor(smoker)          1    18.638 192.98 -2510.8
## - centeredSqrtAge:as.factor(smoker)      1    43.484 217.82 -2348.8
##
## Step:  AIC=-2650.27
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
```

```
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(smoker) + centeredBMI:as.factor(smoker) +
##     centeredBMI:as.factor(region) + as.factor(children):as.factor(smoker) +
##     as.factor(smoker):as.factor(region)
##
##                                       Df Sum of Sq    RSS     AIC
## - as.factor(sex):centeredBMI           1     0.041 174.44 -2651.9
## - centeredSqrtAge:centeredBMI          1     0.183 174.59 -2650.9
## - as.factor(smoker):as.factor(region)  3     0.761 175.16 -2650.4
## <none>                                             174.40 -2650.3
## - as.factor(sex):as.factor(smoker)     1     0.476 174.88 -2648.6
## + as.factor(sex):as.factor(region)     3     0.062 174.34 -2644.7
## - centeredBMI:as.factor(region)        3     1.570 175.97 -2644.3
## + centeredBMI:as.factor(children)      5     0.401 174.00 -2643.3
## + as.factor(sex):as.factor(children)   5     0.189 174.21 -2641.7
## - centeredSqrtAge:as.factor(sex)       1     1.484 175.89 -2640.9
## + as.factor(children):as.factor(region) 15    1.991 172.41 -2635.6
## - centeredSqrtAge:as.factor(region)    3     2.955 177.36 -2633.8
## - as.factor(children):as.factor(smoker) 5     4.806 179.21 -2623.9
## - centeredSqrtAge:as.factor(children)  5     7.179 181.58 -2606.3
## - centeredBMI:as.factor(smoker)        1    18.717 193.12 -2515.9
## - centeredSqrtAge:as.factor(smoker)    1    43.526 217.93 -2354.2
##
## Step:  AIC=-2651.95
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):as.factor(smoker) +
##     centeredBMI:as.factor(smoker) + centeredBMI:as.factor(region) +
##     as.factor(children):as.factor(smoker) + as.factor(smoker):as.factor(region)
##
##                                       Df Sum of Sq    RSS     AIC
## - centeredSqrtAge:centeredBMI          1     0.190 174.63 -2652.5
## - as.factor(smoker):as.factor(region)  3     0.752 175.19 -2652.2
## <none>                                             174.44 -2651.9
## - as.factor(sex):as.factor(smoker)     1     0.476 174.92 -2650.3
## + as.factor(sex):centeredBMI           1     0.041 174.40 -2650.3
## + as.factor(sex):as.factor(region)     3     0.062 174.38 -2646.4
## - centeredBMI:as.factor(region)        3     1.531 175.97 -2646.3
## + centeredBMI:as.factor(children)      5     0.408 174.03 -2645.1
## + as.factor(sex):as.factor(children)   5     0.190 174.25 -2643.4
## - centeredSqrtAge:as.factor(sex)       1     1.562 176.00 -2642.0
## + as.factor(children):as.factor(region) 15    2.001 172.44 -2637.4
## - centeredSqrtAge:as.factor(region)    3     2.934 177.38 -2635.6
## - as.factor(children):as.factor(smoker) 5     4.779 179.22 -2625.8
## - centeredSqrtAge:as.factor(children)  5     7.218 181.66 -2607.7
## - centeredBMI:as.factor(smoker)        1    18.747 193.19 -2517.4
## - centeredSqrtAge:as.factor(smoker)    1    43.781 218.22 -2354.3
##
## Step:  AIC=-2652.5
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
```

```
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:as.factor(children) +
##     centeredSqrtAge:as.factor(smoker) + centeredSqrtAge:as.factor(region) +
##     as.factor(sex):as.factor(smoker) + centeredBMI:as.factor(smoker) +
##     centeredBMI:as.factor(region) + as.factor(children):as.factor(smoker) +
##     as.factor(smoker):as.factor(region)
##
##                                       Df Sum of Sq    RSS     AIC
## - as.factor(smoker):as.factor(region)    3     0.770 175.40 -2652.6
## <none>                                              174.63 -2652.5
## + centeredSqrtAge:centeredBMI            1     0.190 174.44 -2651.9
## + as.factor(sex):centeredBMI             1     0.048 174.59 -2650.9
## - as.factor(sex):as.factor(smoker)       1     0.507 175.14 -2650.6
## + as.factor(sex):as.factor(region)       3     0.057 174.58 -2646.9
## - centeredBMI:as.factor(region)          3     1.539 176.17 -2646.8
## + centeredBMI:as.factor(children)        5     0.404 174.23 -2645.6
## + as.factor(sex):as.factor(children)     5     0.195 174.44 -2644.0
## - centeredSqrtAge:as.factor(sex)         1     1.537 176.17 -2642.8
## - centeredSqrtAge:as.factor(region)      3     2.747 177.38 -2637.6
## + as.factor(children):as.factor(region) 15     1.962 172.67 -2637.6
## - as.factor(children):as.factor(smoker)  5     4.830 179.46 -2626.0
## - centeredSqrtAge:as.factor(children)    5     7.254 181.89 -2608.0
## - centeredBMI:as.factor(smoker)          1    18.837 193.47 -2517.4
## - centeredSqrtAge:as.factor(smoker)      1    43.792 218.43 -2355.1
##
## Step:  AIC=-2652.61
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:as.factor(children) +
##     centeredSqrtAge:as.factor(smoker) + centeredSqrtAge:as.factor(region) +
##     as.factor(sex):as.factor(smoker) + centeredBMI:as.factor(smoker) +
##     centeredBMI:as.factor(region) + as.factor(children):as.factor(smoker)
##
##                                       Df Sum of Sq    RSS     AIC
## <none>                                              175.40 -2652.6
## + as.factor(smoker):as.factor(region)    3     0.770 174.63 -2652.5
## + centeredSqrtAge:centeredBMI            1     0.208 175.19 -2652.2
## + as.factor(sex):centeredBMI             1     0.039 175.37 -2650.9
## - as.factor(sex):as.factor(smoker)       1     0.538 175.94 -2650.5
## + as.factor(sex):as.factor(region)       3     0.062 175.34 -2647.1
## - centeredBMI:as.factor(region)          3     1.649 177.05 -2646.1
## + centeredBMI:as.factor(children)        5     0.411 174.99 -2645.8
## + as.factor(sex):as.factor(children)     5     0.199 175.21 -2644.1
## - centeredSqrtAge:as.factor(sex)         1     1.539 176.94 -2642.9
## - centeredSqrtAge:as.factor(region)      3     2.602 178.01 -2638.9
## + as.factor(children):as.factor(region) 15     2.061 173.34 -2638.4
## - as.factor(children):as.factor(smoker)  5     4.729 180.13 -2627.0
## - centeredSqrtAge:as.factor(children)    5     7.205 182.61 -2608.7
## - centeredBMI:as.factor(smoker)          1    21.781 197.19 -2498.0
## - centeredSqrtAge:as.factor(smoker)      1    44.417 219.82 -2352.6
```

```
allFitAllSBC = step(allFitAll, direction = "both", k=log(dim(dataset)[1]))
```

```
## Start:  AIC=-2269.69
## log(charges) ~ (centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region))^2
```

```
##
##                                          Df Sum of Sq    RSS     AIC
## - as.factor(children):as.factor(region) 15     1.705 173.71 -2364.5
## - centeredBMI:as.factor(children)        5     0.143 172.14 -2304.6
## - as.factor(sex):as.factor(children)     5     0.211 172.21 -2304.0
## - as.factor(sex):as.factor(region)       3     0.073 172.07 -2290.7
## - as.factor(smoker):as.factor(region)    3     0.653 172.65 -2286.2
## - centeredBMI:as.factor(region)          3     1.469 173.47 -2279.9
## - as.factor(sex):centeredBMI             1     0.025 172.03 -2276.7
## - centeredSqrtAge:centeredBMI            1     0.221 172.22 -2275.2
## - as.factor(sex):as.factor(smoker)       1     0.439 172.44 -2273.5
## - as.factor(children):as.factor(smoker)  5     4.308 176.31 -2272.6
## <none>                                             172.00 -2269.7
## - centeredSqrtAge:as.factor(region)      3     2.955 174.96 -2268.5
## - centeredSqrtAge:as.factor(sex)         1     1.368 173.37 -2266.3
## - centeredSqrtAge:as.factor(children)    5     6.665 178.67 -2254.8
## - centeredBMI:as.factor(smoker)          1    18.544 190.54 -2139.9
## - centeredSqrtAge:as.factor(smoker)      1    42.019 214.02 -1984.4
##
## Step:  AIC=-2364.47
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(children) + as.factor(sex):as.factor(smoker) +
##     as.factor(sex):as.factor(region) + centeredBMI:as.factor(children) +
##     centeredBMI:as.factor(smoker) + centeredBMI:as.factor(region) +
##     as.factor(children):as.factor(smoker) + as.factor(smoker):as.factor(region)
##
##                                          Df Sum of Sq    RSS     AIC
## - as.factor(sex):as.factor(children)     5     0.202 173.91 -2398.9
## - centeredBMI:as.factor(children)        5     0.450 174.16 -2397.0
## - as.factor(sex):as.factor(region)       3     0.087 173.79 -2385.4
## - as.factor(smoker):as.factor(region)    3     0.756 174.46 -2380.3
## - centeredBMI:as.factor(region)          3     1.603 175.31 -2373.8
## - as.factor(sex):centeredBMI             1     0.031 173.74 -2371.4
## - centeredSqrtAge:centeredBMI            1     0.190 173.90 -2370.2
## - as.factor(sex):as.factor(smoker)       1     0.455 174.16 -2368.2
## - as.factor(children):as.factor(smoker)  5     4.649 178.35 -2365.1
## <none>                                             173.71 -2364.5
## - centeredSqrtAge:as.factor(region)      3     2.837 176.54 -2364.4
## - centeredSqrtAge:as.factor(sex)         1     1.513 175.22 -2360.1
## - centeredSqrtAge:as.factor(children)    5     6.920 180.63 -2348.2
## + as.factor(children):as.factor(region) 15     1.705 172.00 -2269.7
## - centeredBMI:as.factor(smoker)          1    18.765 192.47 -2234.4
## - centeredSqrtAge:as.factor(smoker)      1    43.602 217.31 -2072.0
##
## Step:  AIC=-2398.92
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
```

```
##     as.factor(sex):as.factor(smoker) + as.factor(sex):as.factor(region) +
##     centeredBMI:as.factor(children) + centeredBMI:as.factor(smoker) +
##     centeredBMI:as.factor(region) + as.factor(children):as.factor(smoker) +
##     as.factor(smoker):as.factor(region)
##
##                                        Df Sum of Sq    RSS     AIC
## - centeredBMI:as.factor(children)       5     0.432 174.34 -2431.6
## - as.factor(sex):as.factor(region)      3     0.094 174.00 -2419.8
## - as.factor(smoker):as.factor(region)   3     0.753 174.66 -2414.7
## - centeredBMI:as.factor(region)         3     1.562 175.47 -2408.6
## - as.factor(sex):centeredBMI            1     0.032 173.94 -2405.9
## - centeredSqrtAge:centeredBMI           1     0.195 174.10 -2404.6
## - as.factor(sex):as.factor(smoker)      1     0.481 174.39 -2402.4
## <none>                                              173.91 -2398.9
## - as.factor(children):as.factor(smoker)  5     4.760 178.67 -2398.8
## - centeredSqrtAge:as.factor(region)     3     2.878 176.78 -2398.6
## - centeredSqrtAge:as.factor(sex)        1     1.512 175.42 -2394.5
## - centeredSqrtAge:as.factor(children)   5     7.039 180.95 -2381.8
## + as.factor(sex):as.factor(children)    5     0.202 173.71 -2364.5
## + as.factor(children):as.factor(region) 15    1.696 172.21 -2304.0
## - centeredBMI:as.factor(smoker)         1    18.861 192.77 -2268.3
## - centeredSqrtAge:as.factor(smoker)     1    43.578 217.49 -2106.9
##
## Step:  AIC=-2431.59
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(smoker) + as.factor(sex):as.factor(region) +
##     centeredBMI:as.factor(smoker) + centeredBMI:as.factor(region) +
##     as.factor(children):as.factor(smoker) + as.factor(smoker):as.factor(region)
##
##                                        Df Sum of Sq    RSS     AIC
## - as.factor(sex):as.factor(region)      3     0.062 174.40 -2452.7
## - as.factor(smoker):as.factor(region)   3     0.761 175.10 -2447.4
## - centeredBMI:as.factor(region)         3     1.571 175.91 -2441.2
## - as.factor(sex):centeredBMI            1     0.041 174.38 -2438.5
## - centeredSqrtAge:centeredBMI           1     0.189 174.53 -2437.3
## - as.factor(sex):as.factor(smoker)      1     0.476 174.82 -2435.1
## <none>                                              174.34 -2431.6
## - as.factor(children):as.factor(smoker)  5     4.844 179.18 -2430.9
## - centeredSqrtAge:as.factor(region)     3     2.960 177.30 -2430.7
## - centeredSqrtAge:as.factor(sex)        1     1.485 175.82 -2427.4
## - centeredSqrtAge:as.factor(children)   5     7.216 181.56 -2413.3
## + centeredBMI:as.factor(children)       5     0.432 173.91 -2398.9
## + as.factor(sex):as.factor(children)    5     0.184 174.16 -2397.0
## + as.factor(children):as.factor(region) 15    1.990 172.35 -2339.0
## - centeredBMI:as.factor(smoker)         1    18.638 192.98 -2302.9
## - centeredSqrtAge:as.factor(smoker)     1    43.484 217.82 -2140.8
##
## Step:  AIC=-2452.71
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
```

```
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(smoker) + centeredBMI:as.factor(smoker) +
##     centeredBMI:as.factor(region) + as.factor(children):as.factor(smoker) +
##     as.factor(smoker):as.factor(region)
##
##                                        Df Sum of Sq    RSS     AIC
## - as.factor(smoker):as.factor(region)   3     0.761 175.16 -2468.5
## - centeredBMI:as.factor(region)         3     1.570 175.97 -2462.3
## - as.factor(sex):centeredBMI            1     0.041 174.44 -2459.6
## - centeredSqrtAge:centeredBMI           1     0.183 174.59 -2458.5
## - as.factor(sex):as.factor(smoker)      1     0.476 174.88 -2456.3
## <none>                                              174.40 -2452.7
## - as.factor(children):as.factor(smoker) 5     4.806 179.21 -2452.3
## - centeredSqrtAge:as.factor(region)     3     2.955 177.36 -2451.8
## - centeredSqrtAge:as.factor(sex)        1     1.484 175.89 -2448.6
## - centeredSqrtAge:as.factor(children)   5     7.179 181.58 -2434.7
## + as.factor(sex):as.factor(region)      3     0.062 174.34 -2431.6
## + centeredBMI:as.factor(children)       5     0.401 174.00 -2419.8
## + as.factor(sex):as.factor(children)    5     0.189 174.21 -2418.2
## + as.factor(children):as.factor(region) 15    1.991 172.41 -2360.1
## - centeredBMI:as.factor(smoker)         1    18.717 193.12 -2323.5
## - centeredSqrtAge:as.factor(smoker)     1    43.526 217.93 -2161.8
##
## Step:  AIC=-2468.48
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(smoker) + centeredBMI:as.factor(smoker) +
##     centeredBMI:as.factor(region) + as.factor(children):as.factor(smoker)
##
##                                        Df Sum of Sq    RSS     AIC
## - centeredBMI:as.factor(region)         3     1.671 176.84 -2477.4
## - as.factor(sex):centeredBMI            1     0.032 175.19 -2475.4
## - centeredSqrtAge:centeredBMI           1     0.202 175.37 -2474.1
## - as.factor(sex):as.factor(smoker)      1     0.505 175.67 -2471.8
## - as.factor(children):as.factor(smoker) 5     4.699 179.86 -2469.1
## - centeredSqrtAge:as.factor(region)     3     2.822 177.99 -2468.7
## <none>                                              175.16 -2468.5
## - centeredSqrtAge:as.factor(sex)        1     1.494 176.66 -2464.3
## + as.factor(smoker):as.factor(region)   3     0.761 174.40 -2452.7
## - centeredSqrtAge:as.factor(children)   5     7.132 182.29 -2451.1
## + as.factor(sex):as.factor(region)      3     0.062 175.10 -2447.4
## + centeredBMI:as.factor(children)       5     0.406 174.76 -2435.6
## + as.factor(sex):as.factor(children)    5     0.192 174.97 -2433.9
## + as.factor(children):as.factor(region) 15    2.090 173.07 -2376.6
## - centeredBMI:as.factor(smoker)         1    21.633 196.80 -2319.9
## - centeredSqrtAge:as.factor(smoker)     1    44.170 219.33 -2174.8
##
## Step:  AIC=-2477.37
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
```

```
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):centeredBMI +
##     as.factor(sex):as.factor(smoker) + centeredBMI:as.factor(smoker) +
##     as.factor(children):as.factor(smoker)
##
##                                        Df Sum of Sq    RSS     AIC
## - as.factor(sex):centeredBMI            1     0.000 176.84 -2484.6
## - centeredSqrtAge:centeredBMI           1     0.217 177.05 -2482.9
## - centeredSqrtAge:as.factor(region)     3     2.512 179.35 -2480.1
## - as.factor(sex):as.factor(smoker)      1     0.595 177.43 -2480.1
## - as.factor(children):as.factor(smoker) 5     4.674 181.51 -2478.5
## <none>                                              176.84 -2477.4
## - centeredSqrtAge:as.factor(sex)        1     1.618 178.45 -2472.4
## + centeredBMI:as.factor(region)         3     1.671 175.16 -2468.5
## + as.factor(smoker):as.factor(region)   3     0.863 175.97 -2462.3
## - centeredSqrtAge:as.factor(children)   5     7.044 183.88 -2461.1
## + as.factor(sex):as.factor(region)      3     0.067 176.77 -2456.3
## + centeredBMI:as.factor(children)       5     0.413 176.42 -2444.5
## + as.factor(sex):as.factor(children)    5     0.156 176.68 -2442.6
## + as.factor(children):as.factor(region) 15    2.242 174.59 -2386.4
## - centeredBMI:as.factor(smoker)         1    21.197 198.03 -2333.1
## - centeredSqrtAge:as.factor(smoker)     1    43.866 220.70 -2188.1
##
## Step:  AIC=-2484.56
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:centeredBMI +
##     centeredSqrtAge:as.factor(children) + centeredSqrtAge:as.factor(smoker) +
##     centeredSqrtAge:as.factor(region) + as.factor(sex):as.factor(smoker) +
##     centeredBMI:as.factor(smoker) + as.factor(children):as.factor(smoker)
##
##                                        Df Sum of Sq    RSS     AIC
## - centeredSqrtAge:centeredBMI           1     0.218 177.05 -2490.1
## - centeredSqrtAge:as.factor(region)     3     2.512 179.35 -2487.3
## - as.factor(sex):as.factor(smoker)      1     0.595 177.43 -2487.3
## - as.factor(children):as.factor(smoker) 5     4.677 181.51 -2485.6
## <none>                                              176.84 -2484.6
## - centeredSqrtAge:as.factor(sex)        1     1.643 178.48 -2479.4
## + as.factor(sex):centeredBMI            1     0.000 176.84 -2477.4
## + centeredBMI:as.factor(region)         3     1.640 175.19 -2475.4
## + as.factor(smoker):as.factor(region)   3     0.861 175.97 -2469.5
## - centeredSqrtAge:as.factor(children)   5     7.054 183.89 -2468.2
## + as.factor(sex):as.factor(region)      3     0.066 176.77 -2463.5
## + centeredBMI:as.factor(children)       5     0.413 176.42 -2451.7
## + as.factor(sex):as.factor(children)    5     0.156 176.68 -2449.8
## + as.factor(children):as.factor(region) 15    2.242 174.59 -2393.7
## - centeredBMI:as.factor(smoker)         1    21.210 198.04 -2340.2
## - centeredSqrtAge:as.factor(smoker)     1    44.002 220.84 -2194.4
##
## Step:  AIC=-2490.12
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
```

```
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:as.factor(children) +
##     centeredSqrtAge:as.factor(smoker) + centeredSqrtAge:as.factor(region) +
##     as.factor(sex):as.factor(smoker) + centeredBMI:as.factor(smoker) +
##     as.factor(children):as.factor(smoker)
##
##                                        Df Sum of Sq    RSS     AIC
## - centeredSqrtAge:as.factor(region)     3     2.305 179.36 -2494.4
## - as.factor(sex):as.factor(smoker)      1     0.634 177.69 -2492.5
## - as.factor(children):as.factor(smoker) 5     4.730 181.78 -2490.8
## <none>                                              177.05 -2490.1
## - centeredSqrtAge:as.factor(sex)        1     1.616 178.67 -2485.2
## + centeredSqrtAge:centeredBMI           1     0.218 176.84 -2484.6
## + as.factor(sex):centeredBMI            1     0.001 177.05 -2482.9
## + centeredBMI:as.factor(region)         3     1.649 175.40 -2481.0
## + as.factor(smoker):as.factor(region)   3     0.881 176.17 -2475.2
## - centeredSqrtAge:as.factor(children)   5     7.093 184.15 -2473.6
## + as.factor(sex):as.factor(region)      3     0.063 176.99 -2469.0
## + centeredBMI:as.factor(children)       5     0.409 176.64 -2457.2
## + as.factor(sex):as.factor(children)    5     0.161 176.89 -2455.3
## + as.factor(children):as.factor(region) 15    2.205 174.85 -2398.9
## - centeredBMI:as.factor(smoker)         1    21.344 198.40 -2345.0
## - centeredSqrtAge:as.factor(smoker)     1    44.016 221.07 -2200.2
##
## Step:  AIC=-2494.41
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:as.factor(children) +
##     centeredSqrtAge:as.factor(smoker) + as.factor(sex):as.factor(smoker) +
##     centeredBMI:as.factor(smoker) + as.factor(children):as.factor(smoker)
##
##                                        Df Sum of Sq    RSS     AIC
## - as.factor(sex):as.factor(smoker)      1     0.660 180.02 -2496.7
## <none>                                              179.36 -2494.4
## - as.factor(children):as.factor(smoker) 5     4.921 184.28 -2494.2
## + centeredSqrtAge:as.factor(region)     3     2.305 177.05 -2490.1
## - centeredSqrtAge:as.factor(sex)        1     1.620 180.98 -2489.6
## + centeredSqrtAge:centeredBMI           1     0.010 179.35 -2487.3
## + as.factor(sex):centeredBMI            1     0.000 179.36 -2487.2
## + centeredBMI:as.factor(region)         3     1.352 178.01 -2482.9
## - centeredSqrtAge:as.factor(children)   5     6.933 186.29 -2479.7
## - as.factor(region)                     3     4.942 184.30 -2479.6
## + as.factor(smoker):as.factor(region)   3     0.735 178.62 -2478.3
## + as.factor(sex):as.factor(region)      3     0.059 179.30 -2473.2
## + centeredBMI:as.factor(children)       5     0.490 178.87 -2462.1
## + as.factor(sex):as.factor(children)    5     0.199 179.16 -2459.9
## + as.factor(children):as.factor(region) 15    2.138 177.22 -2402.5
## - centeredBMI:as.factor(smoker)         1    20.877 200.24 -2354.3
## - centeredSqrtAge:as.factor(smoker)     1    43.570 222.93 -2210.6
##
## Step:  AIC=-2496.69
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:as.factor(children) +
##     centeredSqrtAge:as.factor(smoker) + centeredBMI:as.factor(smoker) +
```

```
##     as.factor(children):as.factor(smoker)
##
##                                         Df Sum of Sq    RSS     AIC
## - as.factor(children):as.factor(smoker)  5     4.673 184.69 -2498.4
## <none>                                               180.02 -2496.7
## + as.factor(sex):as.factor(smoker)       1     0.660 179.36 -2494.4
## + centeredSqrtAge:as.factor(region)      3     2.331 177.69 -2492.5
## - centeredSqrtAge:as.factor(sex)         1     1.586 181.60 -2492.2
## + centeredSqrtAge:centeredBMI            1     0.019 180.00 -2489.6
## + as.factor(sex):centeredBMI             1     0.001 180.02 -2489.5
## + centeredBMI:as.factor(region)          3     1.432 178.59 -2485.8
## - centeredSqrtAge:as.factor(children)    5     6.833 186.85 -2482.8
## - as.factor(region)                      3     4.885 184.90 -2482.5
## + as.factor(smoker):as.factor(region)    3     0.770 179.25 -2480.8
## + as.factor(sex):as.factor(region)       3     0.047 179.97 -2475.4
## + centeredBMI:as.factor(children)        5     0.476 179.54 -2464.2
## + as.factor(sex):as.factor(children)     5     0.223 179.79 -2462.3
## + as.factor(children):as.factor(region) 15     2.155 177.86 -2404.8
## - centeredBMI:as.factor(smoker)          1    22.070 202.09 -2349.2
## - centeredSqrtAge:as.factor(smoker)      1    43.750 223.77 -2212.8
##
## Step:  AIC=-2498.4
## log(charges) ~ centeredSqrtAge + as.factor(sex) + centeredBMI +
##     as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:as.factor(children) +
##     centeredSqrtAge:as.factor(smoker) + centeredBMI:as.factor(smoker)
##
##                                         Df Sum of Sq    RSS     AIC
## <none>                                               184.69 -2498.4
## + as.factor(children):as.factor(smoker)  5     4.673 180.02 -2496.7
## + centeredSqrtAge:as.factor(region)      3     2.510 182.18 -2495.1
## - centeredSqrtAge:as.factor(sex)         1     1.526 186.22 -2494.6
## + as.factor(sex):as.factor(smoker)       1     0.412 184.28 -2494.2
## + centeredSqrtAge:centeredBMI            1     0.021 184.67 -2491.3
## + as.factor(sex):centeredBMI             1     0.011 184.68 -2491.3
## - centeredSqrtAge:as.factor(children)    5     6.249 190.94 -2489.9
## + centeredBMI:as.factor(region)          3     1.414 183.28 -2487.1
## - as.factor(region)                      3     4.913 189.60 -2484.9
## + as.factor(smoker):as.factor(region)    3     0.674 184.02 -2481.7
## + as.factor(sex):as.factor(region)       3     0.006 184.69 -2476.8
## + centeredBMI:as.factor(children)        5     0.588 184.10 -2466.7
## + as.factor(sex):as.factor(children)     5     0.346 184.34 -2464.9
## + as.factor(children):as.factor(region) 15     2.583 182.11 -2409.3
## - centeredBMI:as.factor(smoker)          1    22.167 206.86 -2353.9
## - centeredSqrtAge:as.factor(smoker)      1    47.326 232.02 -2200.4
```

Out first stepwise regression model had 14 predictors which was a lot more than desired maximum of 7 predictors (Dr. Iversen, Olin 113, 2019). For this reason, we performed the step wise regression with SBC criterion. The SBC model yielded a model with 10 predictor variables which is easier to handle than that with 14 predictors so we decided to move forward with this model.

Final model

Numerical Summary

```
summary(allFitAllSBC)
```

```
##
## Call:
## lm(formula = log(charges) ~ centeredSqrtAge + as.factor(sex) +
##     centeredBMI + as.factor(children) + as.factor(smoker) + as.factor(region) +
##     centeredSqrtAge:as.factor(sex) + centeredSqrtAge:as.factor(children) +
##     centeredSqrtAge:as.factor(smoker) + centeredBMI:as.factor(smoker),
##     data = dataset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.60973 -0.14571 -0.09211 -0.02856  2.52207
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                        8.802628   0.026383 333.642  < 2e-16
## centeredSqrtAge                    0.519140   0.015119  34.336  < 2e-16
## as.factor(sex)male                -0.091260   0.020622  -4.425 1.04e-05
## centeredBMI                        0.001554   0.001986   0.783 0.433941
## as.factor(children)1               0.124002   0.026091   4.753 2.23e-06
## as.factor(children)2               0.262155   0.028909   9.068  < 2e-16
## as.factor(children)3               0.238579   0.034501   6.915 7.27e-12
## as.factor(children)4               0.483783   0.076647   6.312 3.76e-10
## as.factor(children)5               0.419767   0.093578   4.486 7.90e-06
## as.factor(smoker)yes               1.536453   0.025637  59.930  < 2e-16
## as.factor(region)northwest        -0.065860   0.029499  -2.233 0.025743
## as.factor(region)southeast        -0.144986   0.029728  -4.877 1.21e-06
## as.factor(region)southwest        -0.152617   0.029581  -5.159 2.86e-07
## centeredSqrtAge:as.factor(sex)male 0.058882   0.017851   3.299 0.000998
## centeredSqrtAge:as.factor(children)1 -0.044065   0.024021  -1.834 0.066811
## centeredSqrtAge:as.factor(children)2 -0.162976   0.028115  -5.797 8.45e-09
## centeredSqrtAge:as.factor(children)3 -0.096724   0.032976  -2.933 0.003413
## centeredSqrtAge:as.factor(children)4 -0.203788   0.077614  -2.626 0.008749
## centeredSqrtAge:as.factor(children)5 -0.172023   0.103628  -1.660 0.097151
## centeredSqrtAge:as.factor(smoker)yes -0.411108   0.022379 -18.371  < 2e-16
## centeredBMI:as.factor(smoker)yes    0.051352   0.004084  12.573  < 2e-16
##
## (Intercept)                       ***
## centeredSqrtAge                   ***
## as.factor(sex)male                ***
## centeredBMI
## as.factor(children)1              ***
## as.factor(children)2              ***
## as.factor(children)3              ***
## as.factor(children)4              ***
## as.factor(children)5              ***
## as.factor(smoker)yes              ***
## as.factor(region)northwest        *
```

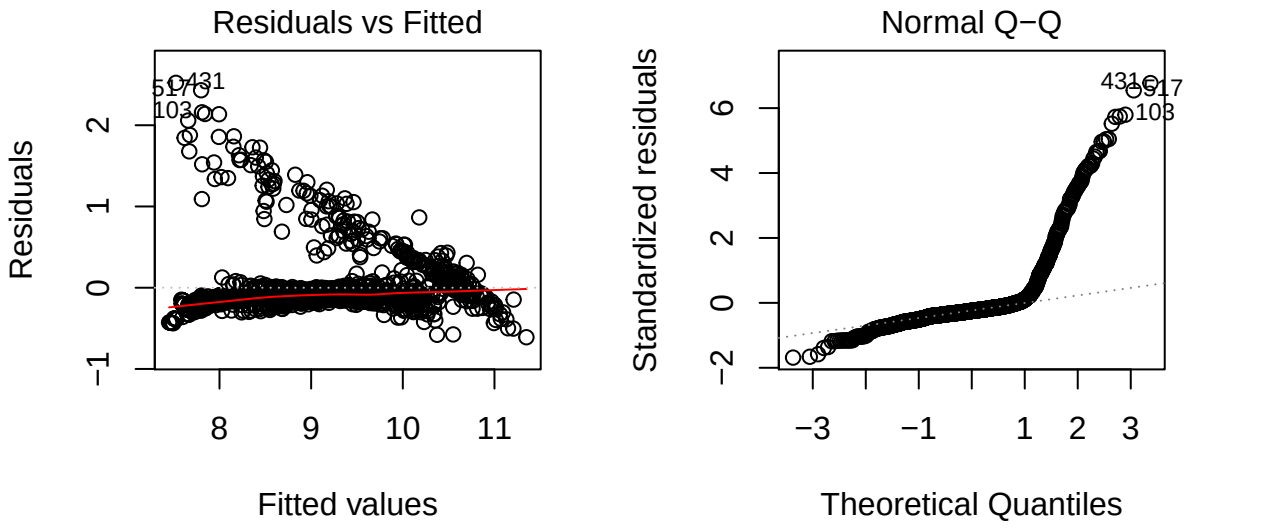
```
## as.factor(region)southeast           ***
## as.factor(region)southwest           ***
## centeredSqrtAge:as.factor(sex)male   ***
## centeredSqrtAge:as.factor(children)1 .
## centeredSqrtAge:as.factor(children)2 ***
## centeredSqrtAge:as.factor(children)3 **
## centeredSqrtAge:as.factor(children)4 **
## centeredSqrtAge:as.factor(children)5 .
## centeredSqrtAge:as.factor(smoker)yes ***
## centeredBMI:as.factor(smoker)yes     ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3745 on 1317 degrees of freedom
## Multiple R-squared:  0.8366, Adjusted R-squared:  0.8341
## F-statistic: 337.2 on 20 and 1317 DF,  p-value: < 2.2e-16
```

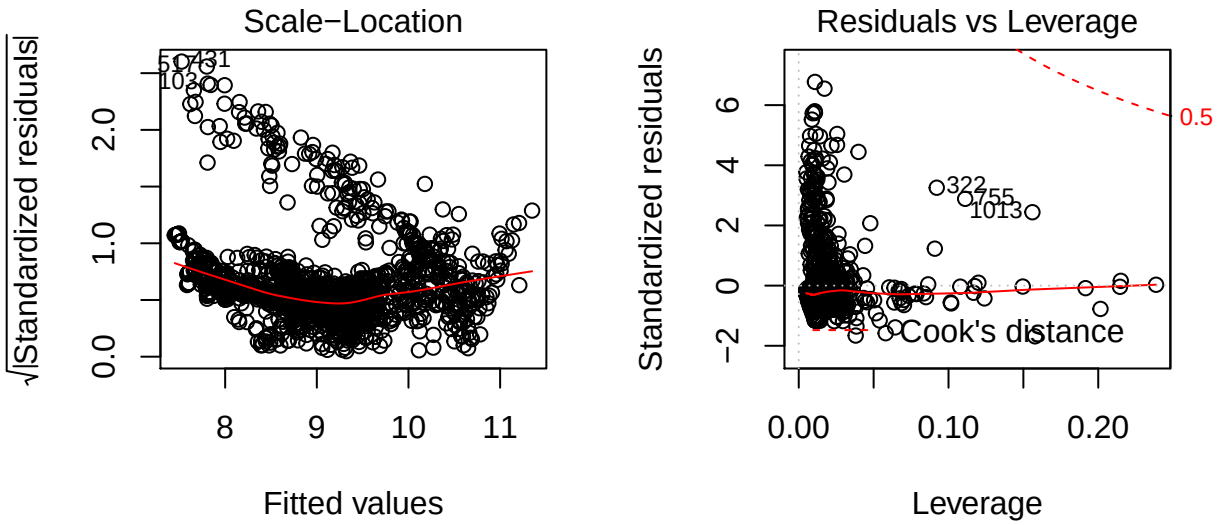
The Adjusted R-squared value is 0.8341 which tells us that 83.41% of the variation in charges is explained by this model. The result is a significant improvement over the previous model which had the Adjusted R-squared value of 0.7685 (76.85% of the variation in charges is explained by the model).

Residual standard error is 0.3745 meaning that on average, predictions of the model are 0.3745 `log`(dollars) away from the real value. The residual standard error went down as compared to the previous model with the value of 0.4425 `log`(dollars).

Residual Plots

```
par(mfrow=c(1,2))
plot(allFitAllSBC)
```





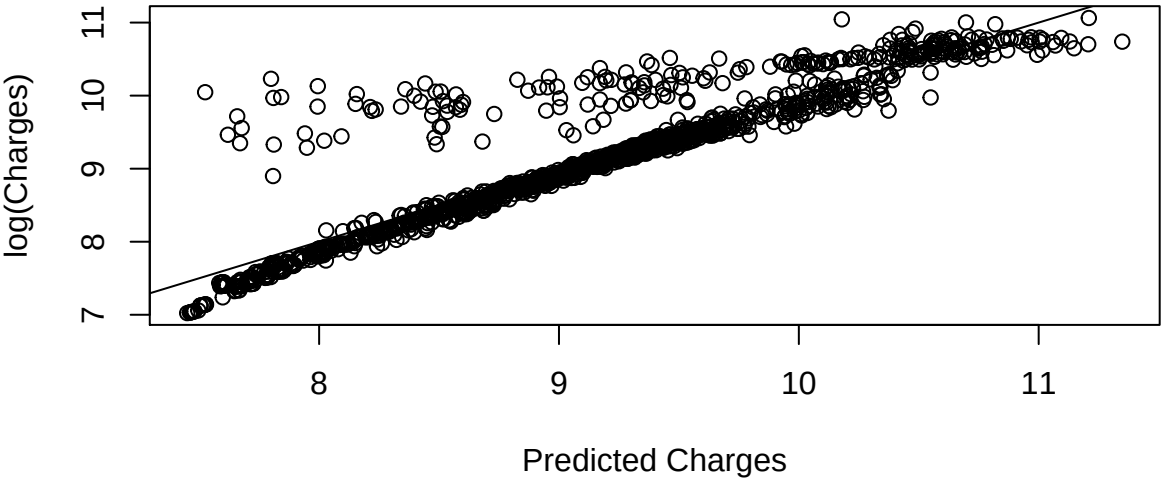
The first plot (Residuals VS Fitted Values) shows virtually no signs of curvature and that the spread is constant. This is a big improvement over the previous model. There is no evidence of heteroscedasticity `allFit1`.

The second plot (Standardized Residuals VS Theoretical Quantiles) does give an evidence that against the claim that both sets of residuals are coming from normal distributions. That said, the shape of the plot looks a lot better than that of the previous model `allFit1`.

The third plot ($\sqrt{\text{Standardized Residuals}}$ VS Fitted values) shows rather straight line with a slight curvature. This suggests that the residuals are, on average, spread equally along the ranges of predictors and that the variance is, on average, constant. This is an improvement over the initial model `allFit1`.

Residuals VS Leverage plot shows no significant outliers.

```
plot(log(dataset$charges) ~ allFitAllSBC$fitted.values, ylab="log(Charges)", xlab="Predicted Charges")
abline(0,1)
```

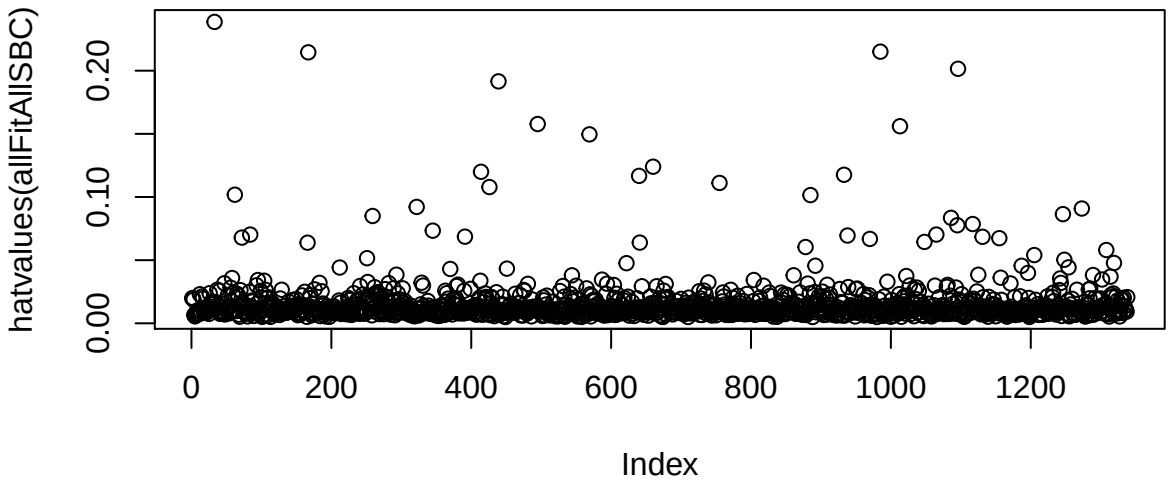


Let's now plot response variables VS fitted values and add the line with intercept 0 and slope 1 to visually evaluate the model.

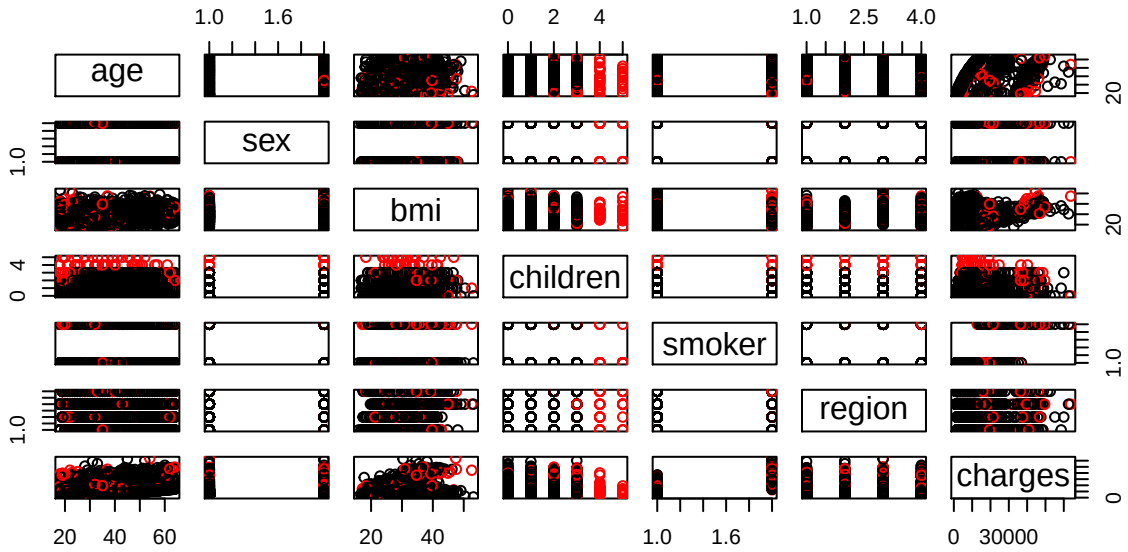
Additional Diagnostics

Hat Matrix Diagonals

```
plot(hatvalues(allFitAllSBC))
```




```
threshold2 = 2 * length(allFitAllSBC$coefficients)/length(dataset$age)
plot(dataset, col=ifelse(hatvalues(allFitAllSBC) > threshold2,2,1))
```

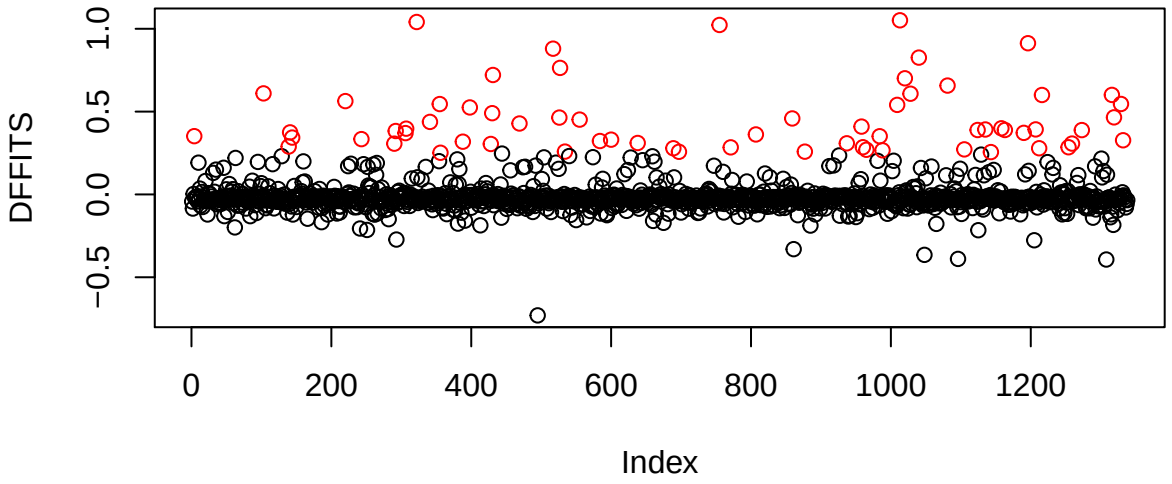


The threshold is $\frac{2p}{n}$. There will almost always be values over the threshold as they're only general outlines. vertical interpretation but not horizontal. The hat matrix diagonals show us that the points with the highest leverage are the individuals with more children. These are explained through the scatterplot matrix(Line 341), in the graph of children vs age.

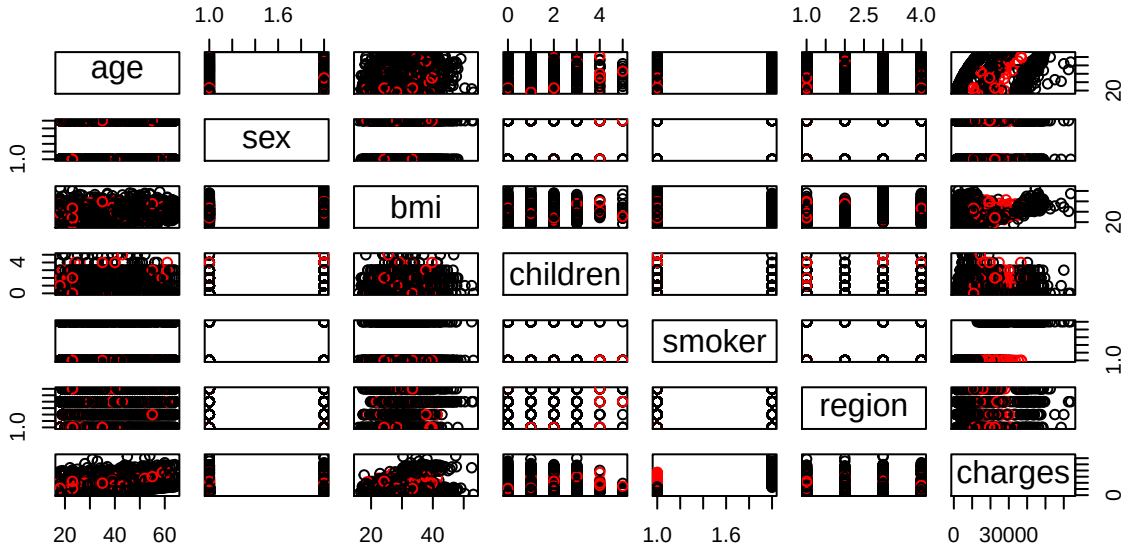
DFFITS

```
threshold <- 2 * sqrt(length(allFitAllSBC$coefficients)/length(dataset$age))
DFFITS <- dffits(allFitAllSBC)
```

```
# Model with colors
plot(DFFITS, col=(DFFITS > threshold) + 1)
```




```
plot(dataset, col=ifelse(dffits(allFitAllSBC) > threshold,2,1))
```



We have a number of outliers as the threshold value calculated by the formula $2\sqrt{\frac{p}{n}}$ and has a value of 0.251.

*We have colored the possible outliers red. The scatterplot matrix (line 354) indicates an odd clustering formed between age and charges. The values over the threshold seem to cluster around the middle values.

VIF (Variance Inflation Factor)

```
library(car)
```

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

```
vif(allFitAllSBC)
```

```
##                                   GVIF Df GVIF^(1/(2*Df))
## centeredSqrtAge               2.905291  1        1.704492
## as.factor(sex)                1.014229  1        1.007089
```

```
## centeredBMI                          1.398220  1        1.182463
## as.factor(children)                  1.166555  5        1.015525
## as.factor(smoker)                    1.021240  1        1.010564
## as.factor(region)                    1.126837  3        1.020102
## centeredSqrtAge:as.factor(sex)       2.050878  1        1.432089
## centeredSqrtAge:as.factor(children) 1.972008  5        1.070264
## centeredSqrtAge:as.factor(smoker)    1.281435  1        1.132005
## centeredBMI:as.factor(smoker)        1.296649  1        1.138705
```

VIF (the second column in the output) values are all less than 5 meaning that there is no indication of a problematic amount of collinearity.

10-Fold Cross Validation (using `caret` package)

Cross-validation is a set of methods that splits the data into a "training" set, on which to fit and select a model, and a "test" set used to evaluate the model fitted values. Here we illustrate 10-fold cross validation. For this purpose, we used package `caret`.

```
# Use libraries
library(caret)
```

```
## Loading required package: lattice
```

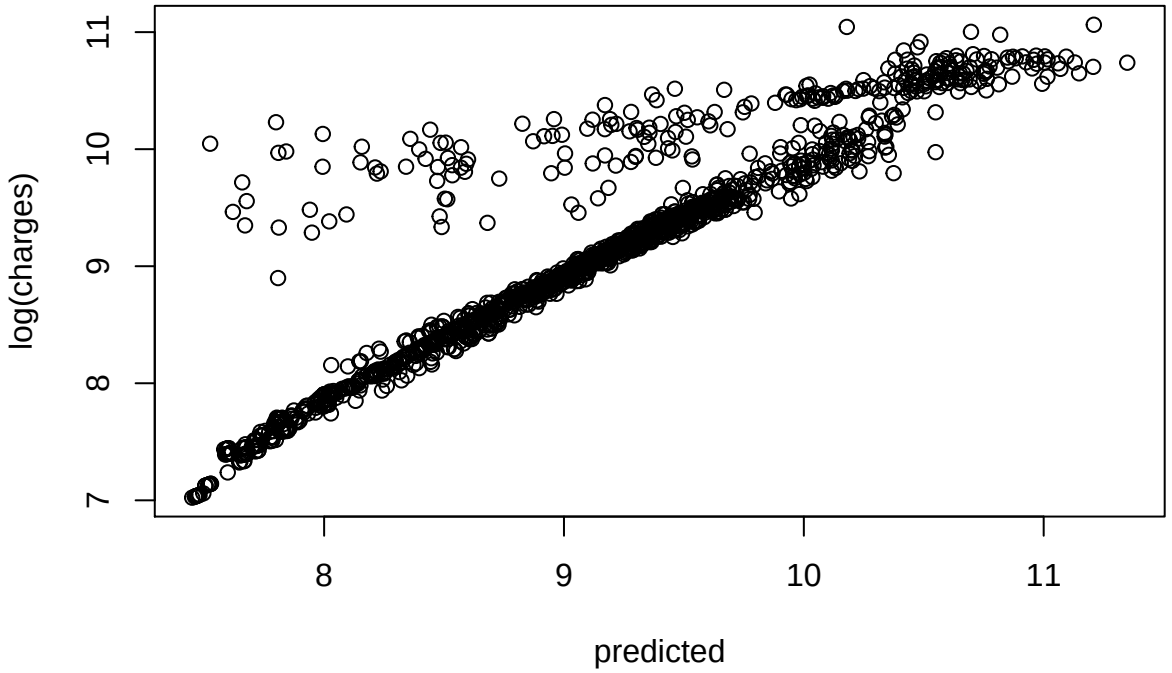
```
## Loading required package: ggplot2
```

```
# Random seed
set.seed(42069)

# Define training control (Leave-One-Out Cross-Validation)
# `number=10` specifies the number of "folds" - 10 in this case hence, 10-fold
# cross-validation
trainContr <- trainControl(method="LOOCV", number=10)

# `caret` complains about columns not being in the actual dataset so we added them
dataset$centeredSqrtAge <- centeredSqrtAge
dataset$centeredBMI <- centeredBMI

# Fits a linear model and does the 10-fold cross-validation
model <- train(log(charges) ~ centeredSqrtAge +
                              as.factor(sex) +
                              centeredBMI +
                              as.factor(children) +
                              as.factor(smoker) +
                              as.factor(region) +
                              centeredSqrtAge:as.factor(sex) +
                              centeredSqrtAge:as.factor(children) +
                              centeredSqrtAge:as.factor(smoker) +
                              centeredBMI:as.factor(smoker),
               trControl=trainContr,
               method="lm",
               data=dataset)

# Summarize results (numerical)
print(model)
```

```
## Linear Regression
```

```
##
## 1338 samples
##    6 predictor
##
## No pre-processing
## Resampling: Leave-One-Out Cross-Validation
## Summary of sample sizes: 1337, 1337, 1337, 1337, 1337, 1337, ...
## Resampling results:
##
##   RMSE       Rsquared  MAE
##   0.3776655  0.831198  0.2112711
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

```r
# Get the predicted values and
# plot actual VS predicted values
predicted <- predict(model, dataset)
plot(log(charges) ~ predicted, data=dataset)
```



Our 10-fold-cross-validated model gave us the R squared value of 0.8311 which is a bit below 0.8341 (to be expected). This verified that our model is indeed performing well and, on average, has the R squared value of 0.8311 (83.11% of the variation in `log`(charges) is explained by the model).

The plot shows a nice correlation between `log`(charges) and predicted values. Recall that we would like the fitted line to be approximately the same as $y = x$ line, which seems to be the case here. Some points do not seem to follow the fitted line, but the number of these points is negligible relative to the ones that are aligned. Overall, the plot further reinforces our numerical observations and, once again, tells us that our model is ready to go!