Project

Hyun Suk (Max) Ryoo (hr2ee)

11/11/2021

```
## Data Processing
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.0.2

## -- Attaching packages --------------------------------------------------------------- tidyverse 1.3

## v ggplot2 3.3.2     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.2
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.4.0     v forcats 0.5.0

## Warning: package 'ggplot2' was built under R version 4.0.2

## Warning: package 'tidyr' was built under R version 4.0.2

## Warning: package 'readr' was built under R version 4.0.2

## Warning: package 'dplyr' was built under R version 4.0.2

## Warning: package 'stringr' was built under R version 4.0.2

## Warning: package 'forcats' was built under R version 4.0.2

## -- Conflicts ------------------------------------------------------------------ tidyverse_conflicts
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(MASS)
```

```
## Warning: package 'MASS' was built under R version 4.0.2

##
## Attaching package: 'MASS'

## The following object is masked from 'package:dplyr':
##
##     select
```

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.0.2
```

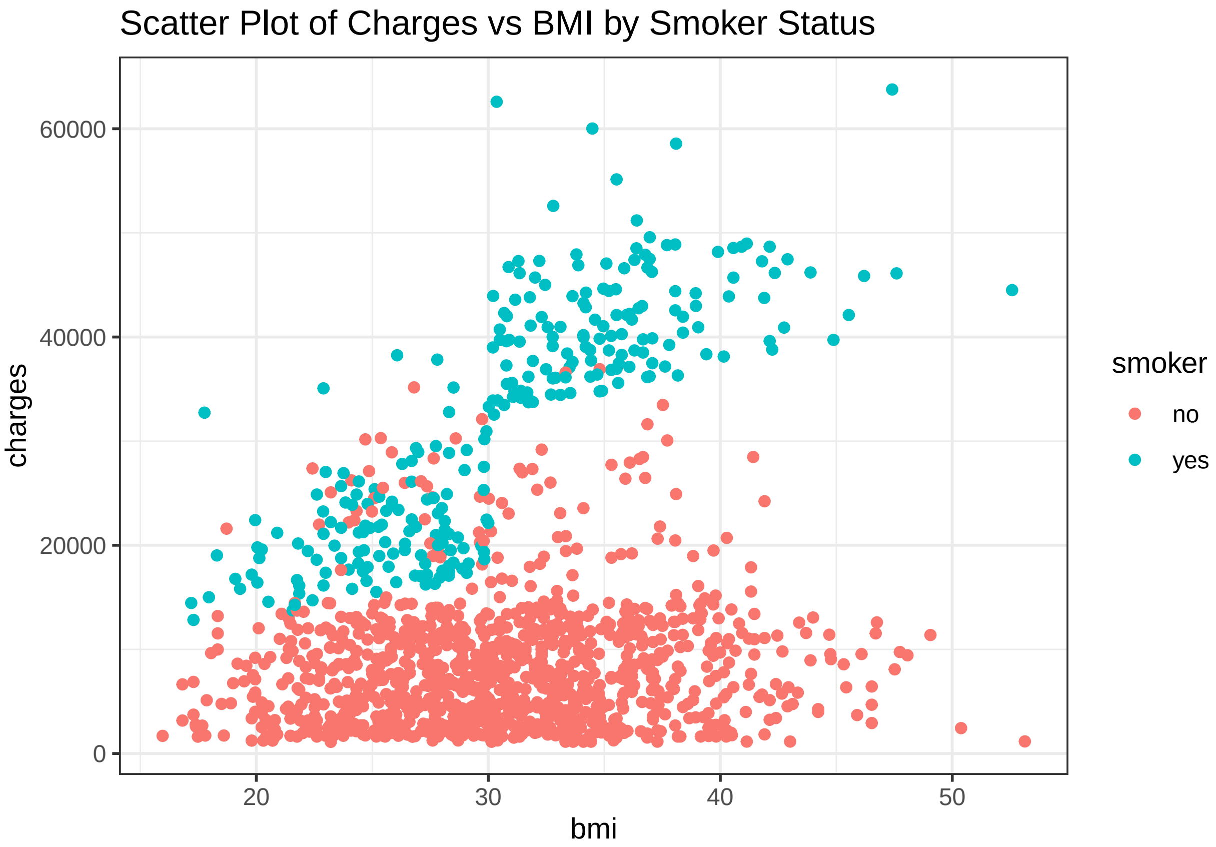
```
setwd("/Users/maxryoo/Documents/MSDS/STAT6021/Project2")
data <- read.csv("data/insurance.csv")
head(data)
```

```
##   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
```

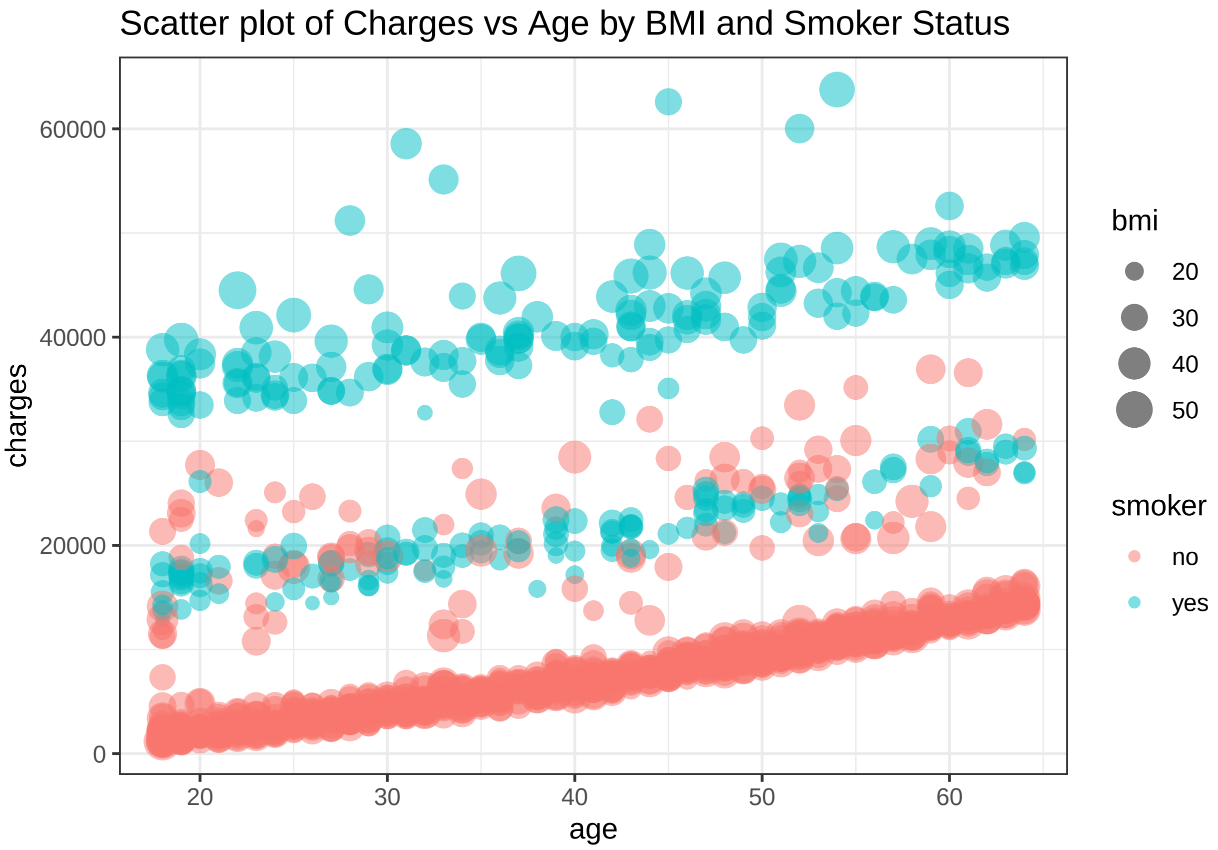
```
data$significant.charge = as.factor(data$charges > median(data$charges))
head(data)
```

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

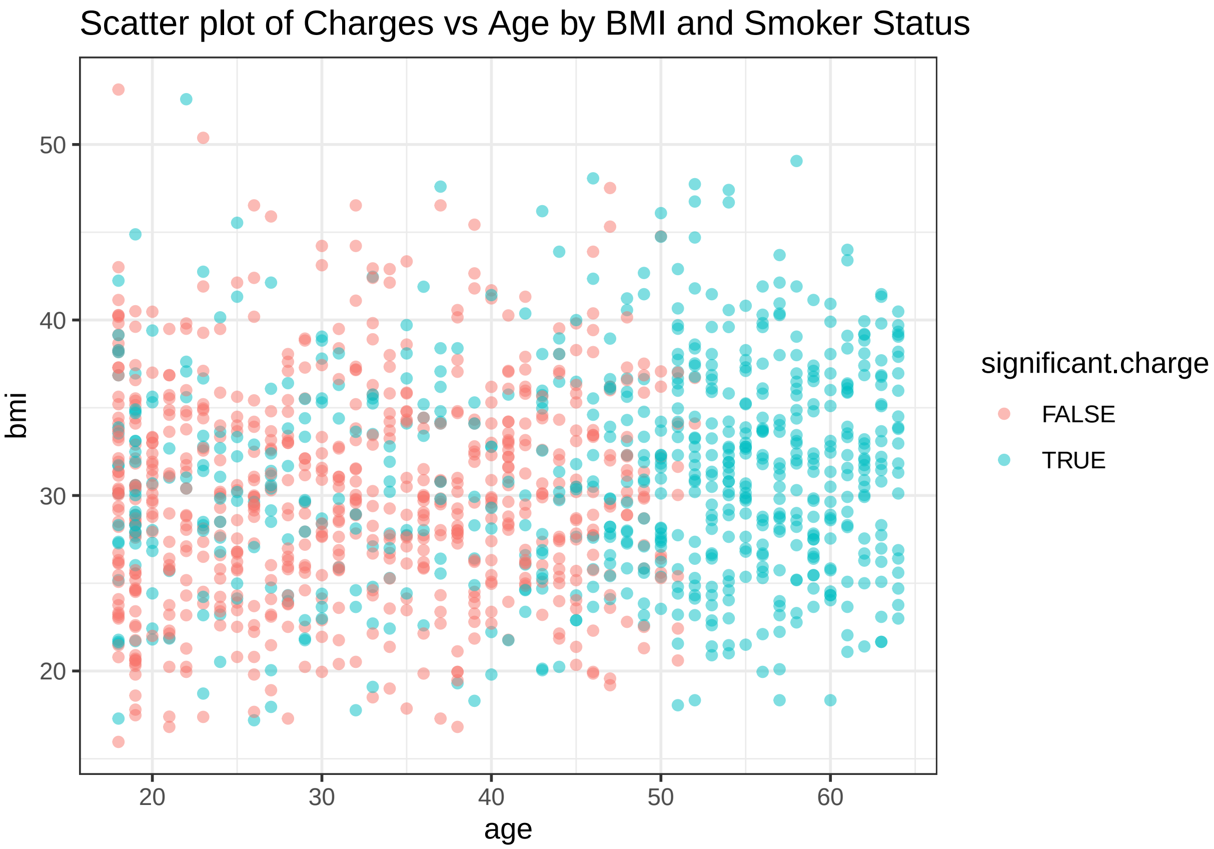
```
ggplot(aes(x=bmi, y=charges, color=smoker), data=data) +
  labs(title="Scatter Plot of Charges vs BMI by Smoker Status") +
  theme_bw() +
  geom_point()
```



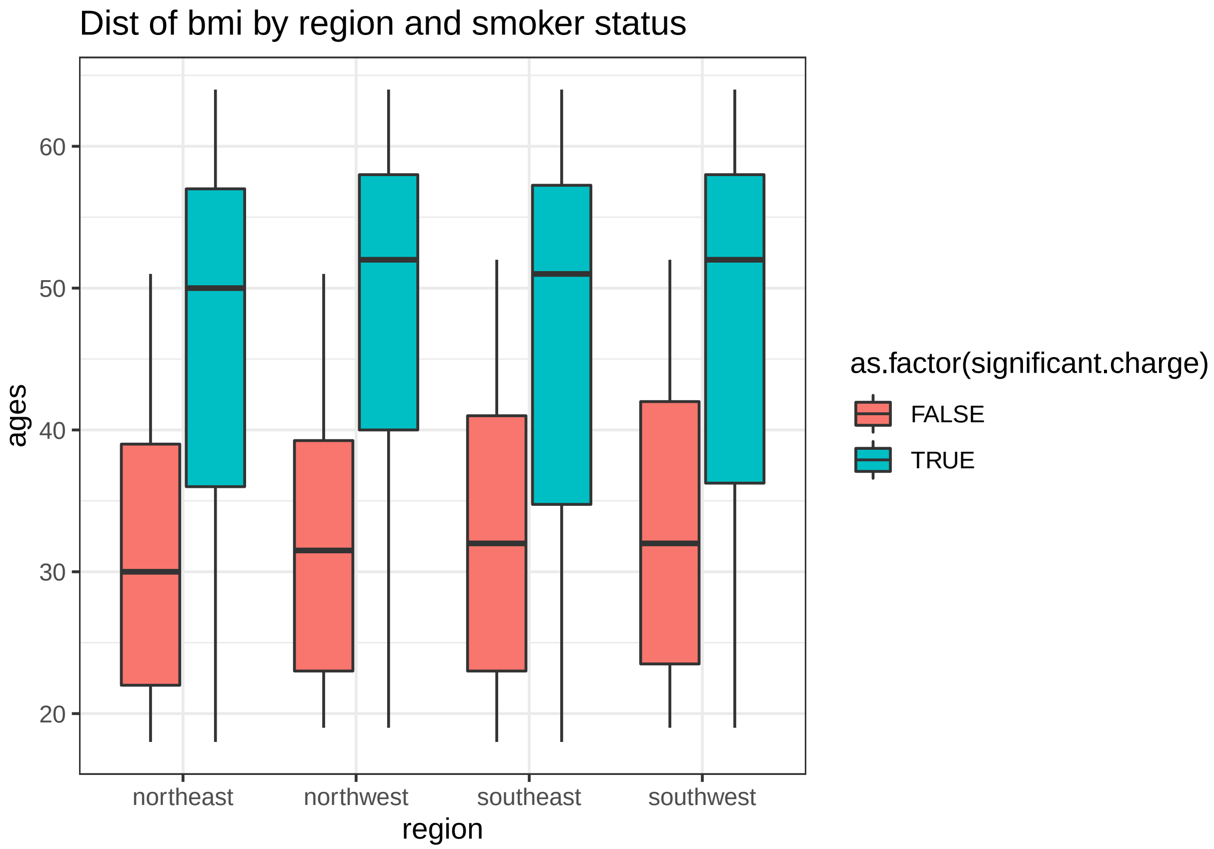
```
ggplot(aes(x=age,y=charges, color=smoker, size=bmi), data=data) +
  labs(title="Scatter plot of Charges vs Age by BMI and Smoker Status") +
  theme_bw() +
  geom_point(alpha=0.5)
```



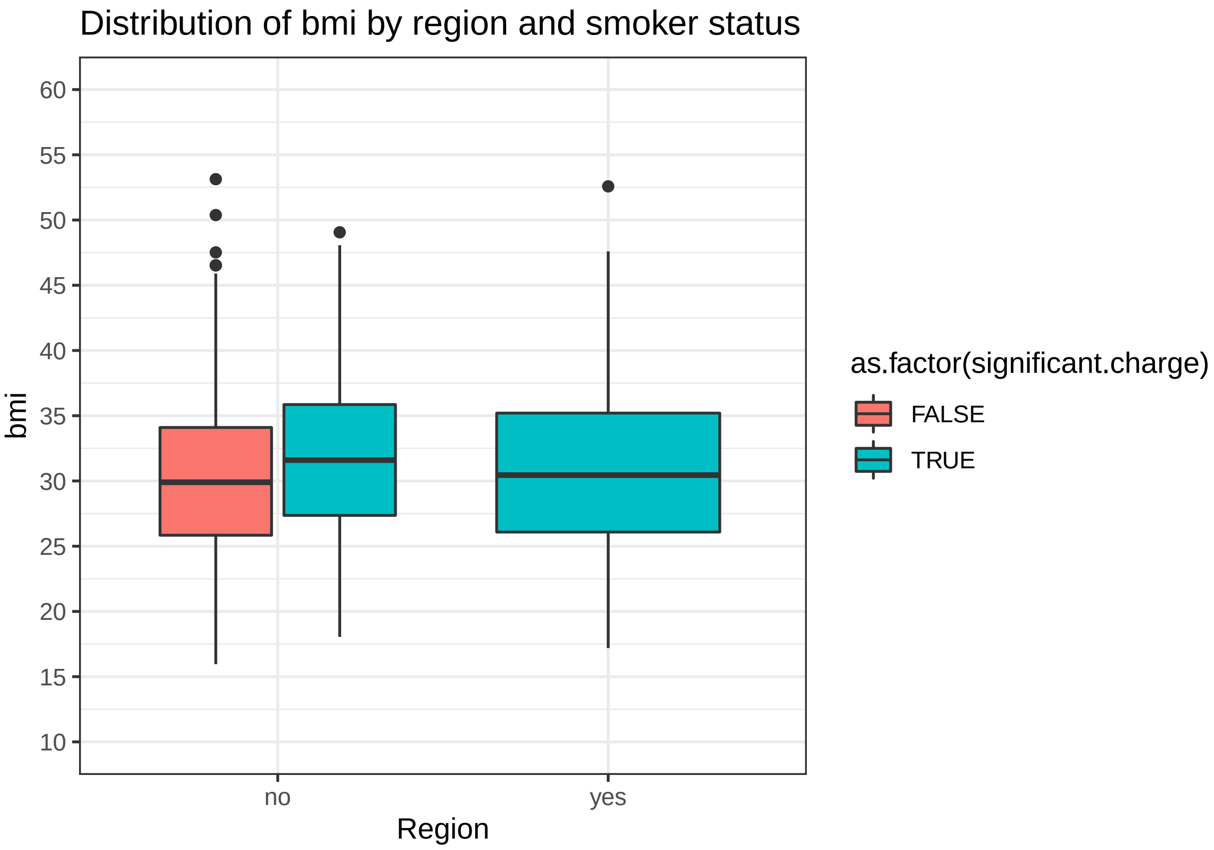
```
ggplot(aes(x=age,y=bmi, color=significant.charge), data=data) +
  labs(title="Scatter plot of Charges vs Age by BMI and Smoker Status") +
  theme_bw() +
  geom_point(alpha=0.5)
```



```
ggplot(data, aes(x=region, y=age, fill=as.factor(significant.charge)))+
  geom_boxplot() +
  theme_bw() +
  labs(x="region", y="ages", title="Dist of bmi by region and smoker status")
```

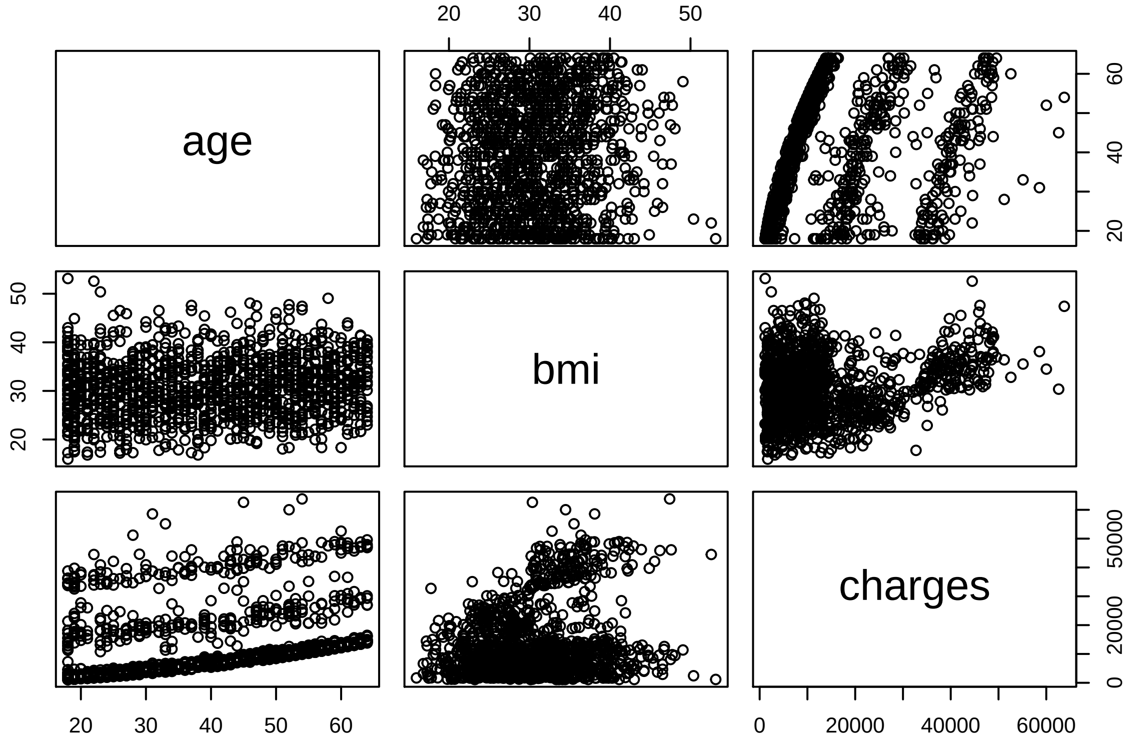


```r
ggplot(data, aes(x=smoker, y=bmi, fill=as.factor(significant.charge)))+
  geom_boxplot() +
  theme_bw() +
  labs(x="Region", y="bmi", title="Distribution of bmi by region and smoker status") +  scale_y_continu
```



Correlation

```
pairs(data[c("age", "bmi", "charges")])
```



```
round(cor(data[c("age", "bmi", "charges")]),4)
```

```
##             age    bmi  charges
## age     1.0000 0.1093  0.2990
## bmi     0.1093 1.0000  0.1983
## charges 0.2990 0.1983  1.0000
```

All possible regressions and pull based on adjusted R square, mallow, and BIC

```
no_class_predictor = data[1:7]
allreg2 <- regsubsets(charges ~., data=no_class_predictor, nbest=2)
summary(allreg2)
```

```
## Subset selection object
## Call: regsubsets.formula(charges ~ ., data = no_class_predictor, nbest = 2)
## 8 Variables  (and intercept)
##                 Forced in Forced out
## age                 FALSE      FALSE
## sexmale             FALSE      FALSE
## bmi                 FALSE      FALSE
## children            FALSE      FALSE
## smokeryes           FALSE      FALSE
## regionnorthwest     FALSE      FALSE
## regionsoutheast     FALSE      FALSE
## regionsouthwest     FALSE      FALSE
## 2 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          age sexmale bmi children smokeryes regionnorthwest regionsoutheast
## 1  ( 1 ) " " " "     " " " "      "*"       " "             " "
## 1  ( 2 ) "*" " "     " " " "      " "       " "             " "
## 2  ( 1 ) "*" " "     " " " "      "*"       " "             " "
## 2  ( 2 ) " " " "     "*" " "      "*"       " "             " "
## 3  ( 1 ) "*" " "     "*" " "      "*"       " "             " "
## 3  ( 2 ) "*" " "     " " "*"      "*"       " "             " "
## 4  ( 1 ) "*" " "     "*" "*"      "*"       " "             " "
## 4  ( 2 ) "*" " "     "*" " "      "*"       " "             "*"
## 5  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 5  ( 2 ) "*" " "     "*" "*"      "*"       " "             " "
## 6  ( 1 ) "*" " "     "*" "*"      "*"       " "             "*"
## 6  ( 2 ) "*" "*"     "*" "*"      "*"       " "             "*"
## 7  ( 1 ) "*" " "     "*" "*"      "*"       "*"             "*"
## 7  ( 2 ) "*" "*"     "*" "*"      "*"       " "             "*"
## 8  ( 1 ) "*" "*"     "*" "*"      "*"       "*"             "*"
##          regionsouthwest
## 1  ( 1 ) " "
## 1  ( 2 ) " "
## 2  ( 1 ) " "
## 2  ( 2 ) " "
## 3  ( 1 ) " "
## 3  ( 2 ) " "
## 4  ( 1 ) " "
## 4  ( 2 ) " "
## 5  ( 1 ) " "
## 5  ( 2 ) "*"
## 6  ( 1 ) "*"
## 6  ( 2 ) " "
## 7  ( 1 ) "*"
## 7  ( 2 ) "*"
## 8  ( 1 ) "*"
```

Best for Adjusted R square

```
coef(allreg2, which.max(summary(allreg2)$adjr2))
```

```
##     (Intercept)             age             bmi        children       smokeryes
##     -12165.3824        257.0064        338.6413        471.5441      23843.8749
## regionsoutheast regionsouthwest
##       -858.4696       -782.7452
```

Best for Mallows

```
coef(allreg2, which.min(summary(allreg2)$cp))
```

```
##     (Intercept)             age             bmi        children       smokeryes
##     -12165.3824        257.0064        338.6413        471.5441      23843.8749
## regionsoutheast regionsouthwest
##       -858.4696       -782.7452
```

Best for BIC

```
coef(allreg2, which.min(summary(allreg2)$bic))
```

```
## (Intercept)         age         bmi    children   smokeryes
## -12102.7694    257.8495    321.8514    473.5023  23811.3998
```

Forward Selection

```
##intercept only model
regnull <- lm(charges~1, data=no_class_predictor)
##model with all predictors
regfull <- lm(charges ~ . , data=no_class_predictor)
```

Forward Selection

```
step(regnull, scope=list(lower=regnull, upper=regfull), direction="forward")
```

```
## Start:  AIC=25160.18
## charges ~ 1
##
##            Df  Sum of Sq        RSS   AIC
## + smoker    1 1.2152e+11 7.4554e+10 23868
## + age       1 1.7530e+10 1.7854e+11 25037
## + bmi       1 7.7134e+09 1.8836e+11 25108
## + children  1 9.0660e+08 1.9517e+11 25156
## + region    3 1.3008e+09 1.9477e+11 25157
## + sex       1 6.4359e+08 1.9543e+11 25158
## <none>                   1.9607e+11 25160
##
## Step:  AIC=23868.38
## charges ~ smoker
##
##            Df  Sum of Sq        RSS   AIC
## + age       1 1.9928e+10 5.4626e+10 23454
## + bmi       1 7.4856e+09 6.7069e+10 23729
## + children  1 7.5272e+08 7.3802e+10 23857
## <none>                   7.4554e+10 23868
## + sex       1 1.4213e+06 7.4553e+10 23870
## + region    3 1.0752e+08 7.4447e+10 23872
##
## Step:  AIC=23454.24
## charges ~ smoker + age
##
##            Df  Sum of Sq        RSS   AIC
## + bmi       1 5112896646 4.9513e+10 23325
## + children  1  459283727 5.4167e+10 23445
## <none>                   5.4626e+10 23454
## + sex       1    2225509 5.4624e+10 23456
## + region    3  138426748 5.4488e+10 23457
##
## Step:  AIC=23324.76
## charges ~ smoker + age + bmi
##
##            Df Sum of Sq        RSS   AIC
## + children  1 434769398 4.9078e+10 23315
## + region    3 232012208 4.9281e+10 23324
## <none>                  4.9513e+10 23325
## + sex       1   3942912 4.9509e+10 23327
##
## Step:  AIC=23314.96
## charges ~ smoker + age + bmi + children
```

```
##
##          Df Sum of Sq        RSS   AIC
## + region  3 233200844 4.8845e+10 23315
## <none>                4.9078e+10 23315
## + sex     1   5486063 4.9073e+10 23317
##
## Step:  AIC=23314.58
## charges ~ smoker + age + bmi + children + region
##
##        Df Sum of Sq        RSS   AIC
## <none>              4.8845e+10 23315
## + sex   1   5716429 4.8840e+10 23316
```

```
##
## Call:
## lm(formula = charges ~ smoker + age + bmi + children + region,
##     data = no_class_predictor)
##
## Coefficients:
##     (Intercept)        smokeryes              age              bmi
##        -11990.3          23836.3            257.0            338.7
##        children  regionnorthwest  regionsoutheast  regionsouthwest
##           474.6           -352.2          -1034.4           -959.4
```

Backwards

```
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
```

```
## Start:  AIC=23316.43
## charges ~ age + sex + bmi + children + smoker + region
##
##            Df  Sum of Sq        RSS   AIC
## - sex       1 5.7164e+06 4.8845e+10 23315
## <none>                   4.8840e+10 23316
## - region    3 2.3343e+08 4.9073e+10 23317
## - children  1 4.3755e+08 4.9277e+10 23326
## - bmi       1 5.1692e+09 5.4009e+10 23449
## - age       1 1.7124e+10 6.5964e+10 23717
## - smoker    1 1.2245e+11 1.7129e+11 24993
##
## Step:  AIC=23314.58
## charges ~ age + bmi + children + smoker + region
##
##            Df  Sum of Sq        RSS   AIC
## <none>                   4.8845e+10 23315
## - region    3 2.3320e+08 4.9078e+10 23315
## - children  1 4.3596e+08 4.9281e+10 23324
## - bmi       1 5.1645e+09 5.4010e+10 23447
## - age       1 1.7151e+10 6.5996e+10 23715
## - smoker    1 1.2301e+11 1.7186e+11 24996
```

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

```
##
## Coefficients:
##     (Intercept)              age              bmi         children
##        -11990.3            257.0            338.7            474.6
##       smokeryes  regionnorthwest  regionsoutheast  regionsouthwest
##         23836.3           -352.2          -1034.4           -959.4
```

Based on forward and backward

We get the same model for forward and backward

Let’s first make a multiple linear regression model with all the predictors.

```
mlr_full = lm(charges ~  age + bmi + children + smoker + region, data=no_class_predictor)
summary(mlr_full)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = no_class_predictor)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11367.2  -2835.4   -979.7   1361.9  29935.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -11990.27     978.76 -12.250  < 2e-16 ***
## age                256.97      11.89  21.610  < 2e-16 ***
## bmi                338.66      28.56  11.858  < 2e-16 ***
## children           474.57     137.74   3.445 0.000588 ***
## smokeryes        23836.30     411.86  57.875  < 2e-16 ***
## regionnorthwest   -352.18     476.12  -0.740 0.459618
## regionsoutheast  -1034.36     478.54  -2.162 0.030834 *
## regionsouthwest   -959.37     477.78  -2.008 0.044846 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6060 on 1330 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7496
## F-statistic: 572.7 on 7 and 1330 DF,  p-value: < 2.2e-16
```

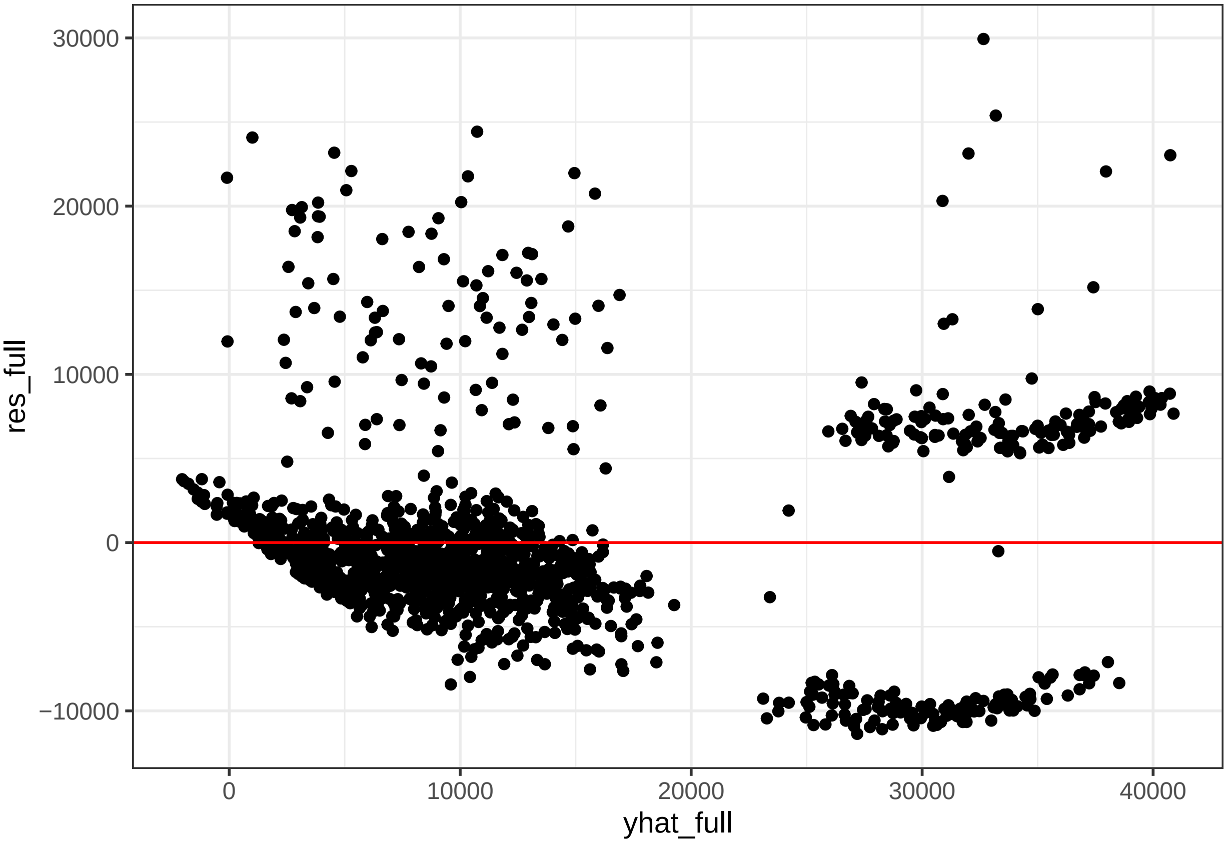
The full regression is as follows.

$$\hat{y} = -11938.5 + 256.9\text{age} - 131.3I_1 + 339.2\text{bmi} + 475.5\text{children} + 23848.5I_2 - 353.0I_3 - 1035.0I_4 - 960.0I_5$$

I_1 indicates whether the sex of the client is male. The value will be 0 for females. I_2 indicates whether that a client smokes. The value will be 0 for non smokers. I_3 indicates that the client is in the northwest region. I_4 indicates that the client is located in the southeast. I_5 indicates that the client is located inthe southwest. If the client is in the northeast I_3, I_4, I_5 will be zero, since this is the reference class.

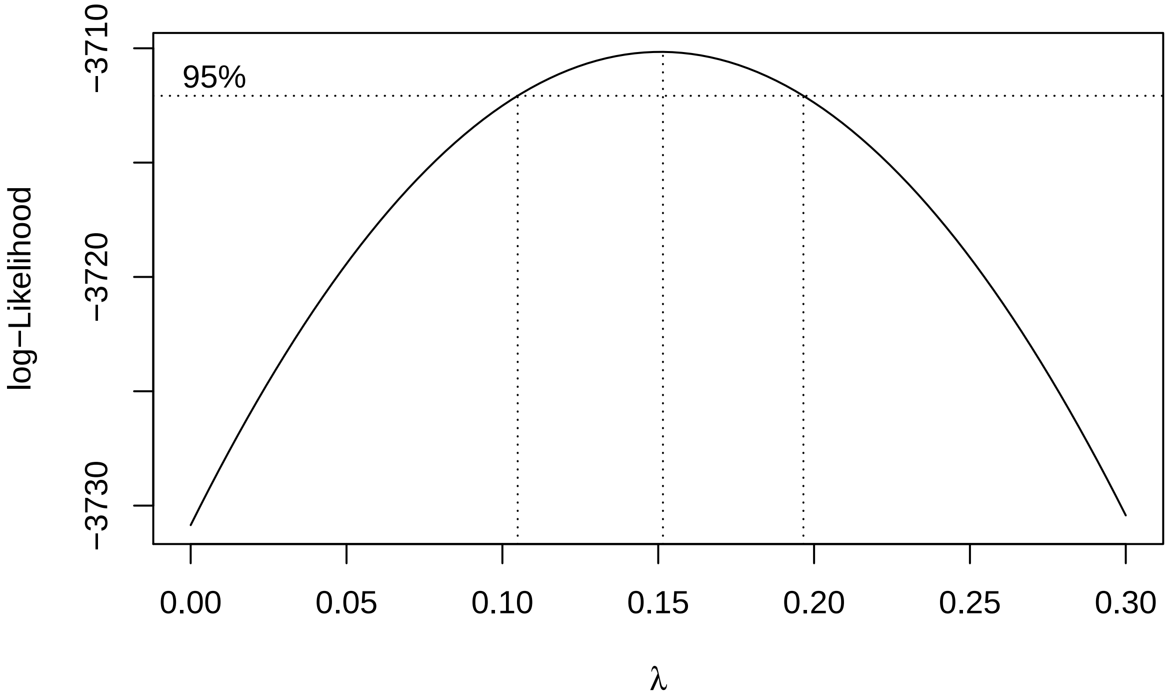
Assumption Check of Full Model

```
yhat_full <- mlr_full$fitted.values
res_full <- mlr_full$residuals
data %>%
  ggplot(aes(yhat_full, res_full)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



The residuals are obviously not evenly scattered, which then we can utilize the boxcox method to give us information about transformation.

```
boxcox(mlr_full, lambda=seq(0,0.3, 0.01))
```



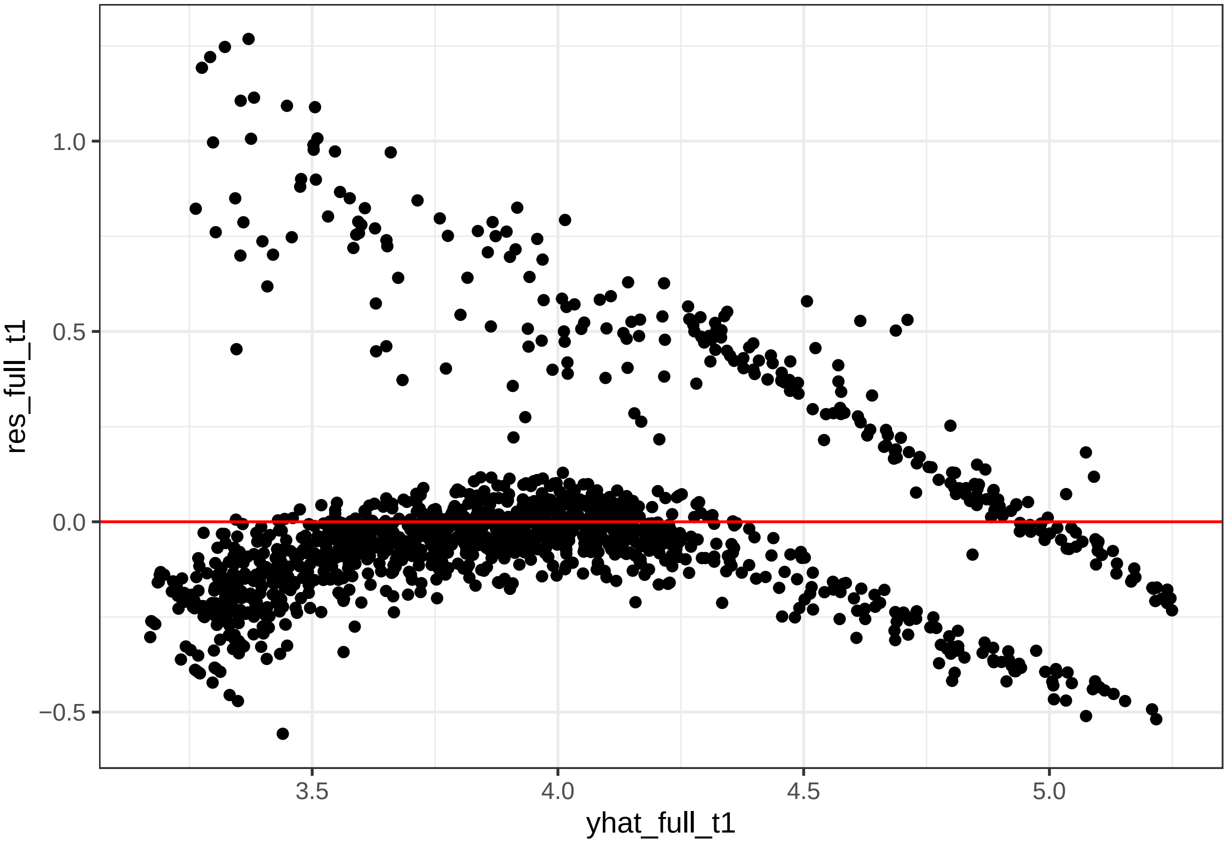
From the boxcox we can try a lambda value of 0.15 for transformation.

```
first_transformation_full <- data
first_transformation_full$charges <- first_transformation_full$charges^0.15
mlr_transform_first <- lm(charges ~ . - significant.charge, data=first_transformation_full)
summary(mlr_transform_first)
```

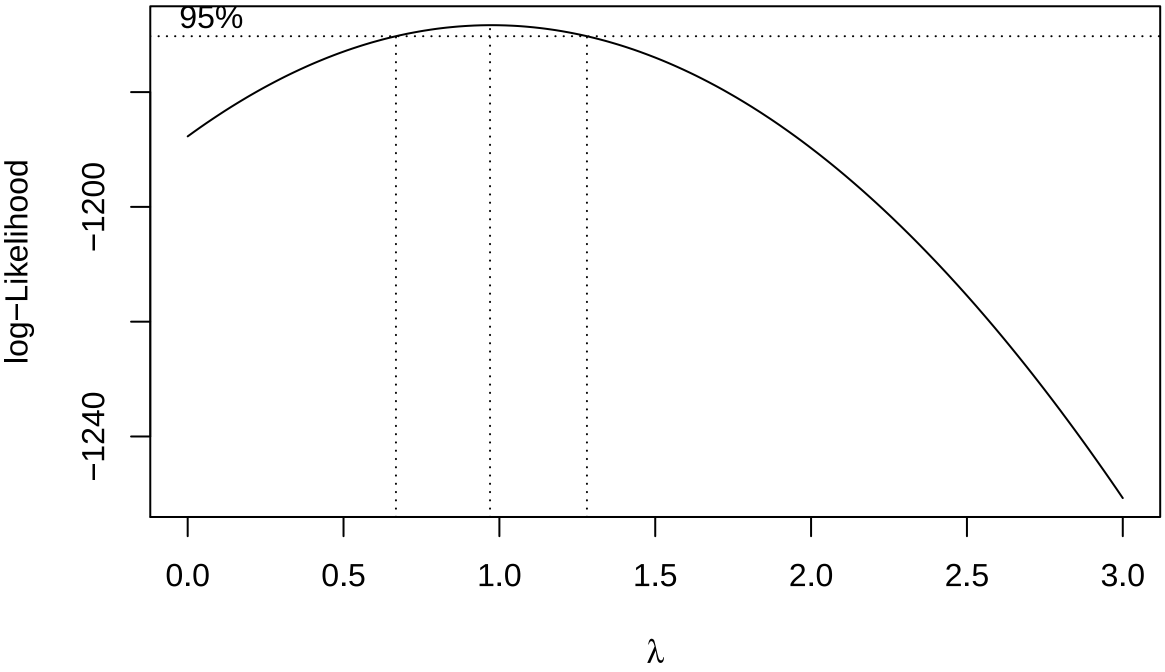
```
##
## Call:
## lm(formula = charges ~ . - significant.charge, data = first_transformation_full)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55700 -0.12467 -0.03934  0.02881  1.26849
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.7385737  0.0419031  65.355  < 2e-16 ***
## age              0.0191413  0.0005047  37.923  < 2e-16 ***
## sexmale         -0.0370677  0.0141235  -2.625  0.00878 **
## bmi              0.0090116  0.0012132   7.428 1.96e-13 ***
## children         0.0527358  0.0058456   9.021  < 2e-16 ***
## smokeryes        0.9595356  0.0175259  54.750  < 2e-16 ***
## regionnorthwest -0.0347484  0.0202035  -1.720  0.08568 .
## regionsoutheast -0.0847137  0.0203060  -4.172 3.22e-05 ***
## regionsouthwest -0.0710849  0.0202738  -3.506  0.00047 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2572 on 1329 degrees of freedom
## Multiple R-squared:  0.7766, Adjusted R-squared:  0.7752
## F-statistic: 577.3 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Residual Plot of the transformed model.

```
yhat_full_t1 <- mlr_transform_first$fitted.values
res_full_t1 <- mlr_transform_first$residuals
data %>%
  ggplot(aes(yhat_full_t1, res_full_t1)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



```
boxcox(mlr_transform_first, lambda=seq(0,3, 0.01))
```


95%
log−Likelihood
−1200
−1240
0.0
0.5
1.0
1.5
2.0
2.5
3.0
λ