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 ---------------------------------------------------------------------------
```

```
## 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 ------------------------------------------------------------------------------------
## 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)
```

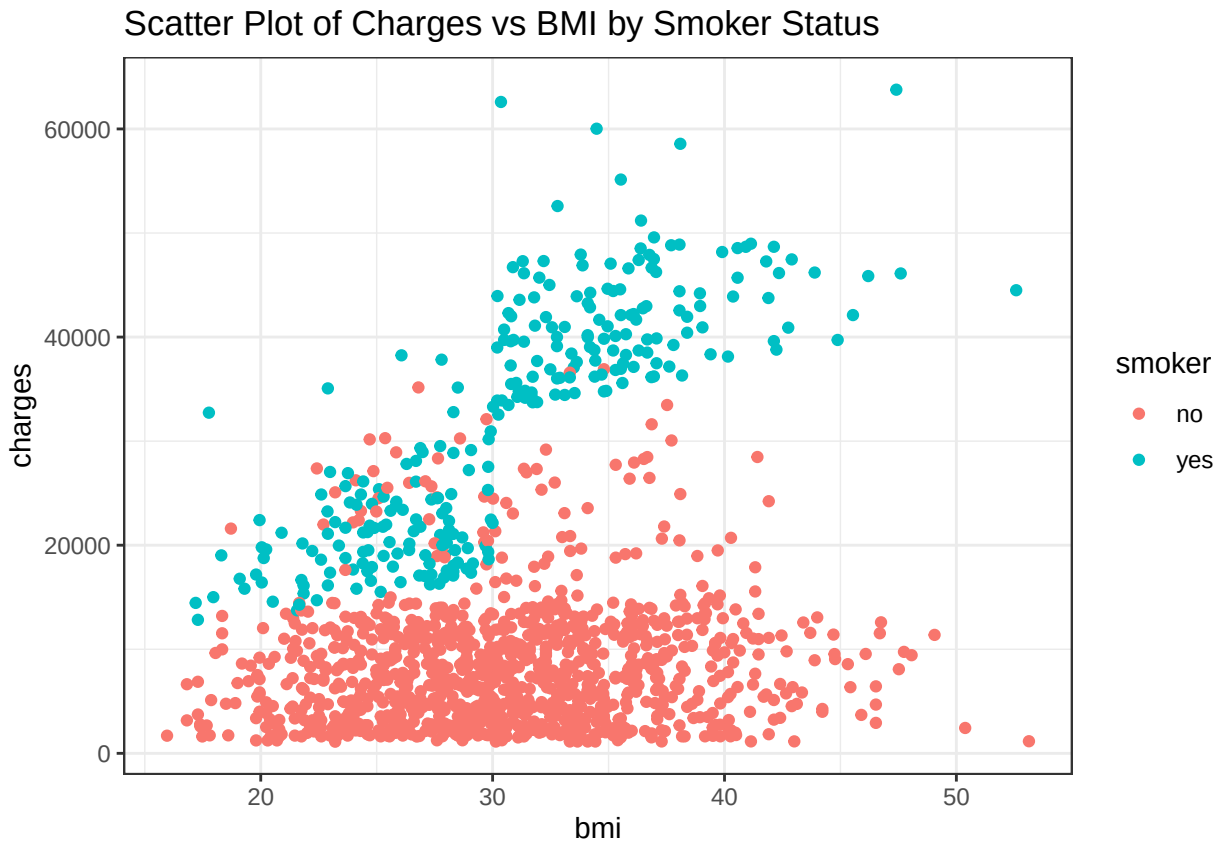
```
##   age    sex    bmi children smoker    region   charges
## 1  19 female 27.900        0    yes southwest 16884.924
```

```
## 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
```

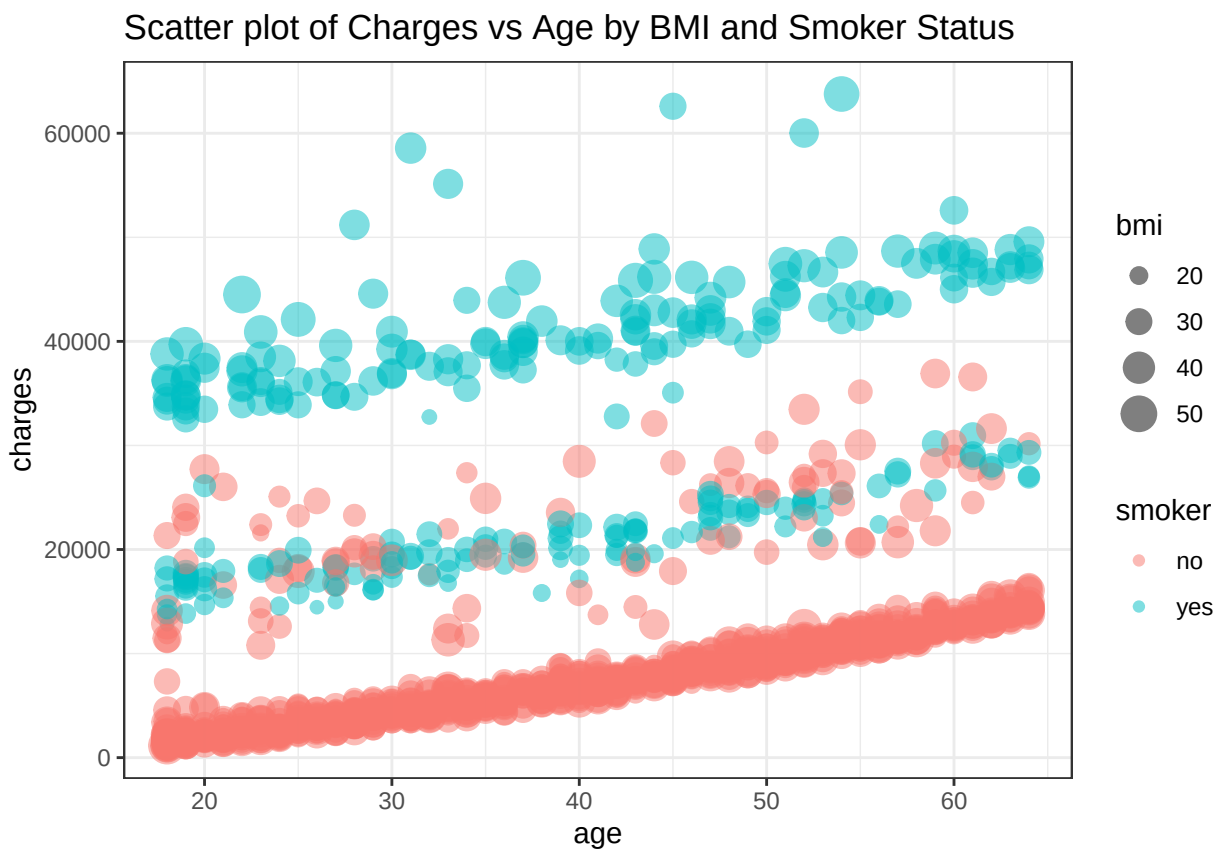
```
data$significant.charge = as.factor(data$charges > median(data$charges))
data$smoker = as.factor(data$smoker)
data$region = as.factor(data$region)
head(data)
```

```
##   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
```

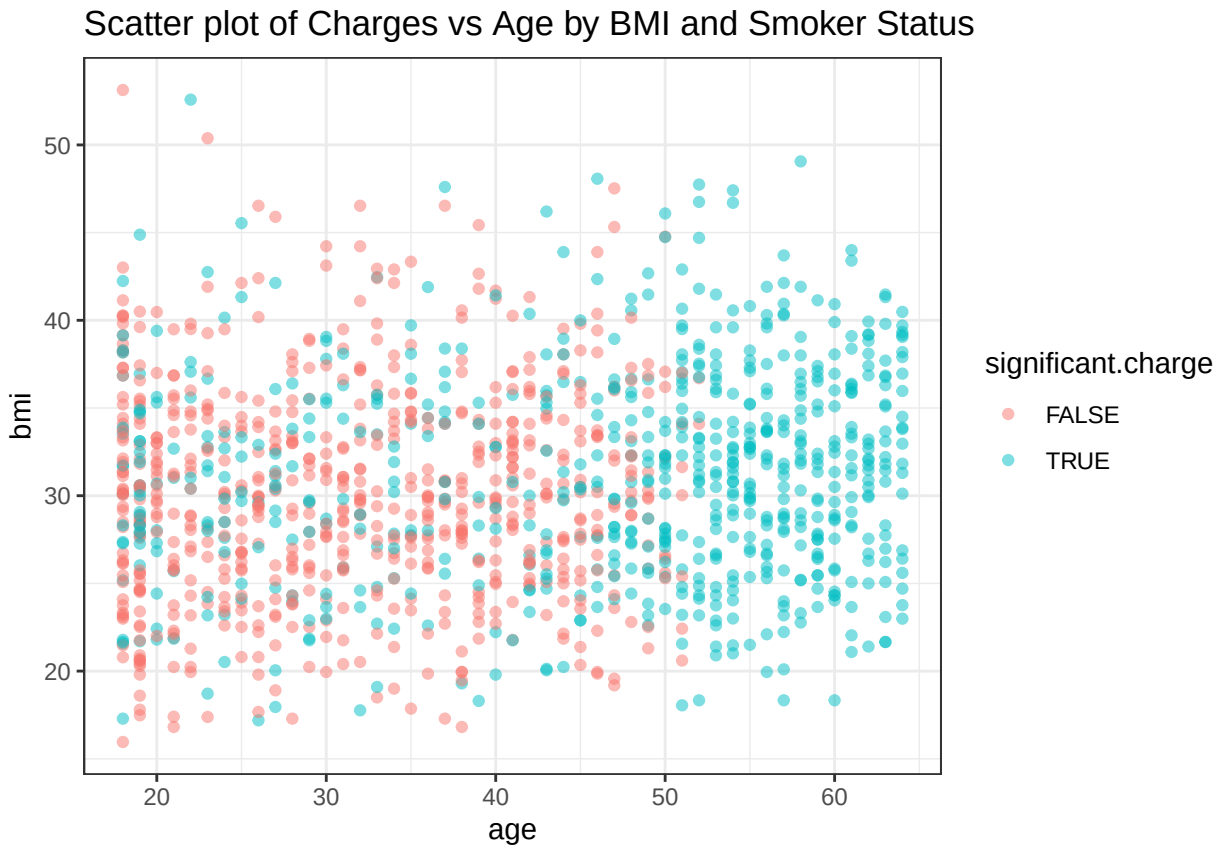
```
ggplot(aes(x=bmi, y=charges, color=smoker), data=data) +
  labs(title="Scatter Plot of Charges vs BMI by Smoker Status") +
  theme_bw() +
  geom_point()
```



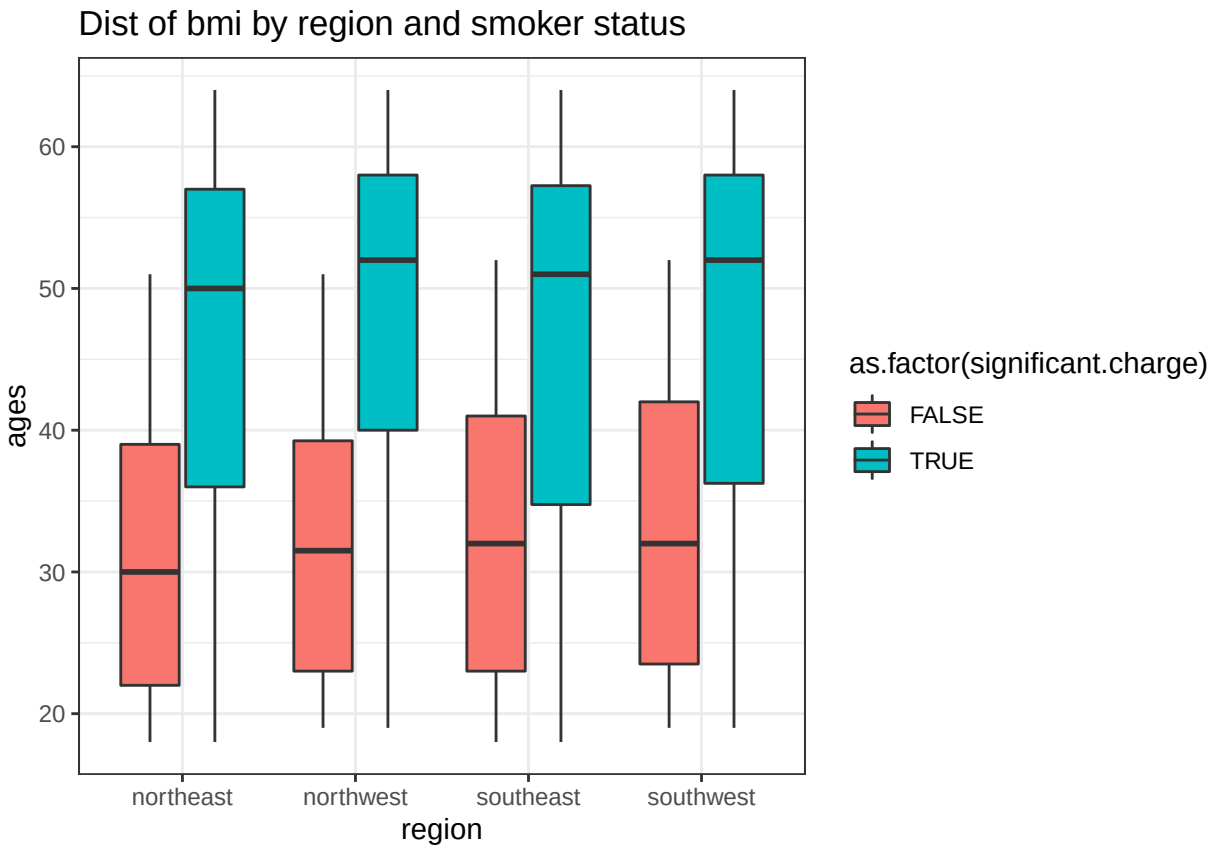

```
ggplot(aes(x=age,y=charges, color=smoker, size=bmi), data=data) +
  labs(title="Scatter plot of Charges vs Age by BMI and Smoker Status") +
  theme_bw() +
  geom_point(alpha=0.5)
```



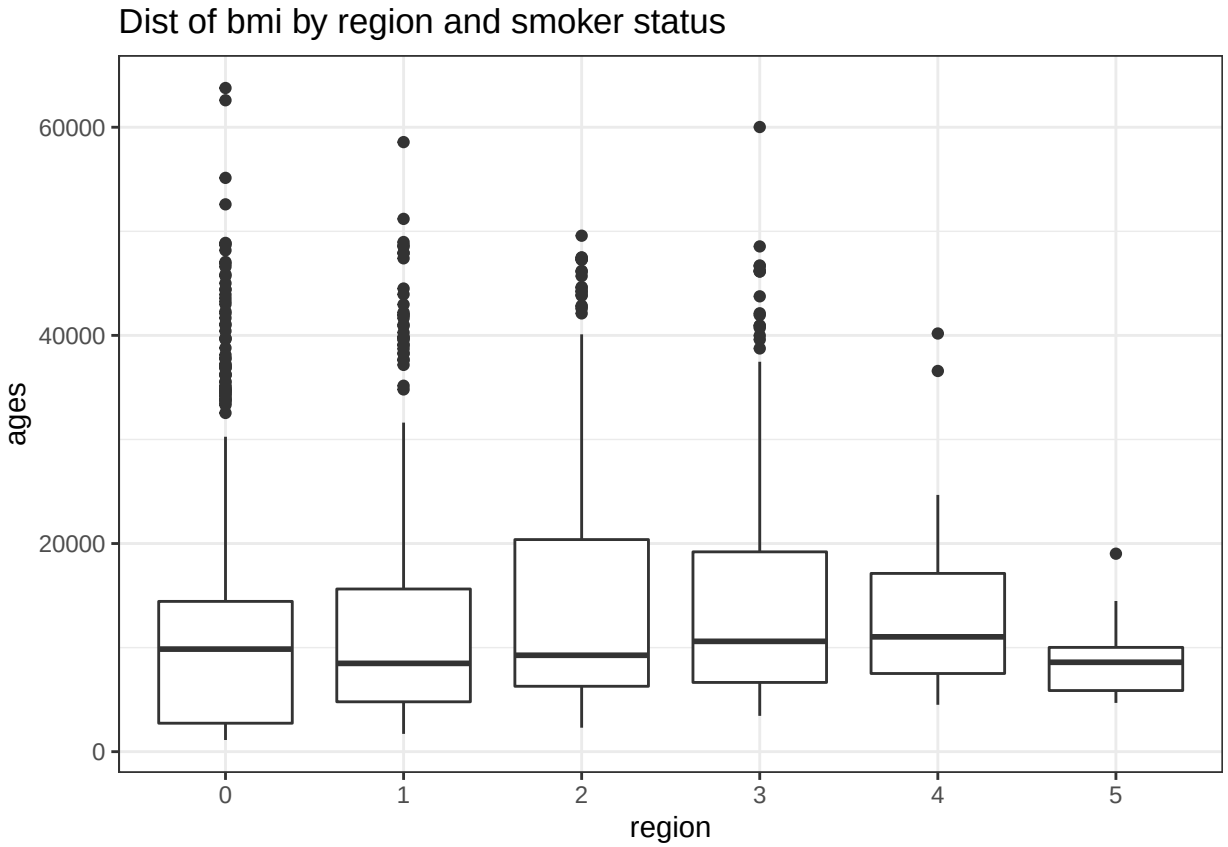

```
ggplot(aes(x=age,y=bmi, color=significant.charge), data=data) +
  labs(title="Scatter plot of Charges vs Age by BMI and Smoker Status") +
  theme_bw() +
  geom_point(alpha=0.5)
```



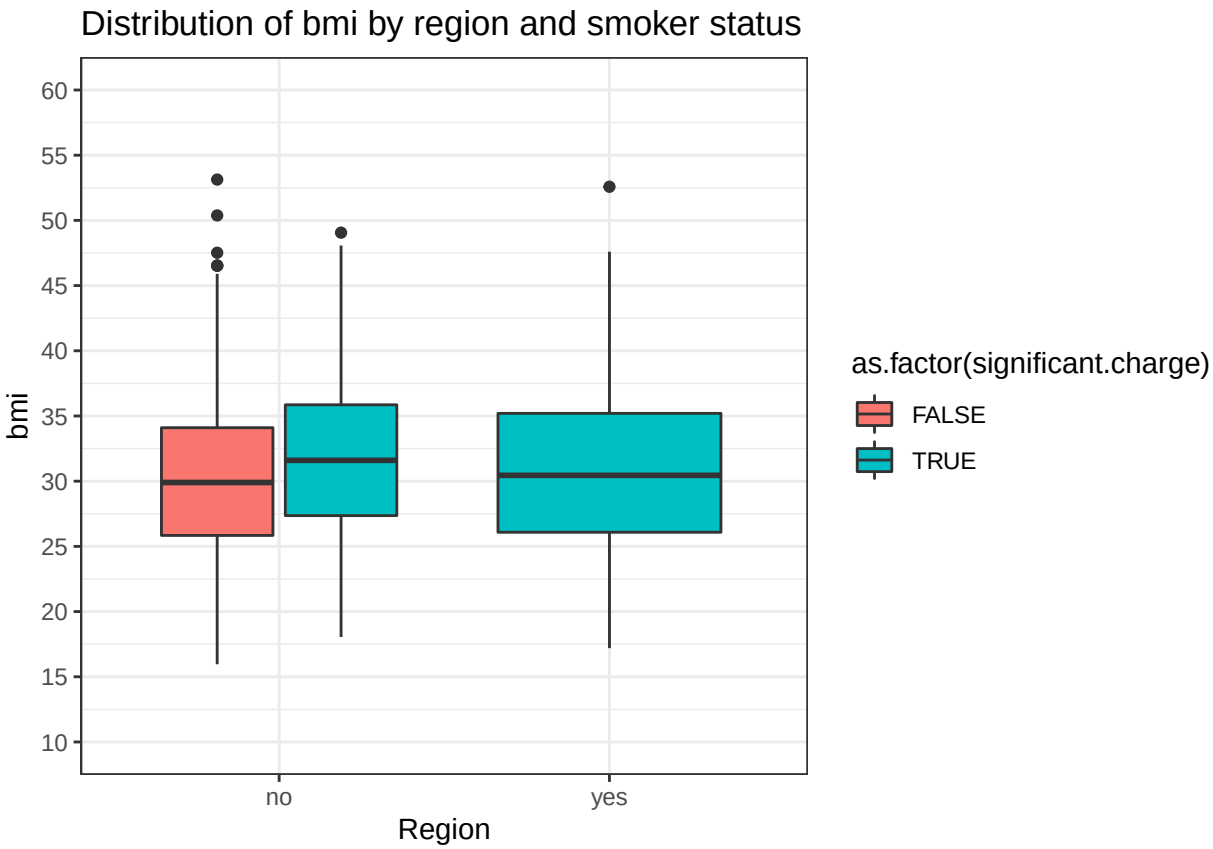

```
ggplot(data, aes(x=region, y=age, fill=as.factor(significant.charge)))+
  geom_boxplot() +
  theme_bw() +
  labs(x="region", y="ages", title="Dist of bmi by region and smoker status")
```




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

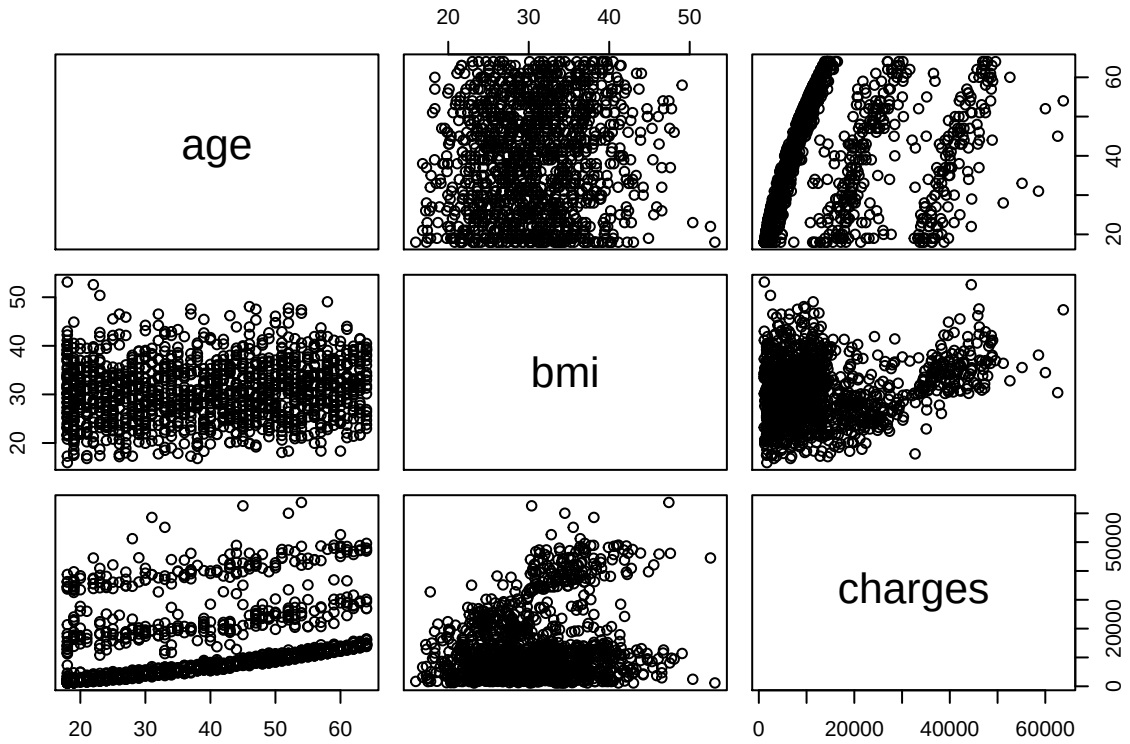



```
ggplot(data, aes(x=smoker, y=bmi, fill=as.factor(significant.charge)))+
  geom_boxplot() +
  theme_bw() +
  labs(x="Region", y="bmi", title="Distribution of bmi by region and smoker status") +  scale_y_continu
```



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
```

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

regionsouthwest -782.7452

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=data)
summary(mlr_full)
```

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

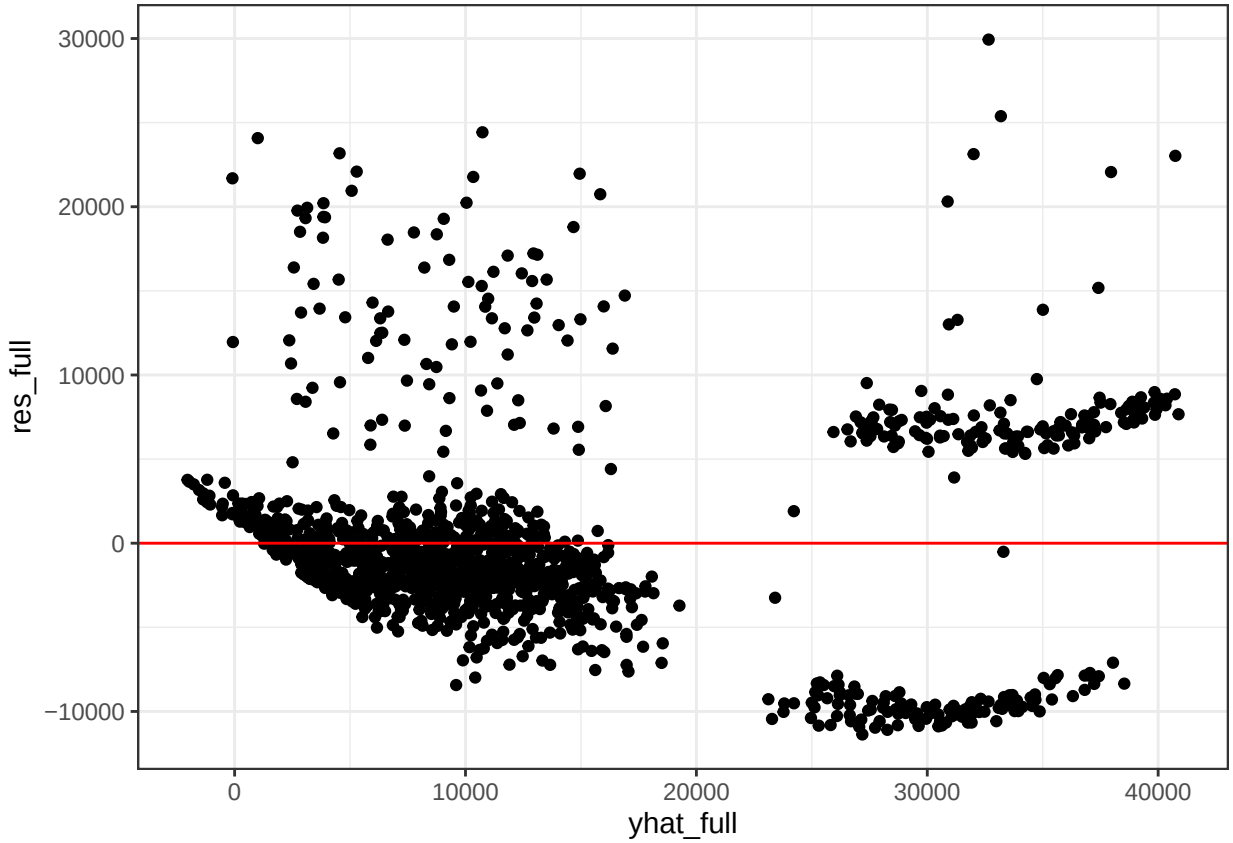
The full regression is as follows.

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

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.

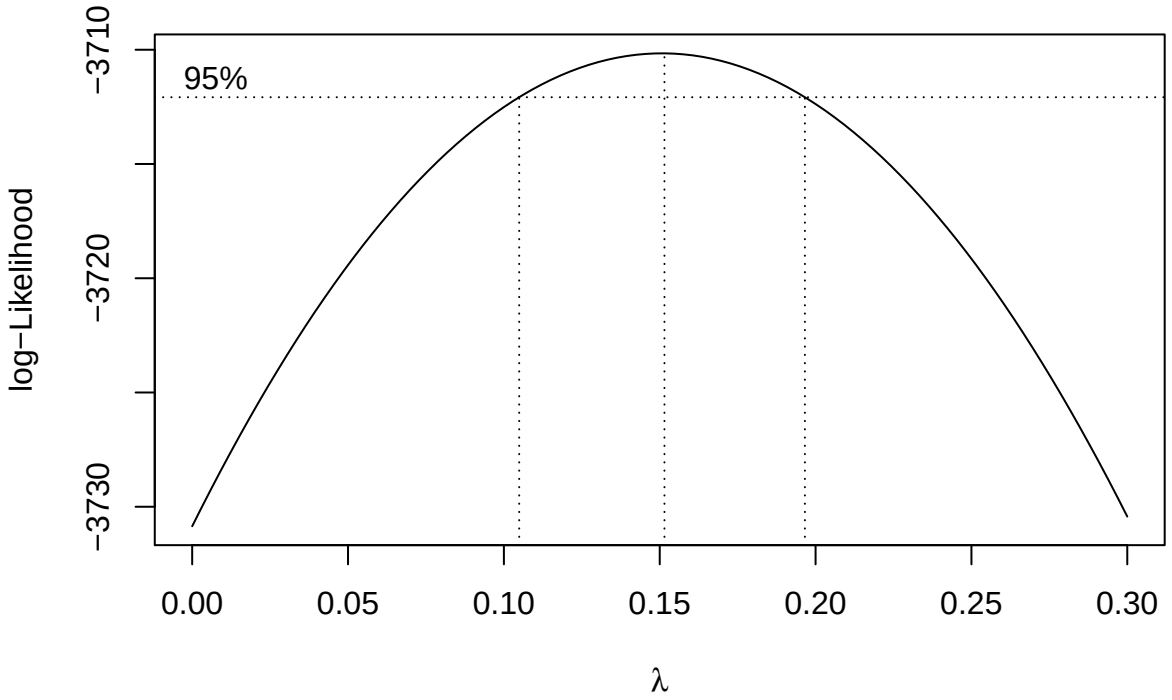
Assumption Check of Full Model

```
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.

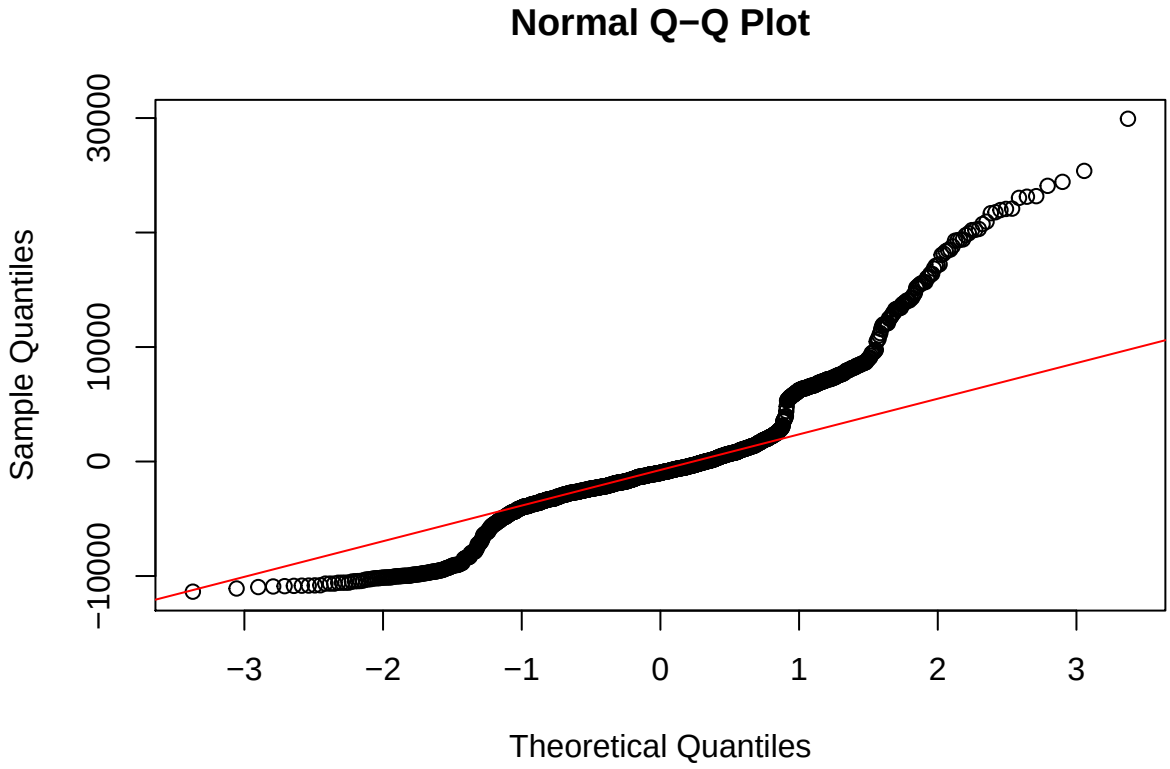
```
boxcox(mlr_full, lambda=seq(0,0.3, 0.01))
```



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

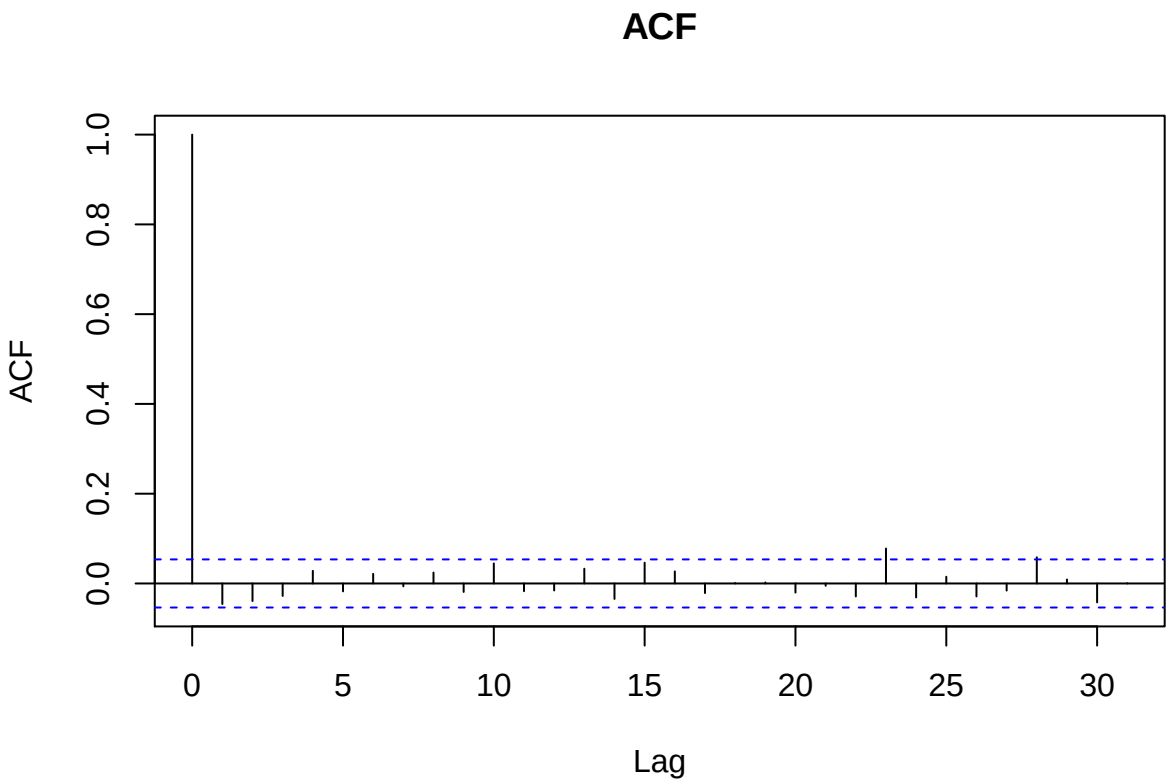
QQPlot

```
{
  qqnorm(mlr_full$residuals)
  qqline(mlr_full$residuals, col="red")
}
```



ACF

```
acf(mlr_full$residuals, main="ACF")
```



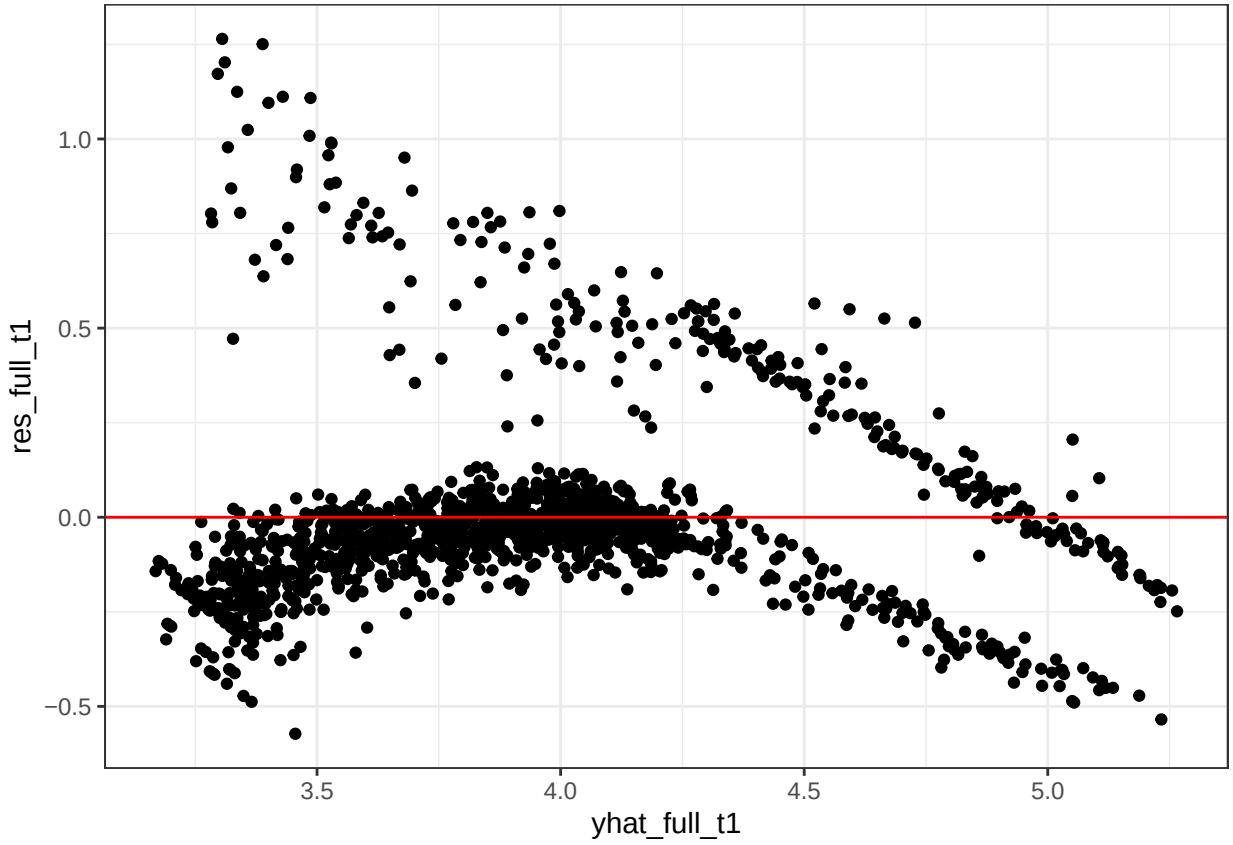

```
first_transformation_full <- data
first_transformation_full$charges <- first_transformation_full$charges^0.15
mlr_transform_first <- lm(charges ~ age + bmi + children + smoker + region, data=first_transformation_fu
summary(mlr_transform_first)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker + region,
##     data = first_transformation_full)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57232 -0.12513 -0.04165  0.03000  1.26454
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.7239709  0.0416239  65.443  < 2e-16 ***
## age              0.0191744  0.0005057  37.916  < 2e-16 ***
## bmi              0.0088624  0.0012145   7.297 5.04e-13 ***
## children         0.0524721  0.0058577   8.958  < 2e-16 ***
## smokeryes        0.9560821  0.0175151  54.586  < 2e-16 ***
## regionnorthwest -0.0345277  0.0202480  -1.705   0.0884 .
## regionsoutheast -0.0845268  0.0203508  -4.153 3.48e-05 ***
## regionsouthwest -0.0708940  0.0203185  -3.489   0.0005 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2577 on 1330 degrees of freedom
## Multiple R-squared:  0.7754, Adjusted R-squared:  0.7742
## F-statistic: 655.9 on 7 and 1330 DF,  p-value: < 2.2e-16
```

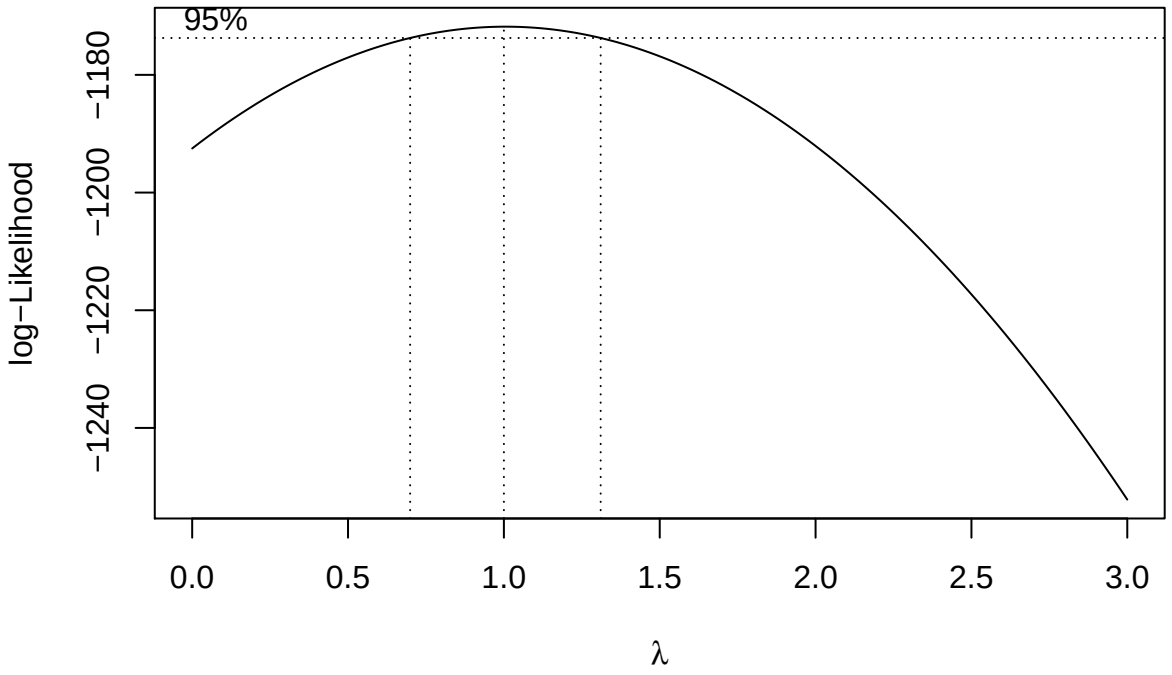
Residual Plot of the transformed model.

```
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")
```



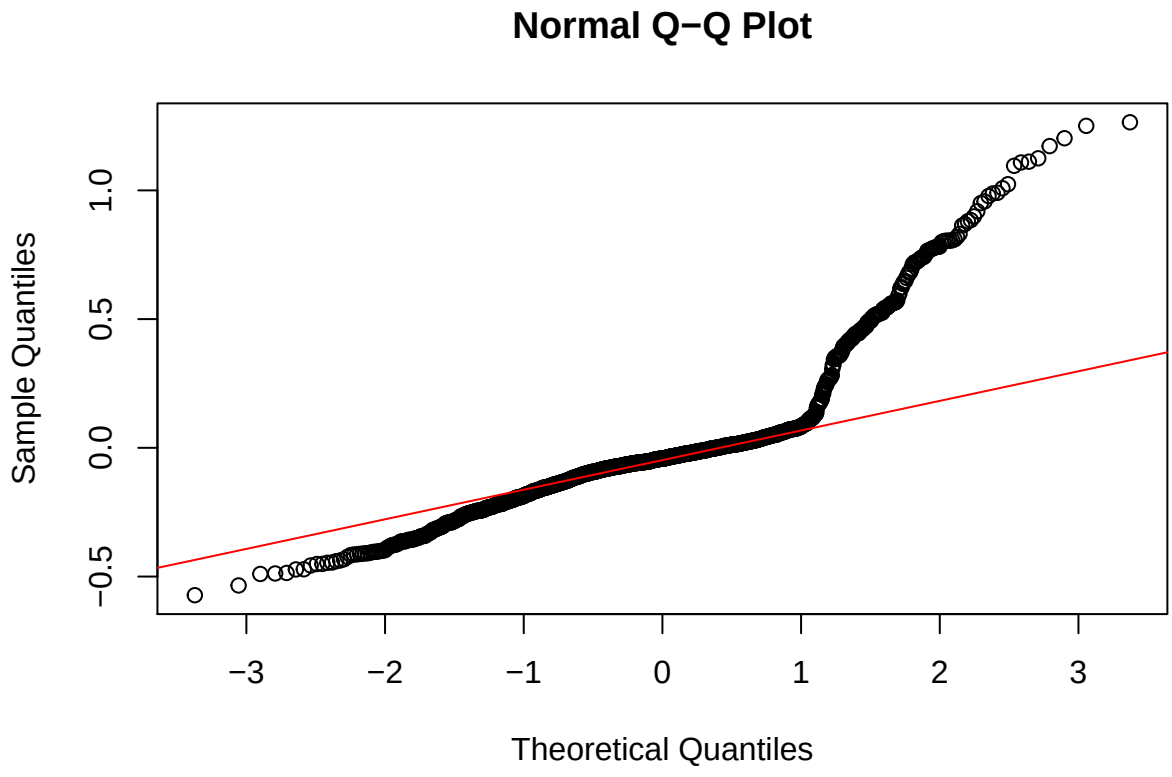
Violation in constant variance

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



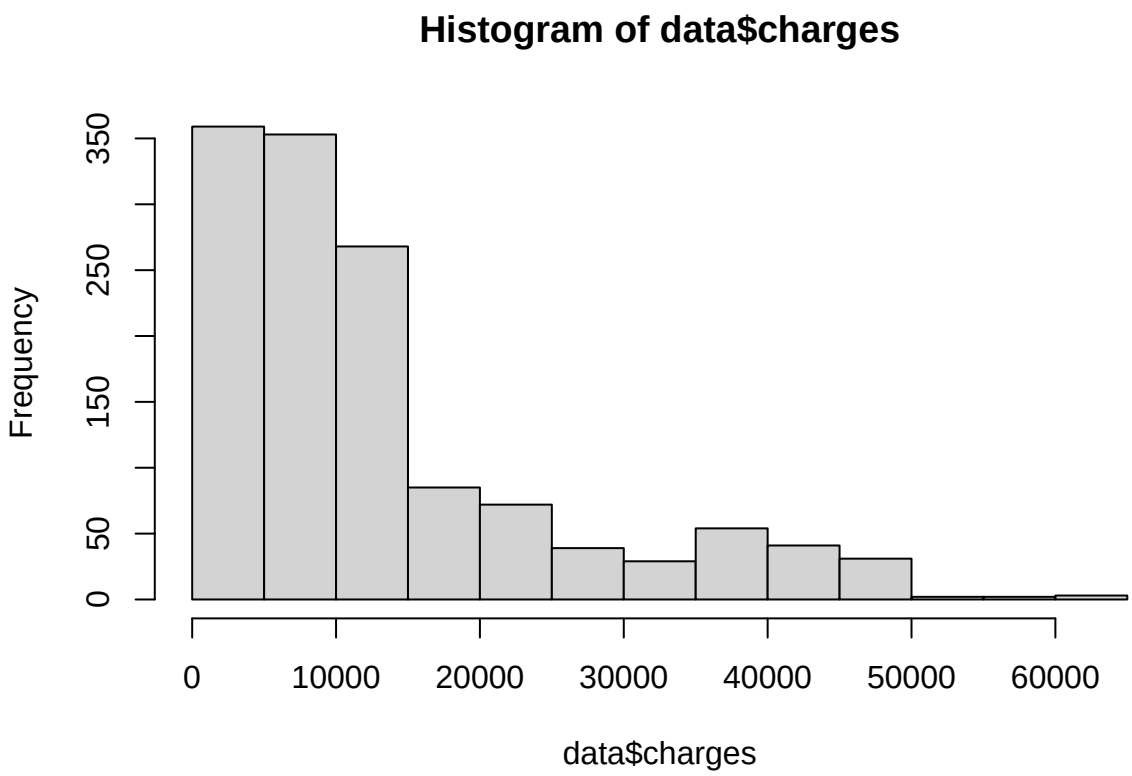
QQPLOT

```
{
  qqnorm(mlr_transform_first$residuals)
  qqline(mlr_transform_first$residuals, col="red")
}
```

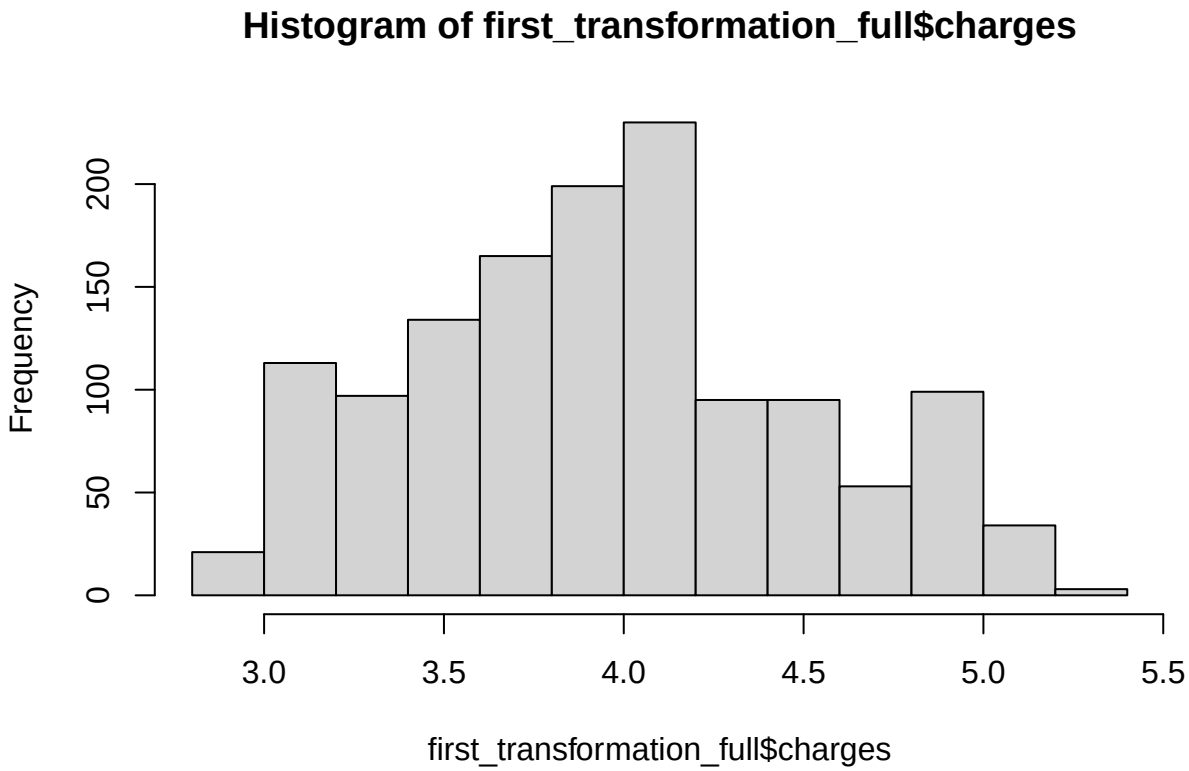


Why is this happening? Is there some weird behavior in the response variable?

```
hist(data$charges)
```




```
hist(first_transformation_full$charges)
```



Trial of other predictors to fullfill the linearity assumption.

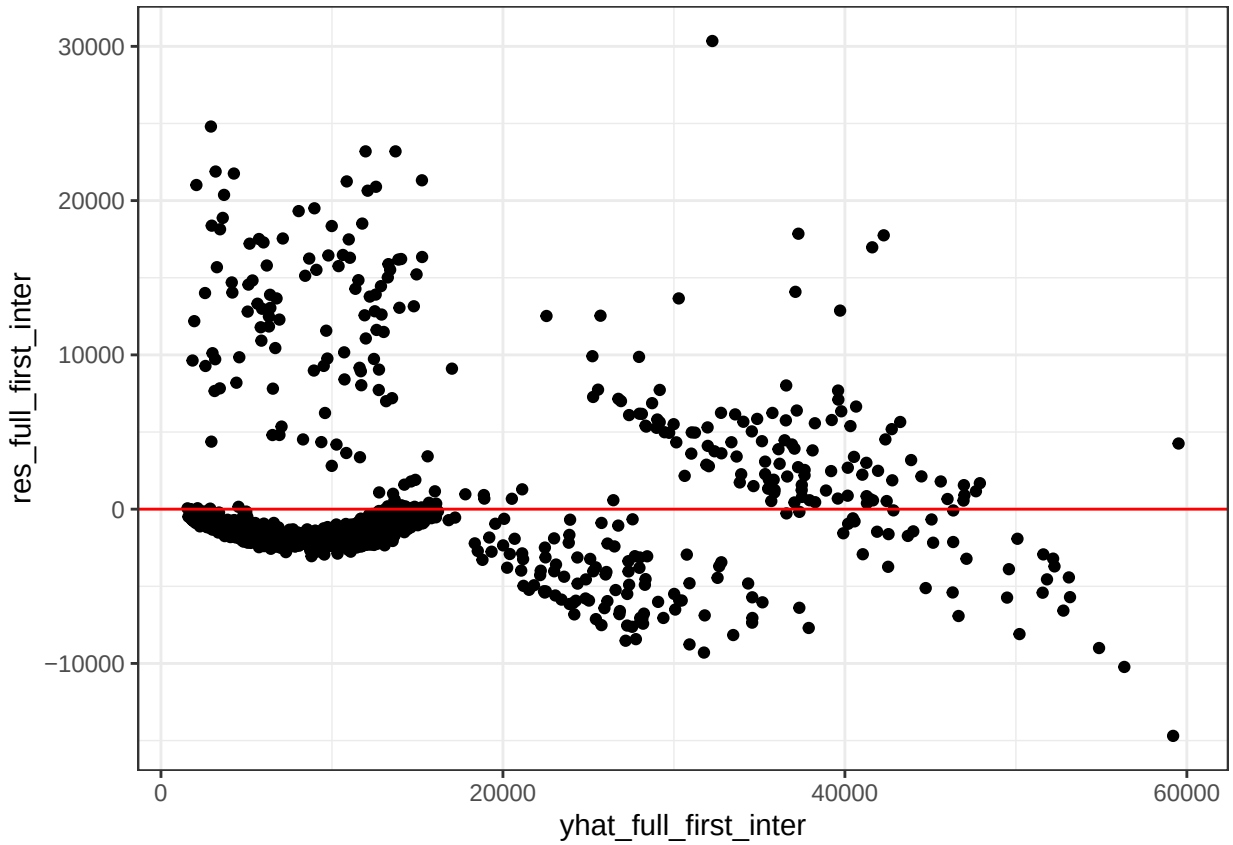
Maybe we can add some interaction terms to the model to see if we can fix the linearity assumption.

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

```
##
## Call:
## lm(formula = charges ~ age * smoker + bmi * smoker + children +
##     region, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14695.2  -1918.6  -1316.2   -480.3  30345.6
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      -2469.101    870.166  -2.838  0.00462 **
## age                264.558     10.672  24.791  < 2e-16 ***
## smokeryes        -20223.654   1831.889 -11.040  < 2e-16 ***
## bmi                 22.444     25.679   0.874  0.38228
## children           512.956    110.331   4.649 3.66e-06 ***
## regionnorthwest   -581.232    381.383  -1.524  0.12774
## regionsoutheast  -1205.652    383.462  -3.144  0.00170 **
## regionsouthwest  -1228.623    382.837  -3.209  0.00136 **
## age:smokeryes       -2.542     23.711  -0.107  0.91464
## smokeryes:bmi     1438.525     52.793  27.249  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4853 on 1328 degrees of freedom
## Multiple R-squared:  0.8405, Adjusted R-squared:  0.8394
## F-statistic: 777.5 on 9 and 1328 DF,  p-value: < 2.2e-16
```

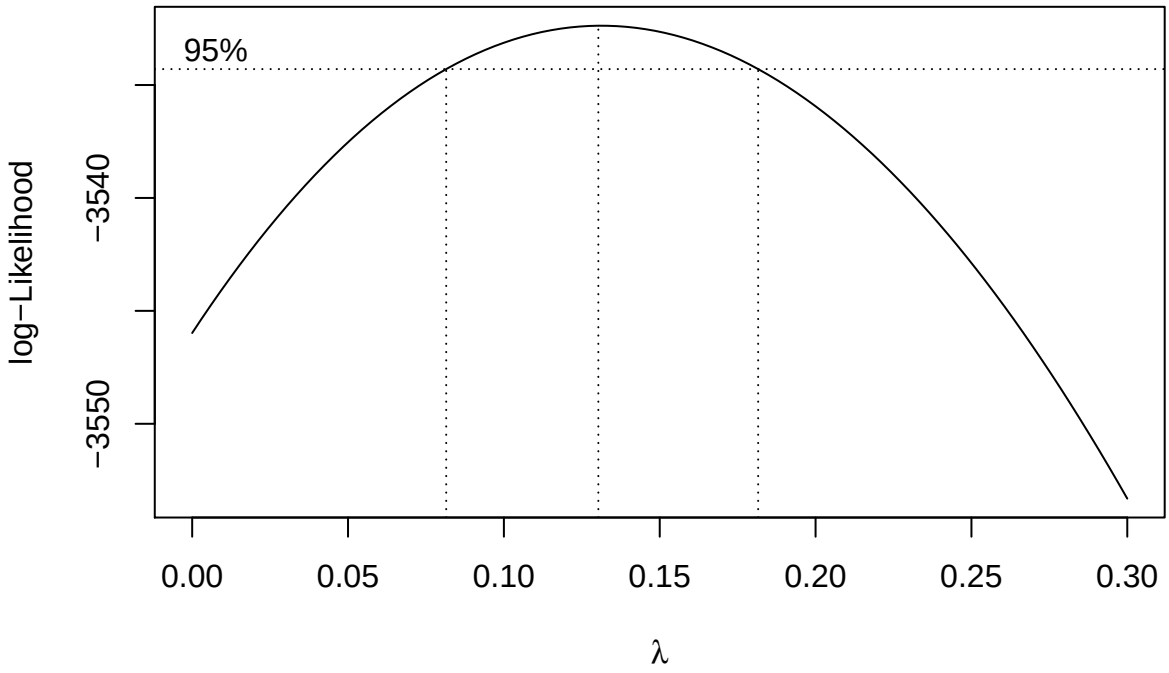
Residual Plot

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



This residual plot is a little better, lets see if we can transform the response with this new equation.

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



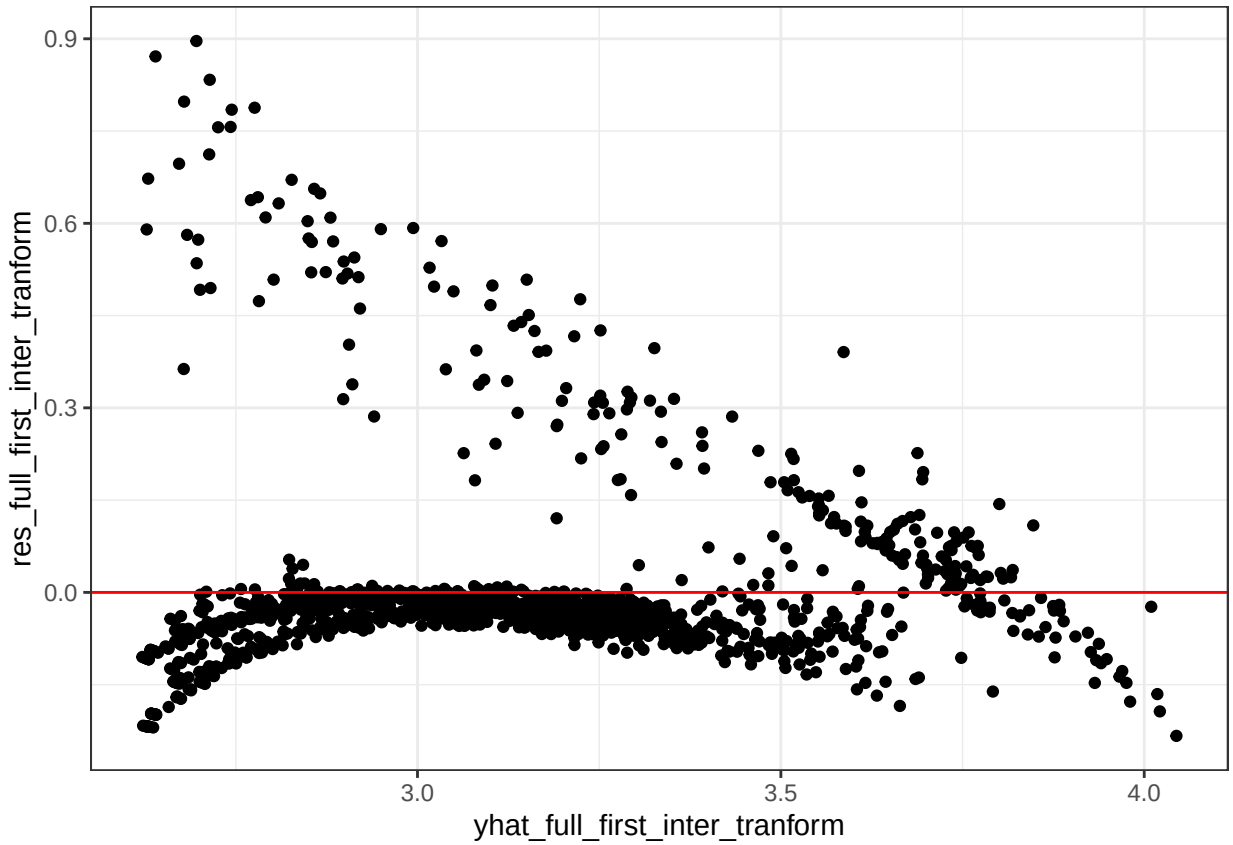
Maybe we can use a lambda value of 0.125

```r
interaction_transform <- data
interaction_transform$charges <- interaction_transform$charges^0.125
mlr_interaction_tranform <- lm(charges ~  age*smoker + bmi*smoker + children  + region, data=interactio
summary(mlr_interaction_tranform)
```

```
##
## Call:
## lm(formula = charges ~ age * smoker + bmi * smoker + children +
##     region, data = interaction_transform)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23325 -0.05925 -0.03216 -0.00578  0.89638
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.3898138  0.0268794  88.909  < 2e-16 ***
## age              0.0152818  0.0003296  46.358  < 2e-16 ***
## smokeryes        0.3867297  0.0565870   6.834 1.25e-11 ***
## bmi              0.0004622  0.0007932   0.583   0.5602
## children         0.0371914  0.0034081  10.913  < 2e-16 ***
## regionnorthwest -0.0243318  0.0117809  -2.065   0.0391 *
## regionsoutheast -0.0531652  0.0118451  -4.488 7.80e-06 ***
## regionsouthwest -0.0559589  0.0118258  -4.732 2.46e-06 ***
## age:smokeryes   -0.0115120  0.0007324 -15.717  < 2e-16 ***
## smokeryes:bmi    0.0223756  0.0016308  13.721  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1499 on 1328 degrees of freedom
## Multiple R-squared:  0.827,  Adjusted R-squared:  0.8259
## F-statistic: 705.6 on 9 and 1328 DF,  p-value: < 2.2e-16
```

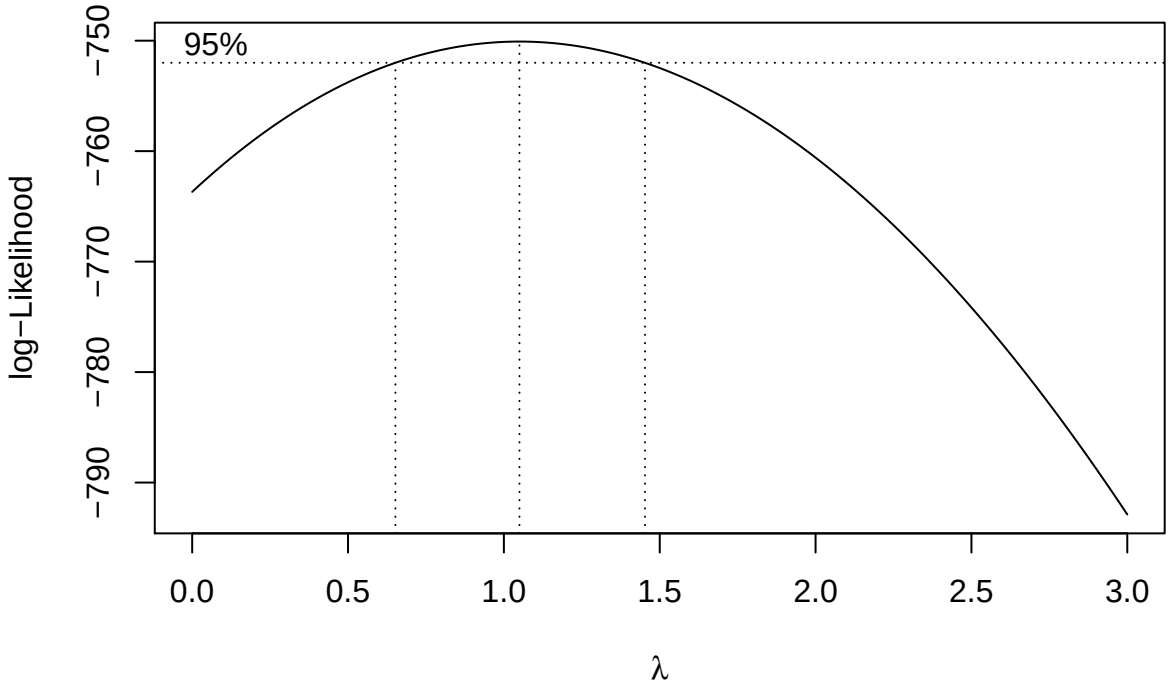
Recheck Residual Plot

```r
yhat_full_first_inter_tranform <- mlr_interaction_tranform$fitted.values
res_full_first_inter_tranform <- mlr_interaction_tranform$residuals
data %>%
  ggplot(aes(yhat_full_first_inter_tranform, res_full_first_inter_tranform)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



Still see the same without adding the interaction terms.

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



Still no luck. We retried this many times, but weren't lucky.

Partial F test of the interaction vs simple model after two transformation of response variable

```
full <- mlr_interaction_tranform
reduced <- lm(charges ~ age + bmi + children + smoker + region, data=interaction_transform)
anova(reduced, full)
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ age + bmi + children + smoker + region
## Model 2: charges ~ age * smoker + bmi * smoker + children + region
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1330 38.959
## 2   1328 29.842  2    9.1174 202.87 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can't drop the interaction terms.

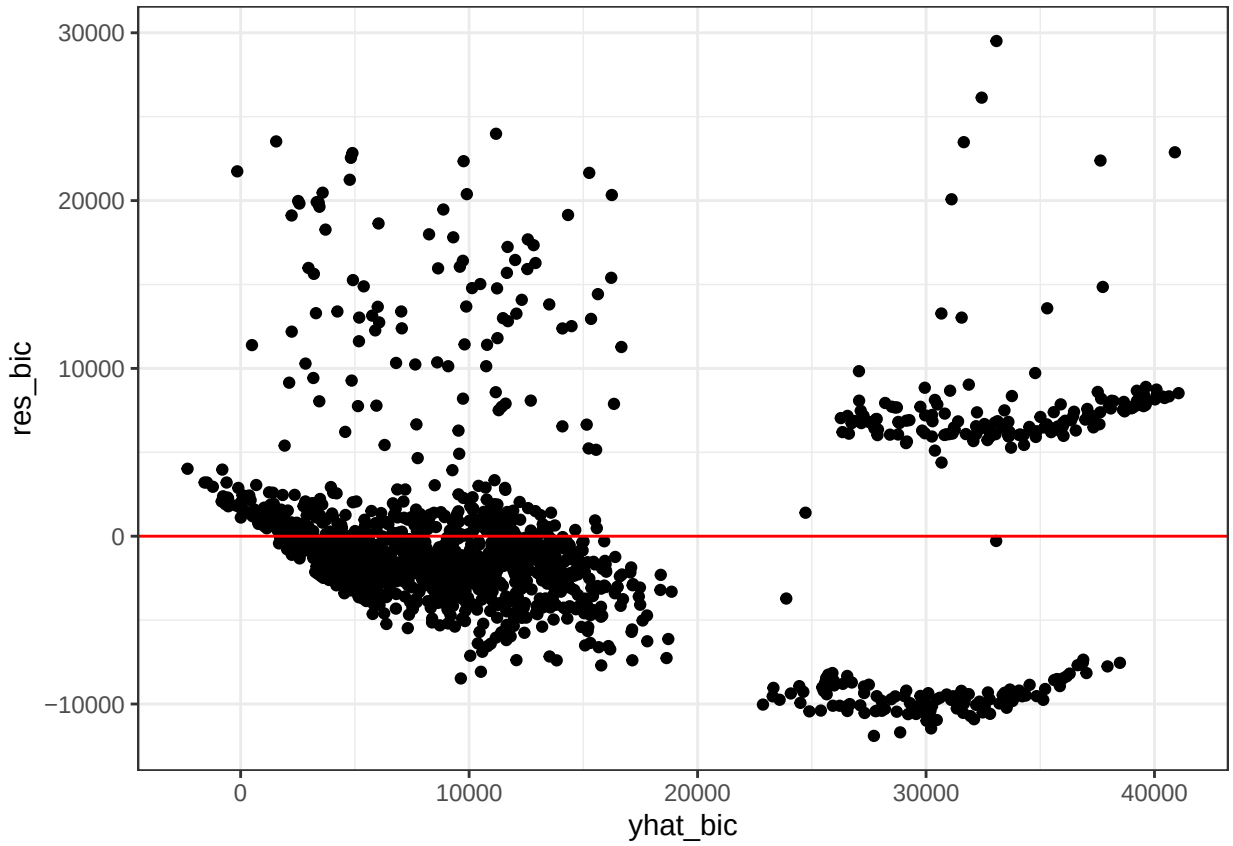
BIC Model selection model might be better

```
bic_selection_model = lm(charges ~  age + bmi + children  + smoker, data=data)
summary(bic_selection_model)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11897.9  -2920.8   -986.6   1392.2  29509.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -12102.77     941.98 -12.848  < 2e-16 ***
## age            257.85      11.90  21.675  < 2e-16 ***
## bmi            321.85      27.38  11.756  < 2e-16 ***
## children       473.50     137.79   3.436 0.000608 ***
## smokeryes    23811.40     411.22  57.904  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6068 on 1333 degrees of freedom
## Multiple R-squared:  0.7497, Adjusted R-squared:  0.7489
## F-statistic: 998.1 on 4 and 1333 DF,  p-value: < 2.2e-16
```

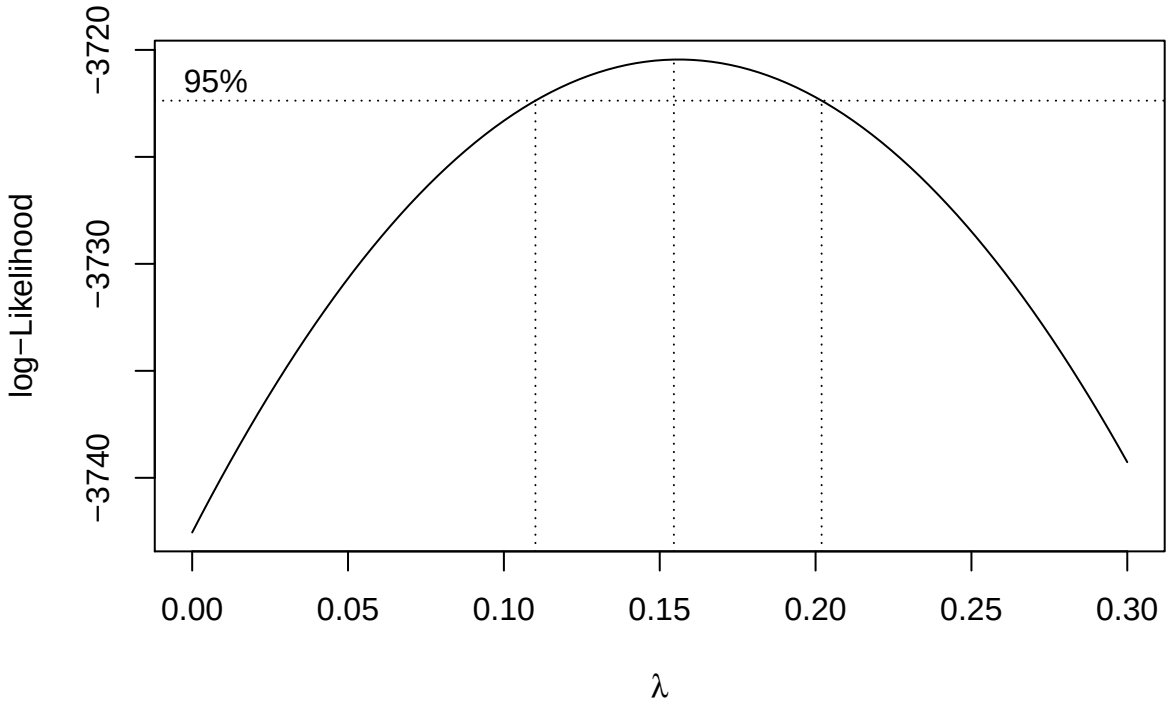
Residual Plot

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



We see a similar plot. Transformation?

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



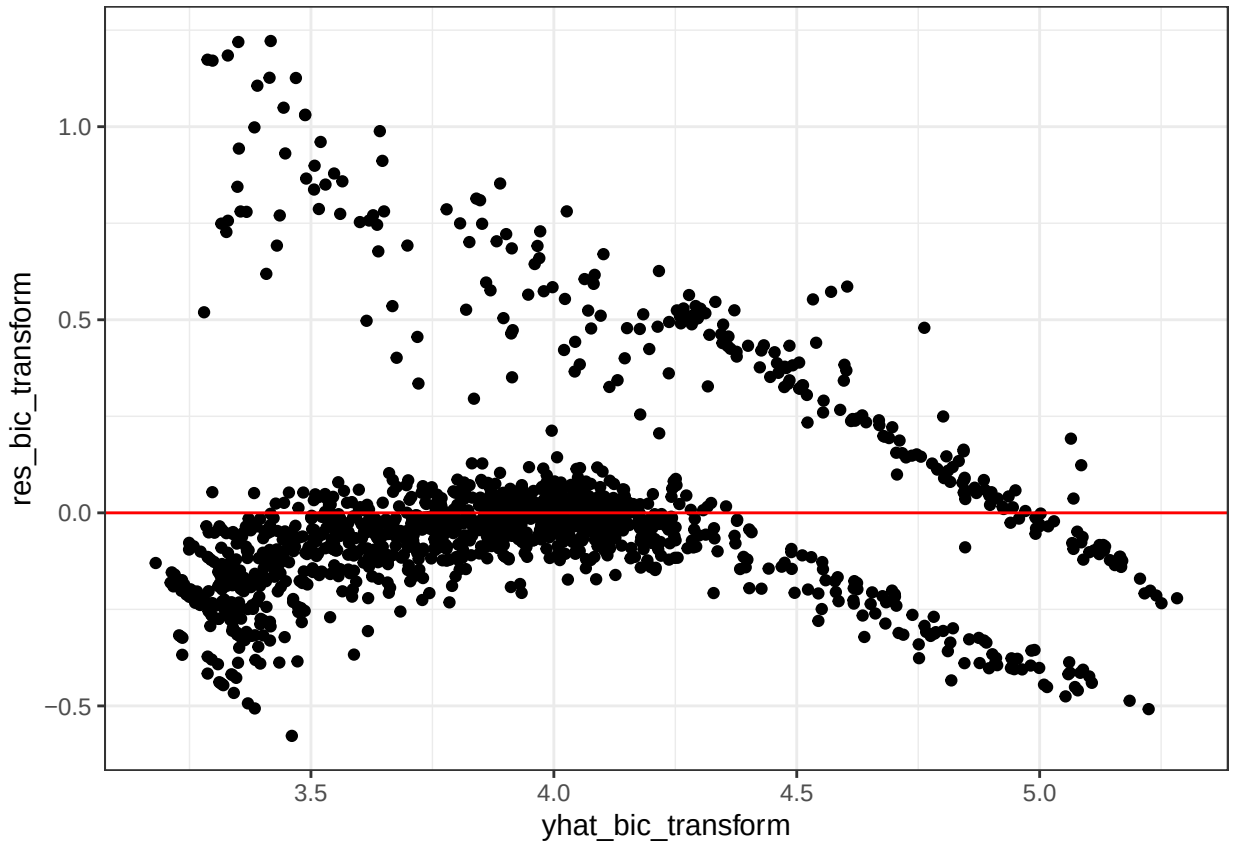
Again lambda of 0.15

```r
bic_transform <- data
bic_transform$charges <- bic_transform$charges^(0.15)
bic_selection_model_transform = lm(charges ~  age + bmi + children  + smoker, data=bic_transform)
summary(bic_selection_model_transform)
```

```
##
## Call:
## lm(formula = charges ~ age + bmi + children + smoker, data = bic_transform)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.57755 -0.12028 -0.03776  0.03505  1.22187
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.7136331  0.0402741  67.379  < 2e-16 ***
## age         0.0192458  0.0005086  37.839  < 2e-16 ***
## bmi         0.0075402  0.0011705   6.442 1.65e-10 ***
## children    0.0523899  0.0058912   8.893  < 2e-16 ***
## smokeryes   0.9539751  0.0175815  54.260  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2594 on 1333 degrees of freedom
## Multiple R-squared:  0.7719, Adjusted R-squared:  0.7712
## F-statistic:  1128 on 4 and 1333 DF,  p-value: < 2.2e-16
```

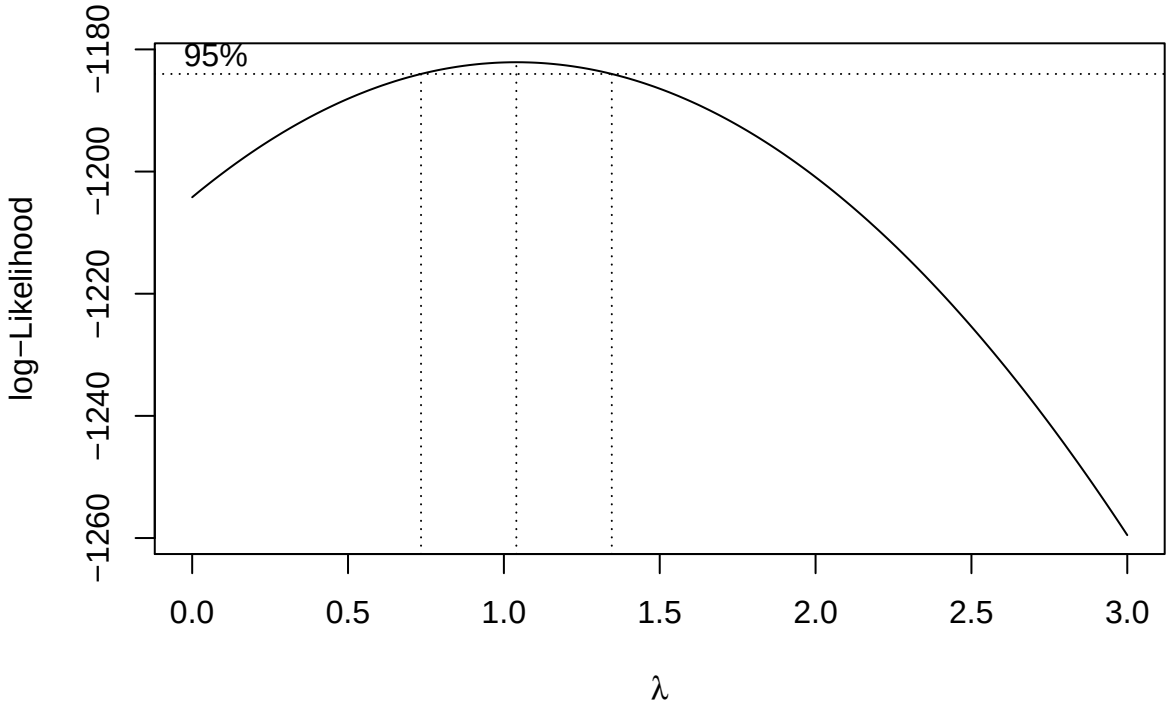
Residual Plot

```r
yhat_bic_transform <- bic_selection_model_transform$fitted.values
res_bic_transform <- bic_selection_model_transform$residuals
data %>%
  ggplot(aes(yhat_bic_transform, res_bic_transform)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



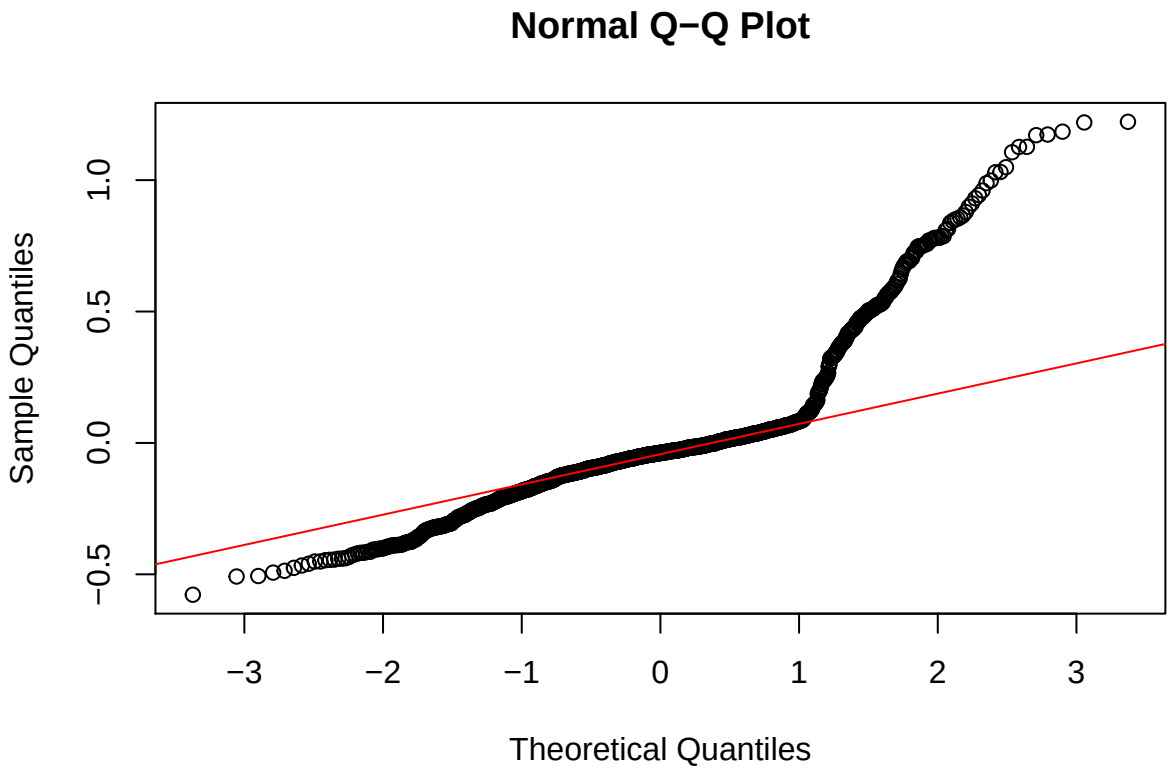
Same Stuff happening.

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



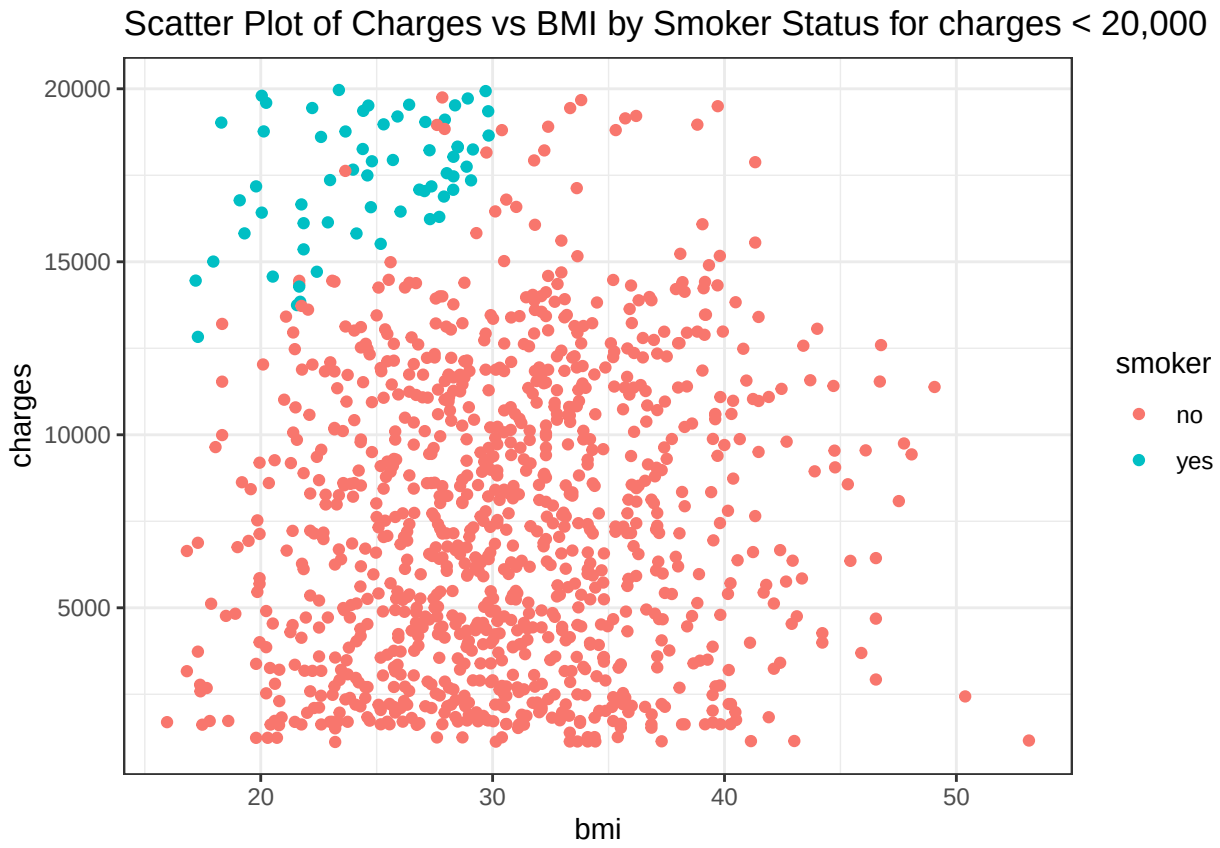
QQPlot

```
{
  qqnorm(bic_selection_model_transform$residuals)
  qqline(bic_selection_model_transform$residuals, col="red")
}
```

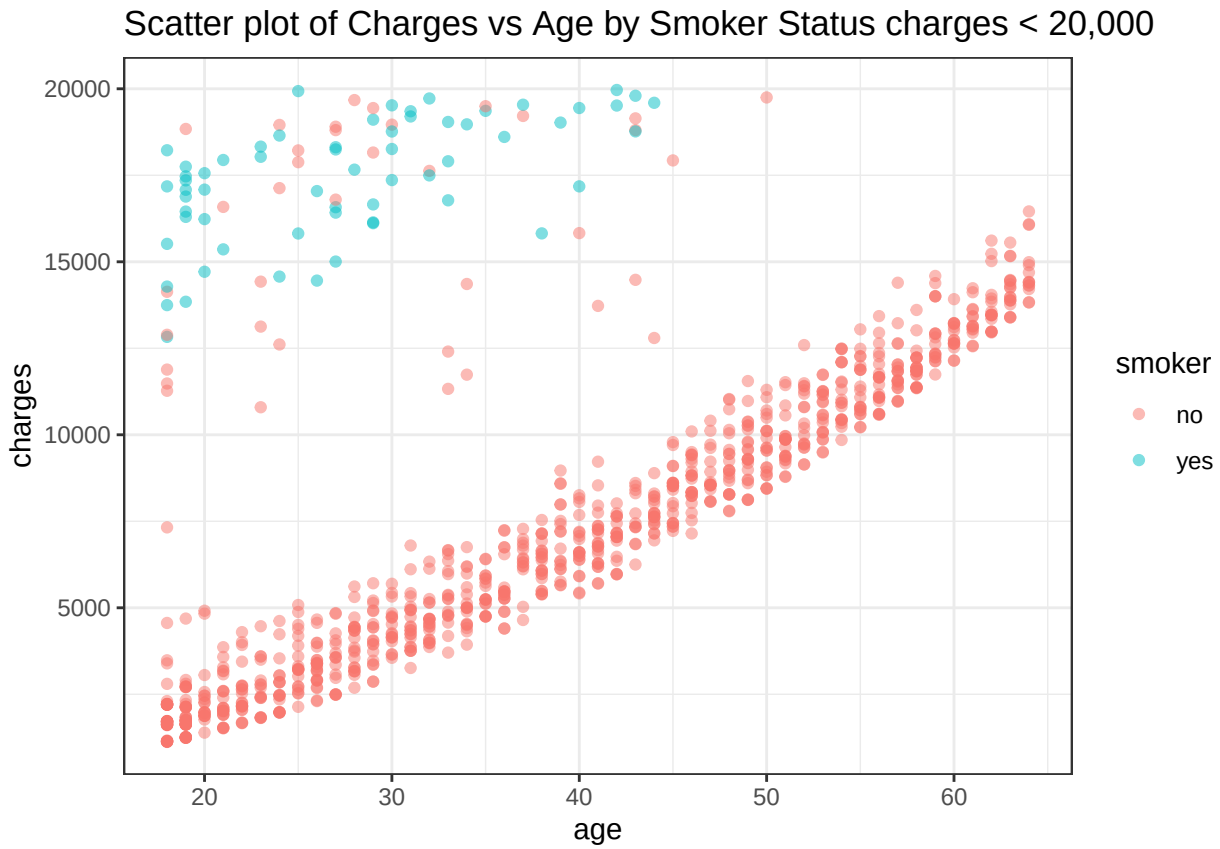


Change the data split the data ?

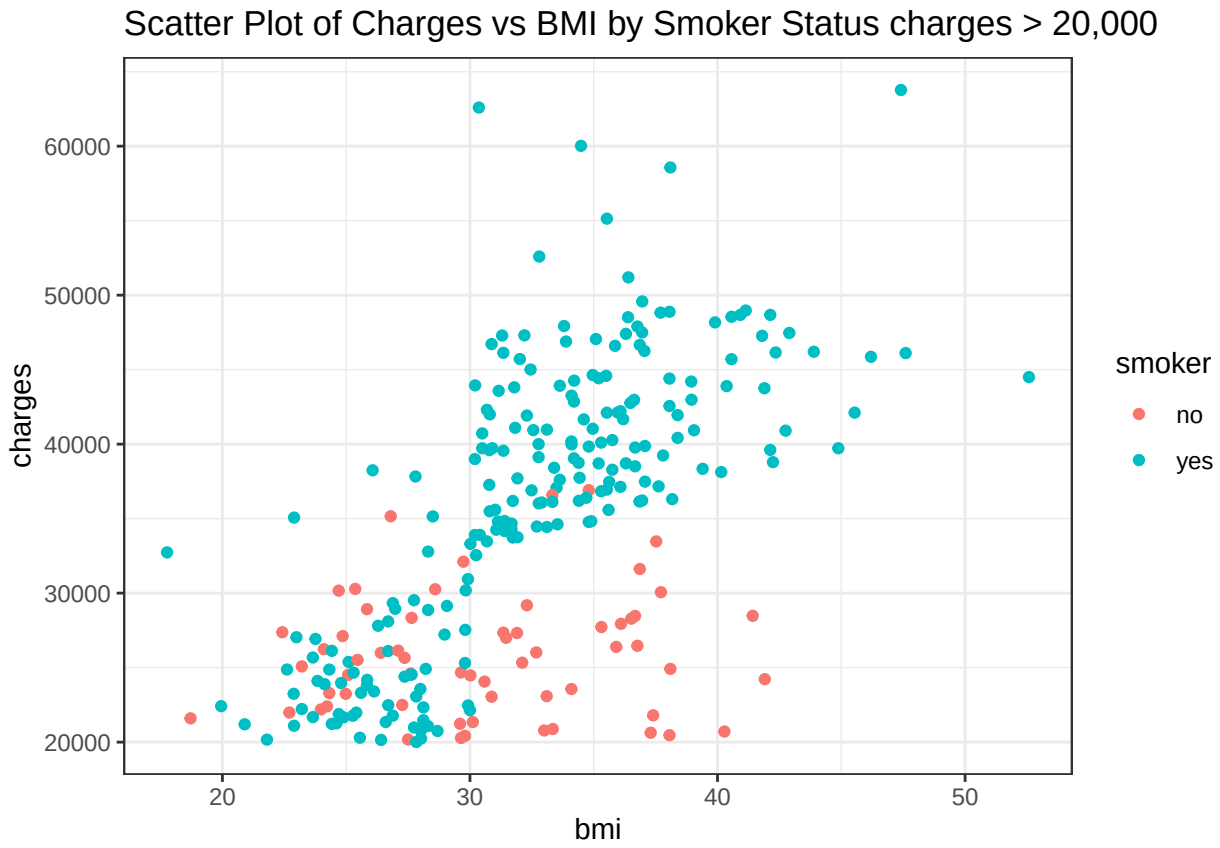
```
less_charge <- data[data$charges < 20000,]
more_change <- data[data$charges >= 20000, ]
library(tidyverse)
ggplot(aes(x=bmi, y=charges, color=smoker), data=less_charge) +
  labs(title="Scatter Plot of Charges vs BMI by Smoker Status for charges < 20,000") +
  theme_bw() +
  geom_point()
```



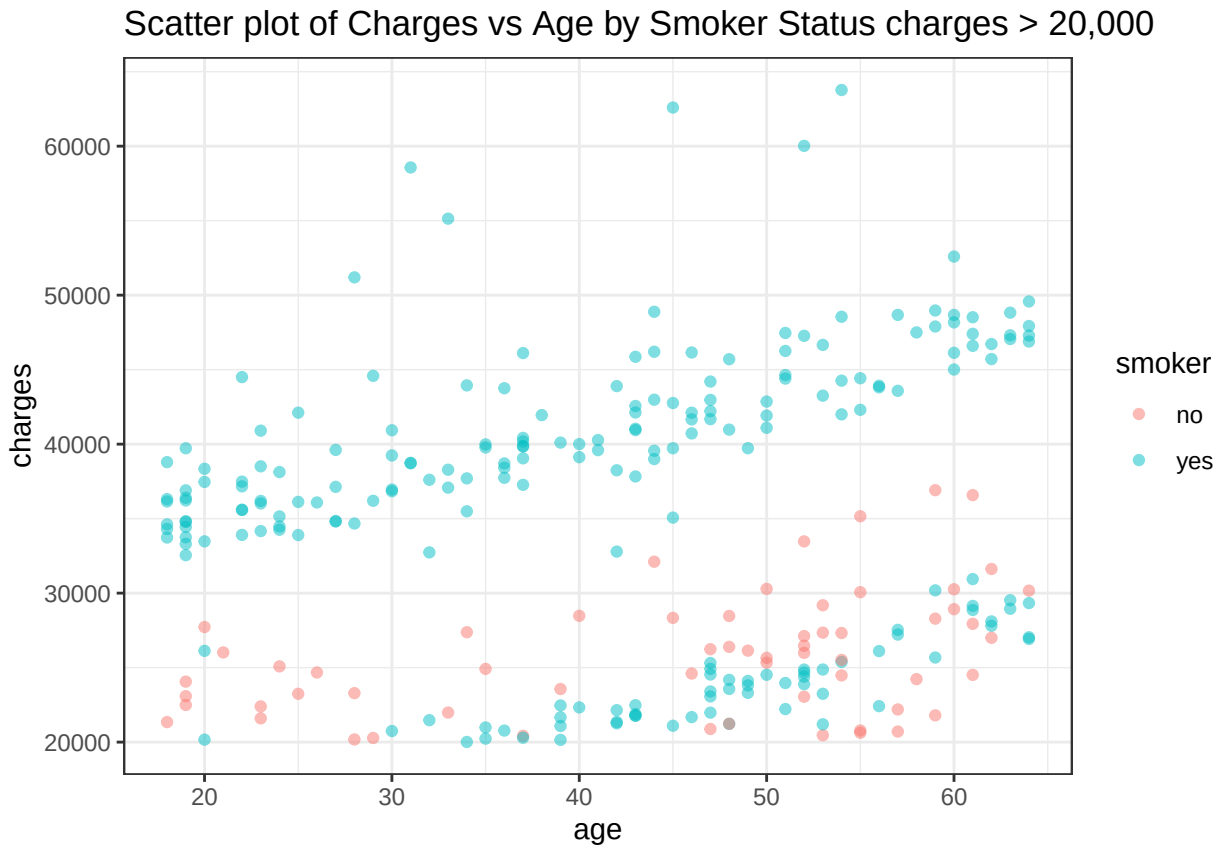

```
ggplot(aes(x=age,y=charges, color=smoker), data=less_charge) +
  labs(title="Scatter plot of Charges vs Age by Smoker Status charges < 20,000") +
  theme_bw() +
  geom_point(alpha=0.5)
```



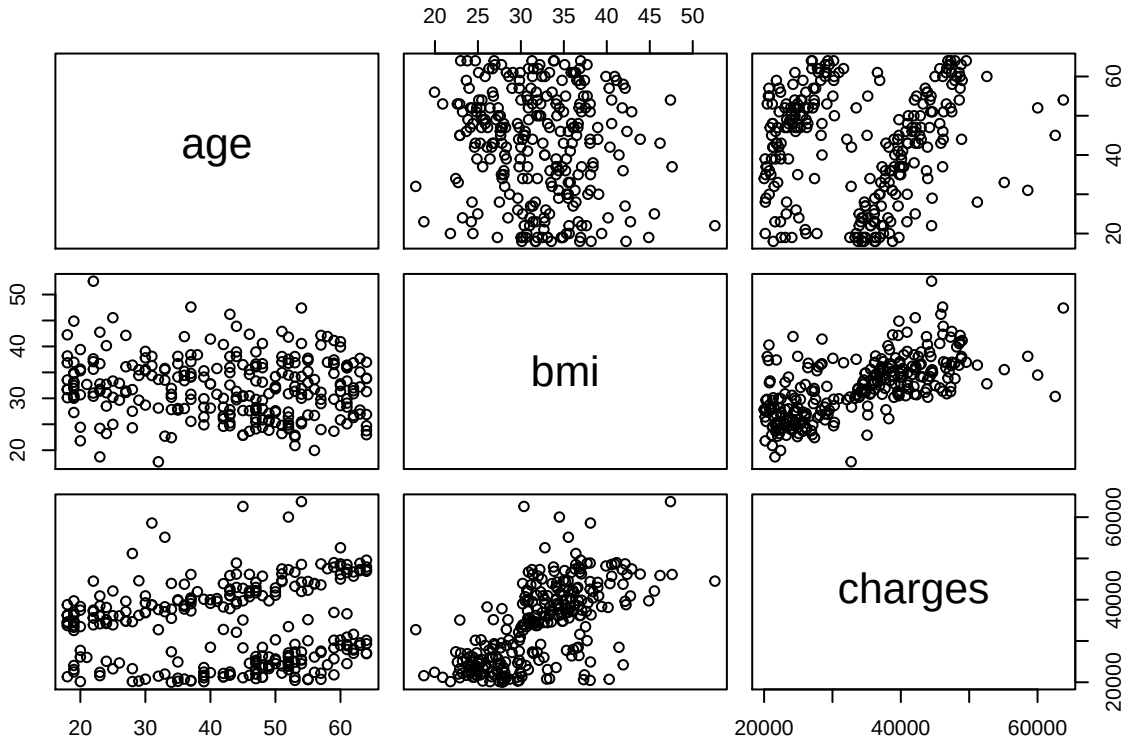

```
ggplot(aes(x=bmi, y=charges, color=smoker), data=more_change) +
  labs(title="Scatter Plot of Charges vs BMI by Smoker Status charges > 20,000") +
  theme_bw() +
  geom_point()
```



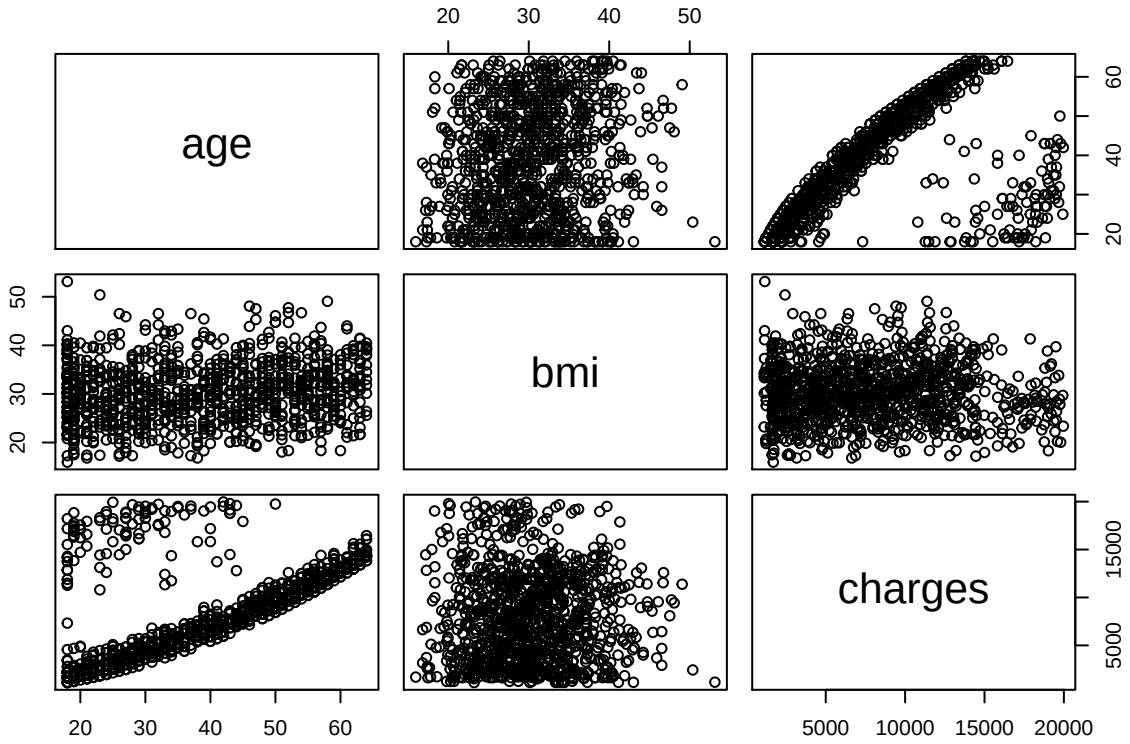

```
ggplot(aes(x=age,y=charges, color=smoker), data=more_change) +
  labs(title="Scatter plot of Charges vs Age by Smoker Status charges > 20,000") +
  theme_bw() +
  geom_point(alpha=0.5)
```




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



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



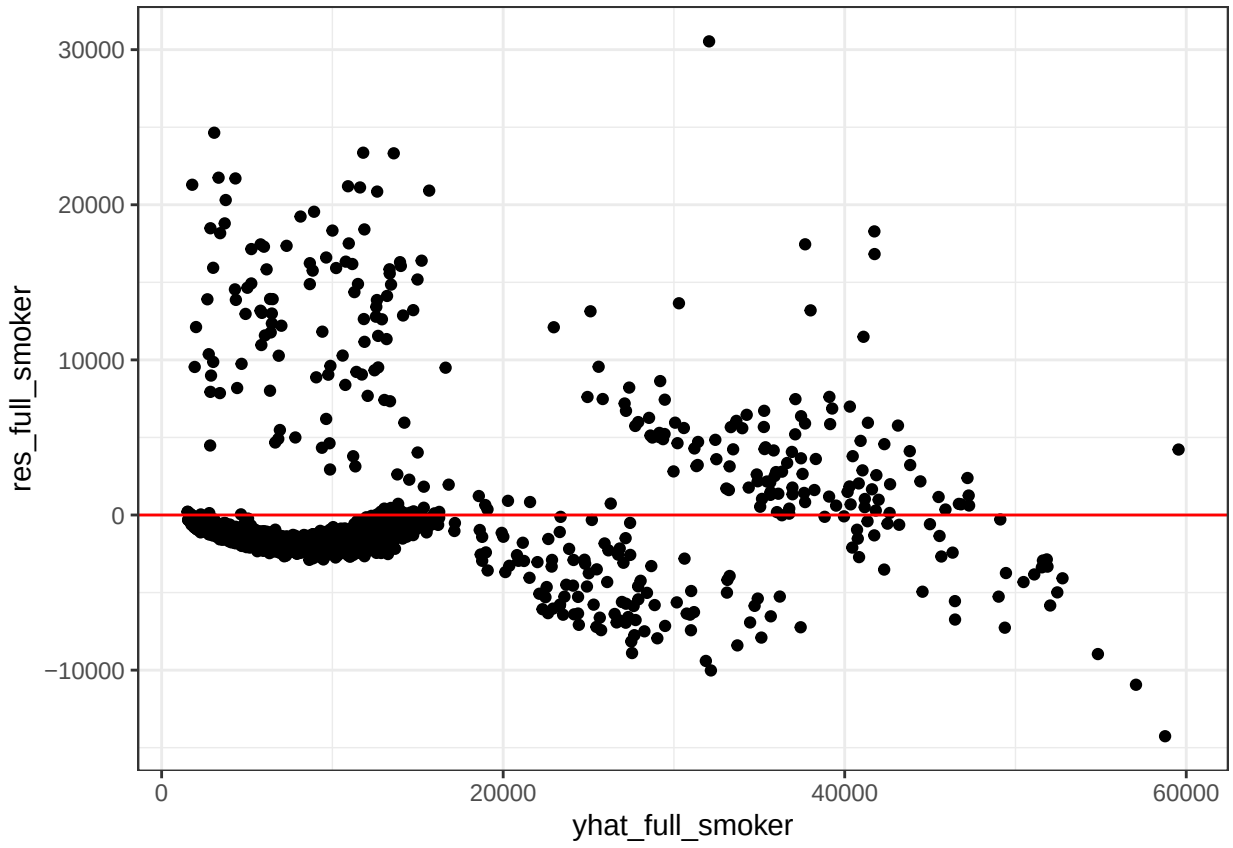

```r
mlr_full_smoker = lm(charges ~  (age + bmi + children + region) * smoker, data=data)
summary(mlr_full_smoker)
```

```
##
## Call:
## lm(formula = charges ~ (age + bmi + children + region) * smoker,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14260.4  -1922.0  -1299.4   -410.1  30534.4
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   -2450.179    878.080  -2.790  0.00534 **
## age                             264.947     10.671  24.828  < 2e-16 ***
## bmi                              17.939     25.910   0.692  0.48883
## children                        586.903    122.242   4.801 1.76e-06 ***
## regionnorthwest                 -552.249    423.912  -1.303  0.19289
## regionsoutheast                 -989.701    435.141  -2.274  0.02310 *
## regionsouthwest                -1385.975    424.969  -3.261  0.00114 **
## smokeryes                     -20310.836   1907.109 -10.650  < 2e-16 ***
## age:smokeryes                     2.673     23.819   0.112  0.91067
## bmi:smokeryes                  1453.239     55.209  26.323  < 2e-16 ***
## children:smokeryes             -417.913    283.549  -1.474  0.14076
## regionnorthwest:smokeryes      -105.947    969.919  -0.109  0.91303
## regionsoutheast:smokeryes      -898.633    920.894  -0.976  0.32933
## regionsouthwest:smokeryes       935.973    977.752   0.957  0.33861
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4850 on 1324 degrees of freedom
## Multiple R-squared:  0.8412, Adjusted R-squared:  0.8396
## F-statistic: 539.5 on 13 and 1324 DF,  p-value: < 2.2e-16
```

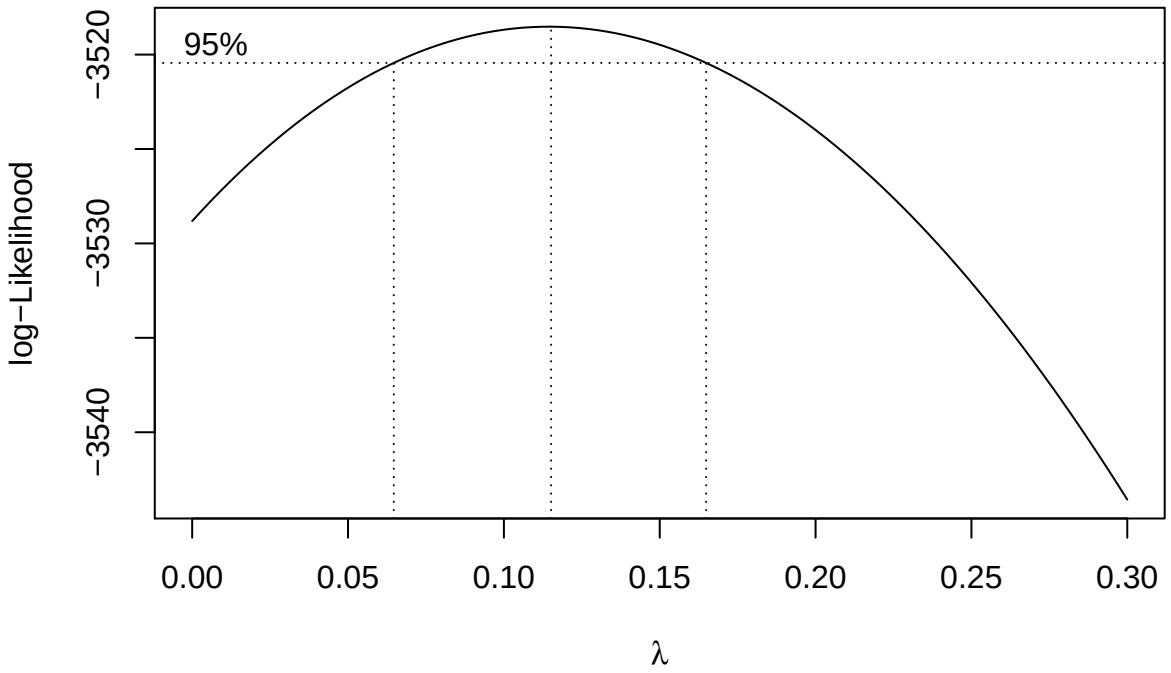
Assumption Check of Full Model

```
yhat_full_smoker <- mlr_full_smoker$fitted.values
res_full_smoker <- mlr_full_smoker$residuals
data %>%
  ggplot(aes(yhat_full_smoker, res_full_smoker)) +
  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_smoker, lambda=seq(0,0.3, 0.01))
```



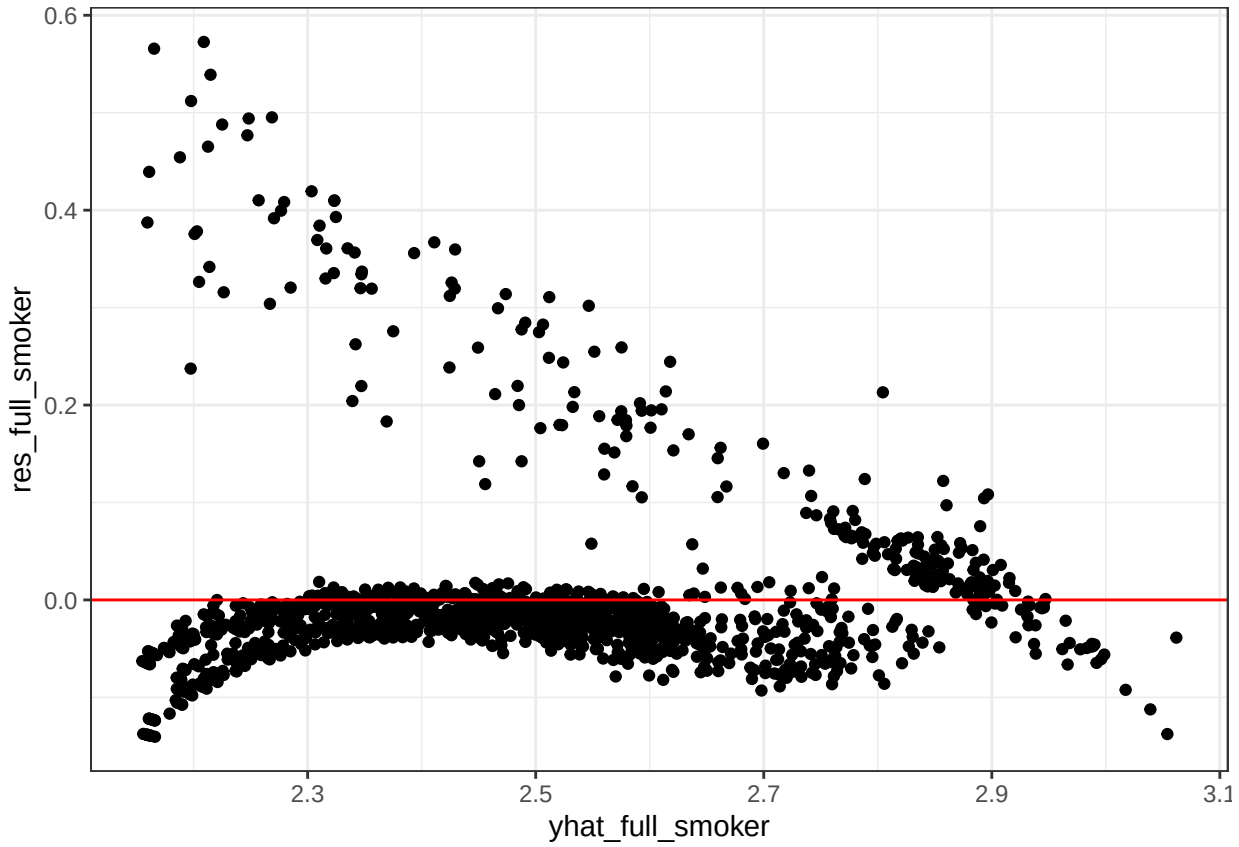
From the boxcox we can try a lambda value of 0.1 for transformation.

```
interaction_transform_smoker <- data
interaction_transform_smoker$charges <- interaction_transform_smoker$charges^0.1
mlr_interaction_tranform_smoker <- lm(charges ~  (age + bmi + children + region) * smoker, data=interac
summary(mlr_interaction_tranform_smoker)
```

```
##
## Call:
## lm(formula = charges ~ (age + bmi + children + region) * smoker,
##     data = interaction_transform_smoker)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.14041 -0.03918 -0.01973 -0.00012  0.57262
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 2.0103535  0.0171453 117.254  < 2e-16 ***
## age                         0.0098368  0.0002084  47.209  < 2e-16 ***
## bmi                         0.0003437  0.0005059   0.679   0.4970
## children                    0.0292642  0.0023869  12.260  < 2e-16 ***
## regionnorthwest            -0.0180006  0.0082773  -2.175   0.0298 *
## regionsoutheast            -0.0395694  0.0084965  -4.657 3.53e-06 ***
## regionsouthwest            -0.0433088  0.0082979  -5.219 2.08e-07 ***
## smokeryes                   0.2783934  0.0372380   7.476 1.39e-13 ***
## age:smokeryes              -0.0073089  0.0004651 -15.715  < 2e-16 ***
## bmi:smokeryes               0.0134167  0.0010780  12.446  < 2e-16 ***
## children:smokeryes         -0.0273920  0.0055365  -4.947 8.48e-07 ***
## regionnorthwest:smokeryes   0.0142893  0.0189385   0.755   0.4507
## regionsoutheast:smokeryes   0.0237658  0.0179813   1.322   0.1865
## regionsouthwest:smokeryes   0.0413423  0.0190915   2.165   0.0305 *
## ---
```

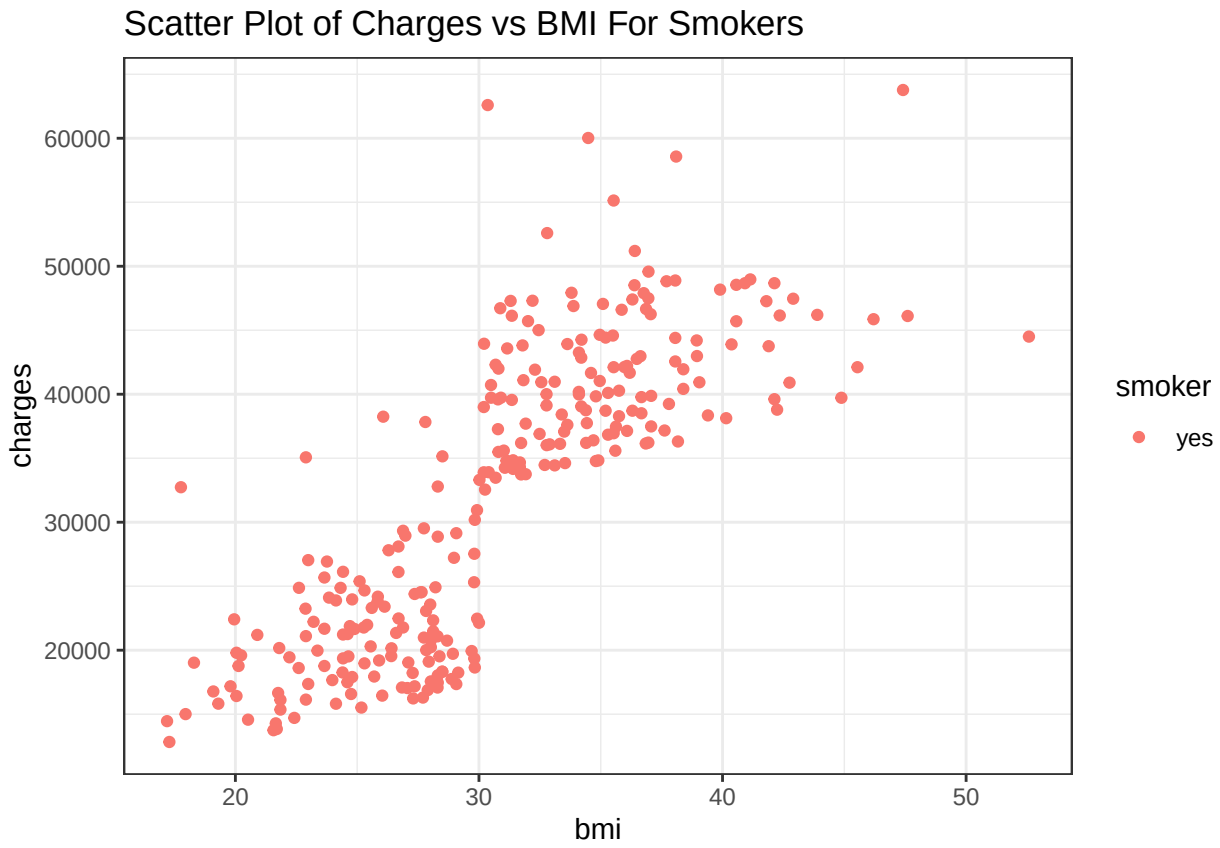
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09469 on 1324 degrees of freedom
## Multiple R-squared:   0.83,  Adjusted R-squared:  0.8283
## F-statistic: 497.3 on 13 and 1324 DF,  p-value: < 2.2e-16
```

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

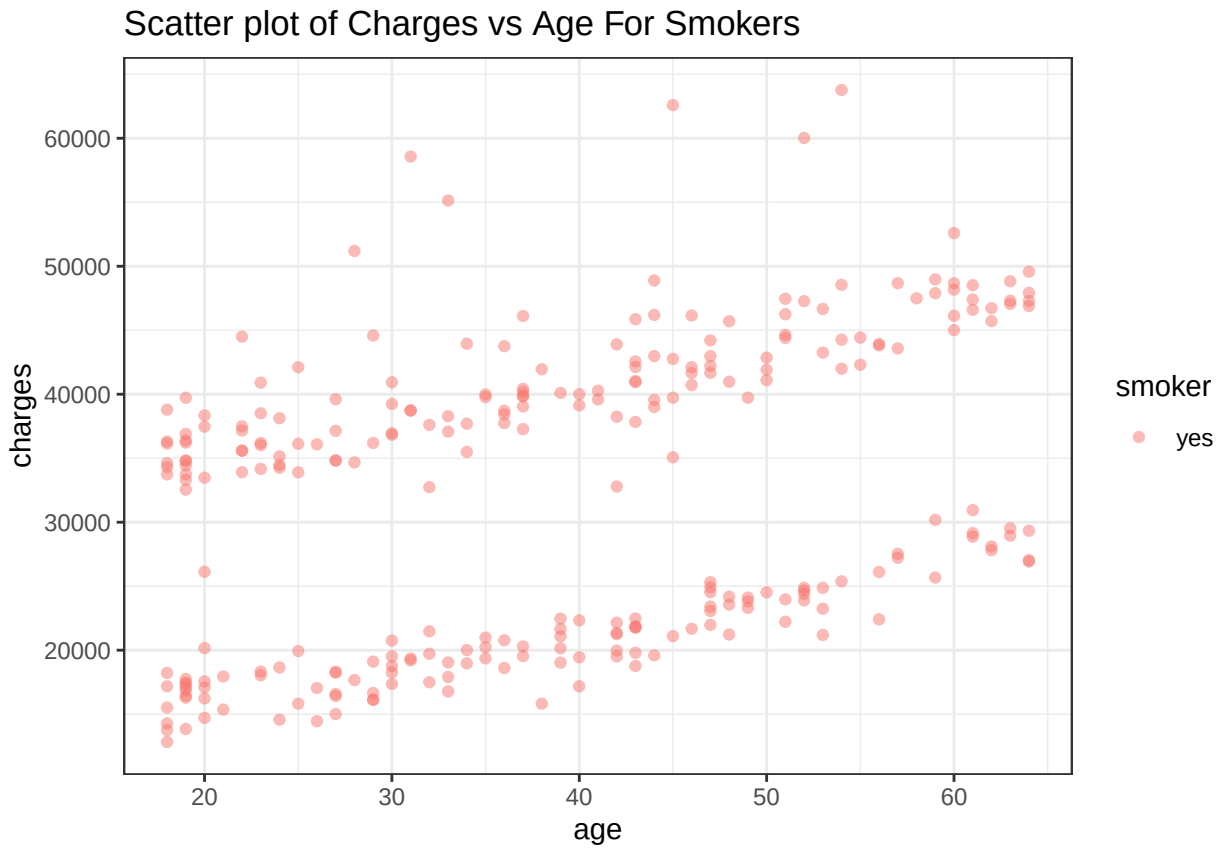


Lets SPLIT BY SMOKERS

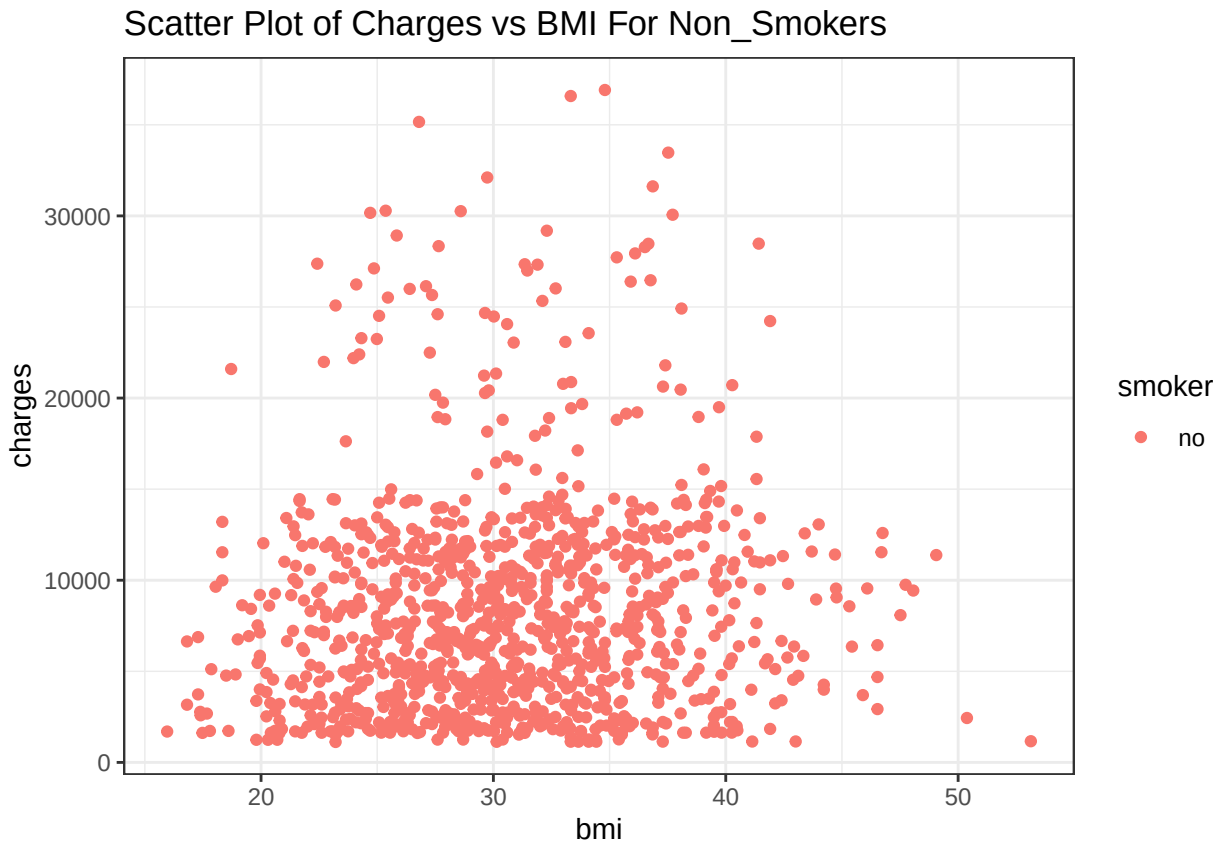
```
smokers <- data[data$smoker == 'yes',]
non_smokers <- data[data$smoker != 'yes',]

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



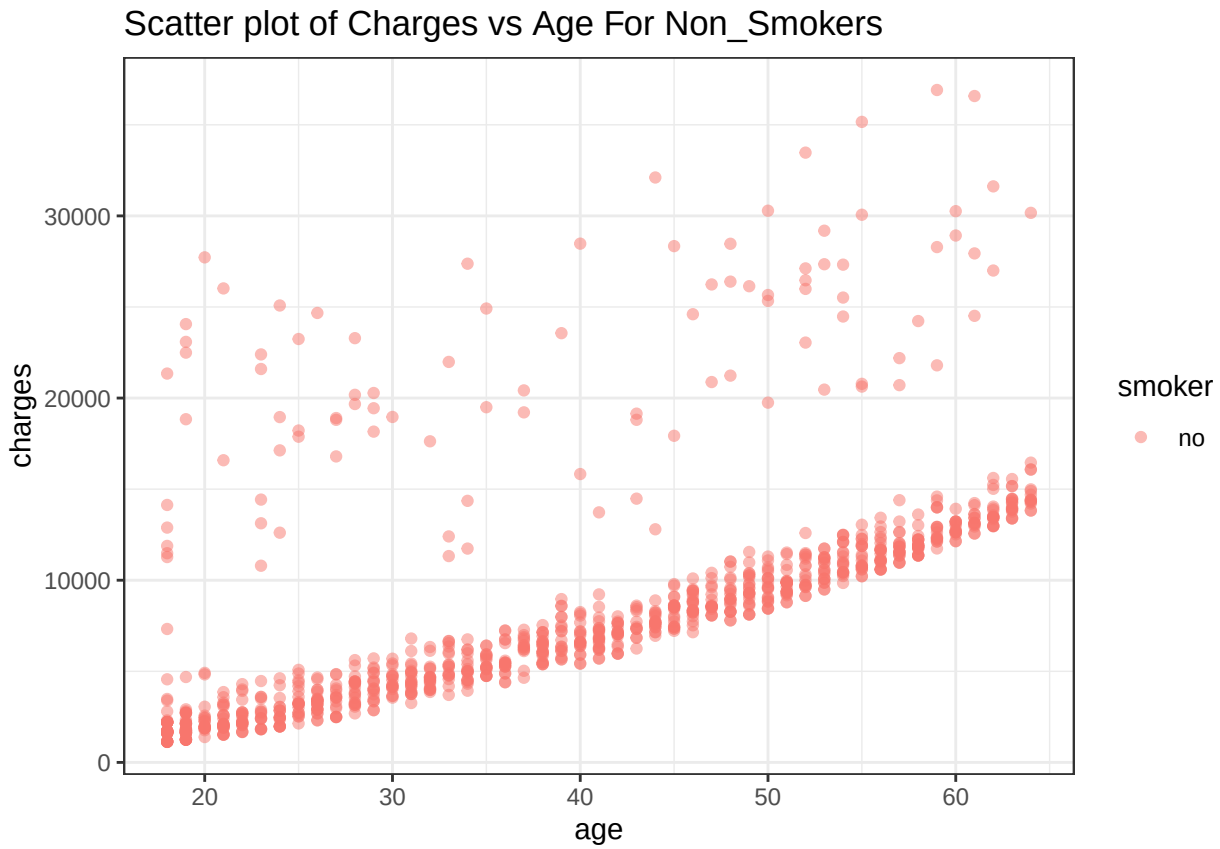

```
ggplot(aes(x=age,y=charges, color=smoker), data=smokers) +
  labs(title="Scatter plot of Charges vs Age For Smokers") +
  theme_bw() +
  geom_point(alpha=0.5)
```




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




```
ggplot(aes(x=age,y=charges, color=smoker), data=non_smokers) +
  labs(title="Scatter plot of Charges vs Age For Non_Smokers") +
  theme_bw() +
  geom_point(alpha=0.5)
```




```
## Smokers
##intercept only model
regnull_smoker <- lm(charges~1, data=smokers)
##model with all predictors
regfull_smoker <- lm(charges ~ age + sex + bmi + children + region , data=smokers)
```

Forward Selection

```
step(regnull_smoker, scope=list(lower=regnull_smoker, upper=regfull_smoker), direction="forward")
```

```
## Start:  AIC=5126.83
## charges ~ 1
##
##             Df  Sum of Sq        RSS    AIC
## + bmi        1 2.3653e+10 1.2713e+10 4840.9
## + age        1 4.9308e+09 3.1435e+10 5088.9
## + region     3 1.2923e+09 3.5073e+10 5122.9
## + sex        1 3.7263e+08 3.5993e+10 5126.0
## <none>                    3.6366e+10 5126.8
## + children   1 4.6986e+07 3.6319e+10 5128.5
##
## Step:  AIC=4840.86
## charges ~ bmi
##
##             Df  Sum of Sq        RSS    AIC
## + age        1 3739448620 8.9736e+09 4747.4
## <none>                    1.2713e+10 4840.9
## + children   1   77371010 1.2636e+10 4841.2
```

```
## + sex       1   12609906 1.2700e+10 4842.6
## + region    3  112969895 1.2600e+10 4844.4
##
## Step:  AIC=4747.41
## charges ~ bmi + age
##
##            Df Sum of Sq        RSS    AIC
## <none>                  8973564816 4747.4
## + region    3 149563719 8824001097 4748.8
## + children  1  14356763 8959208053 4749.0
## + sex       1   7002694 8966562122 4749.2
```

```
##
## Call:
## lm(formula = charges ~ bmi + age, data = smokers)
##
## Coefficients:
## (Intercept)          bmi          age
##    -22367.4       1438.1        266.3
```

```
#Non smokers
regnull_non_smokers <- lm(charges~1, data=non_smokers)
##model with all predictors
regfull_non_smokers <- lm(charges ~ age + sex + bmi + children + region , data=non_smokers)
```

Forward Selection

```
step(regnull_non_smokers, scope=list(lower=regnull_non_smokers, upper=regfull_non_smokers), direction="
```

```
## Start:  AIC=18511.36
## charges ~ 1
##
##            Df  Sum of Sq        RSS   AIC
## + age       1 1.5058e+10 2.3130e+10 17980
## + children  1 7.3709e+08 3.7452e+10 18493
## + bmi       1 2.6969e+08 3.7919e+10 18506
## + sex       1 1.2113e+08 3.8068e+10 18510
## + region    3 2.3153e+08 3.7957e+10 18511
## <none>                   3.8189e+10 18511
##
## Step:  AIC=17979.87
## charges ~ age
##
##            Df Sum of Sq        RSS   AIC
## + children  1 531956489 2.2598e+10 17957
## + region    3 248975743 2.2881e+10 17974
## + sex       1  68342342 2.3062e+10 17979
## <none>                   2.3130e+10 17980
## + bmi       1   1914187 2.3128e+10 17982
##
## Step:  AIC=17957.12
## charges ~ age + children
##
##          Df Sum of Sq        RSS   AIC
## + region  3 262141210 2.2336e+10 17951
## + sex     1  69429619 2.2529e+10 17956
```

```
## <none>                2.2598e+10 17957
## + bmi     1   1065536 2.2597e+10 17959
##
## Step:  AIC=17950.7
## charges ~ age + children + region
##
##        Df Sum of Sq        RSS   AIC
## + sex   1  72093470 2.2264e+10 17949
## <none>              2.2336e+10 17951
## + bmi   1  11274021 2.2325e+10 17952
##
## Step:  AIC=17949.26
## charges ~ age + children + region + sex
##
##        Df Sum of Sq        RSS   AIC
## <none>              2.2264e+10 17949
## + bmi   1  12614493 2.2251e+10 17951
```

```
##
## Call:
## lm(formula = charges ~ age + children + region + sex, data = non_smokers)
##
## Coefficients:
##     (Intercept)              age         children  regionnorthwest
##         -1695.9            265.5            589.1           -550.2
## regionsoutheast  regionsouthwest          sexmale
##          -913.2          -1373.0           -521.0
```

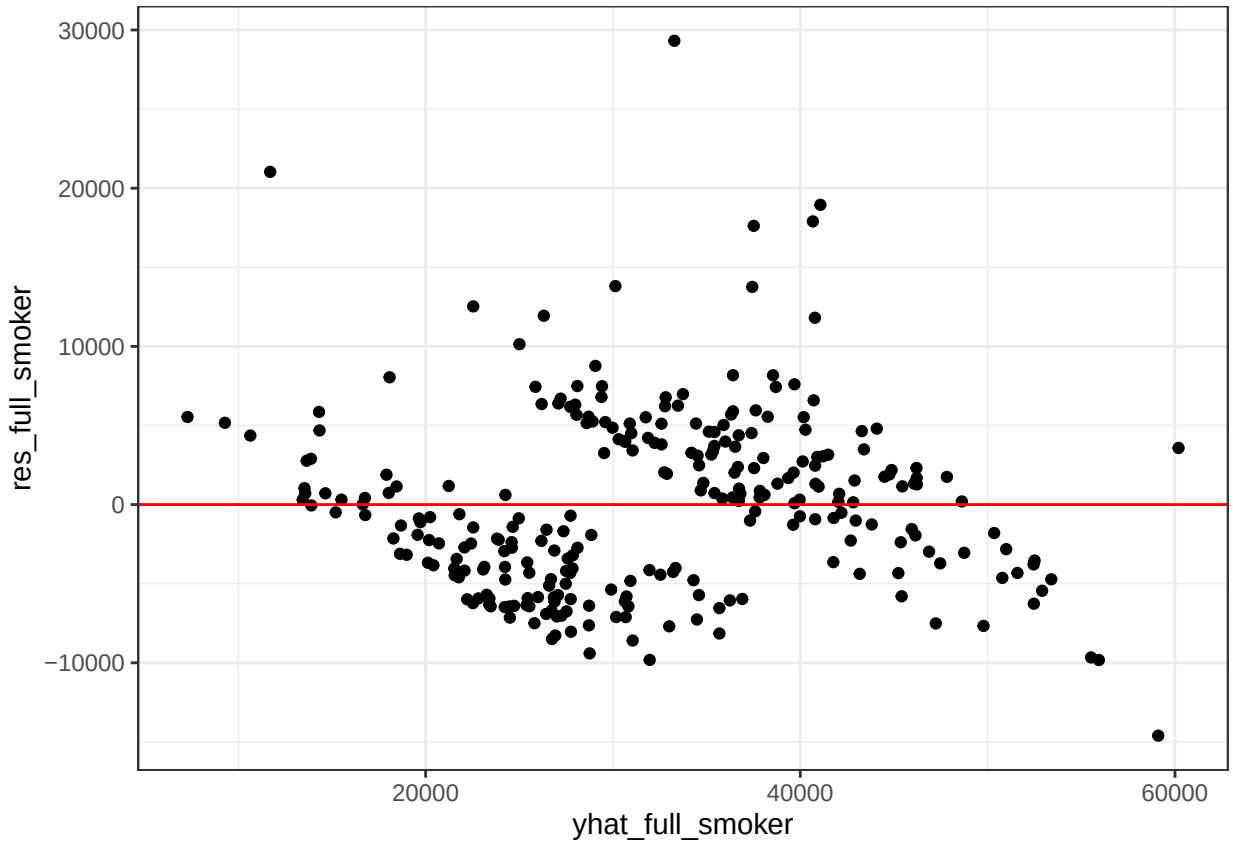
We get different models

```
mlr_full_smoker = lm(charges ~  bmi+age , data=smokers)
summary(mlr_full_smoker)
```

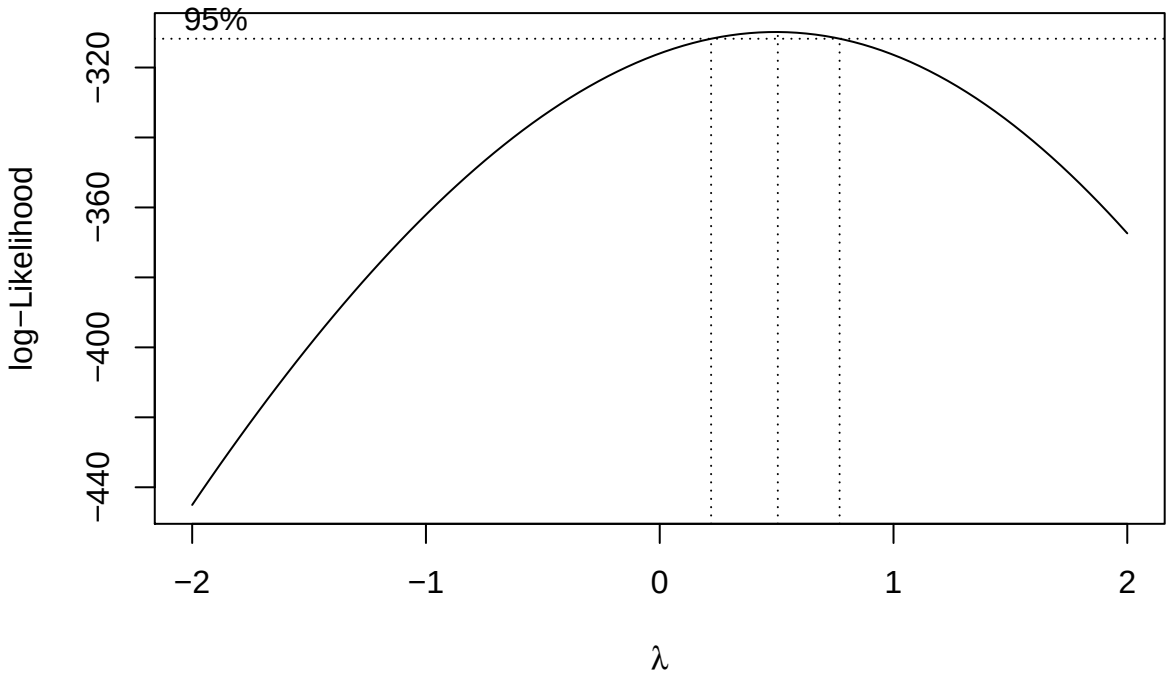
```
##
## Call:
## lm(formula = charges ~ bmi + age, data = smokers)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14604.4  -4315.1   -240.5   3638.0  29316.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -22367.45    1931.86  -11.58   <2e-16 ***
## bmi           1438.09      55.22   26.05   <2e-16 ***
## age            266.29      25.06   10.63   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5754 on 271 degrees of freedom
## Multiple R-squared:  0.7532, Adjusted R-squared:  0.7514
## F-statistic: 413.6 on 2 and 271 DF,  p-value: < 2.2e-16
```

```
yhat_full_smoker <- mlr_full_smoker$fitted.values
res_full_smoker <- mlr_full_smoker$residuals
```

```r
smokers %>%
  ggplot(aes(yhat_full_smoker, res_full_smoker)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```




```r
boxcox(mlr_full_smoker)
```



Rseponse needs to be transformed

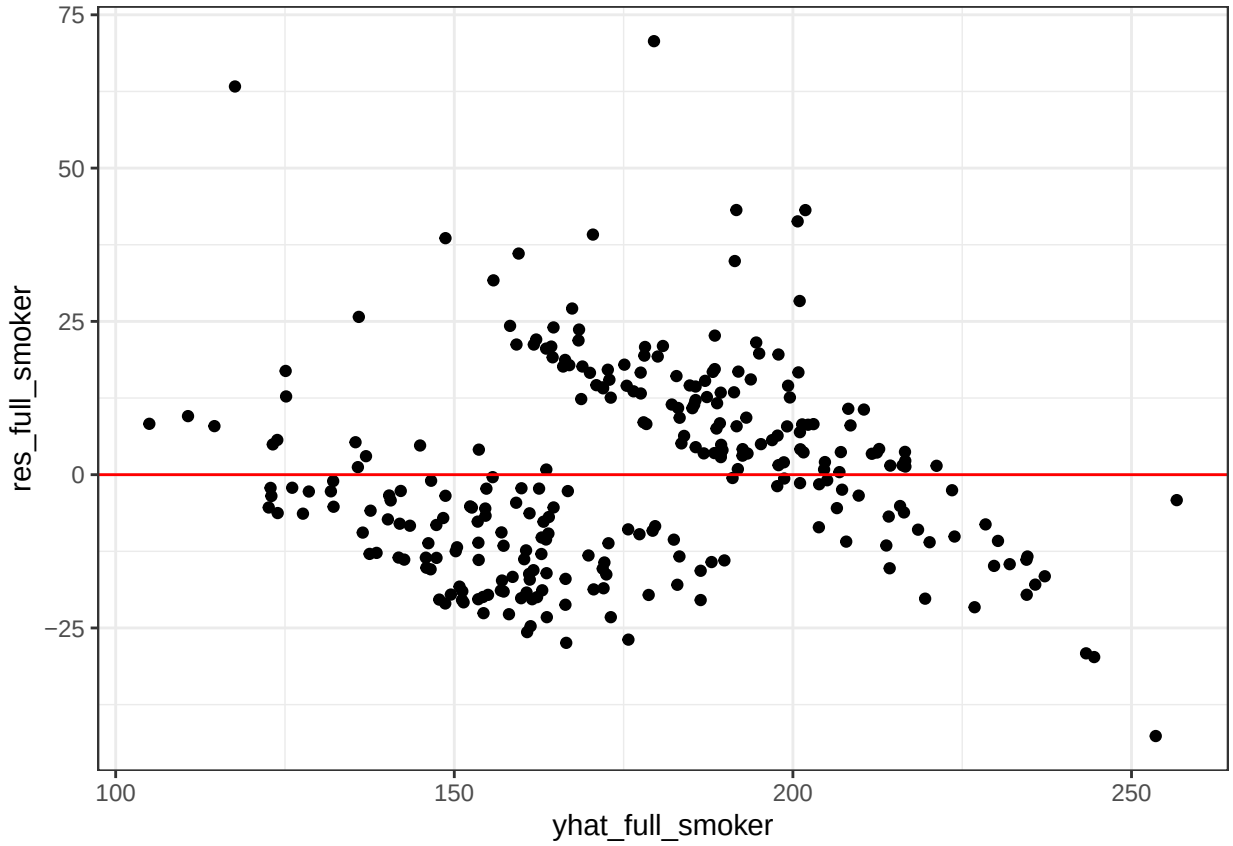
```
smokers_transform <- smokers
smokers_transform$charges <- smokers_transform$charges^0.5
mlr_full_smoker_transform = lm(charges ~  bmi+age , data=smokers_transform)
mlr_full_smoker_transform_full = lm(charges ~  bmi+age + sex + region , data=smokers_transform)

summary(mlr_full_smoker_transform)
```

```
##
## Call:
## lm(formula = charges ~ bmi + age, data = smokers_transform)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.622 -12.877  -1.715  10.868  70.699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.9145     5.4586   3.648 0.000317 ***
## bmi           4.1245     0.1560  26.436  < 2e-16 ***
## age           0.7634     0.0708  10.781  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.26 on 271 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.7569
## F-statistic: 426.1 on 2 and 271 DF,  p-value: < 2.2e-16
```

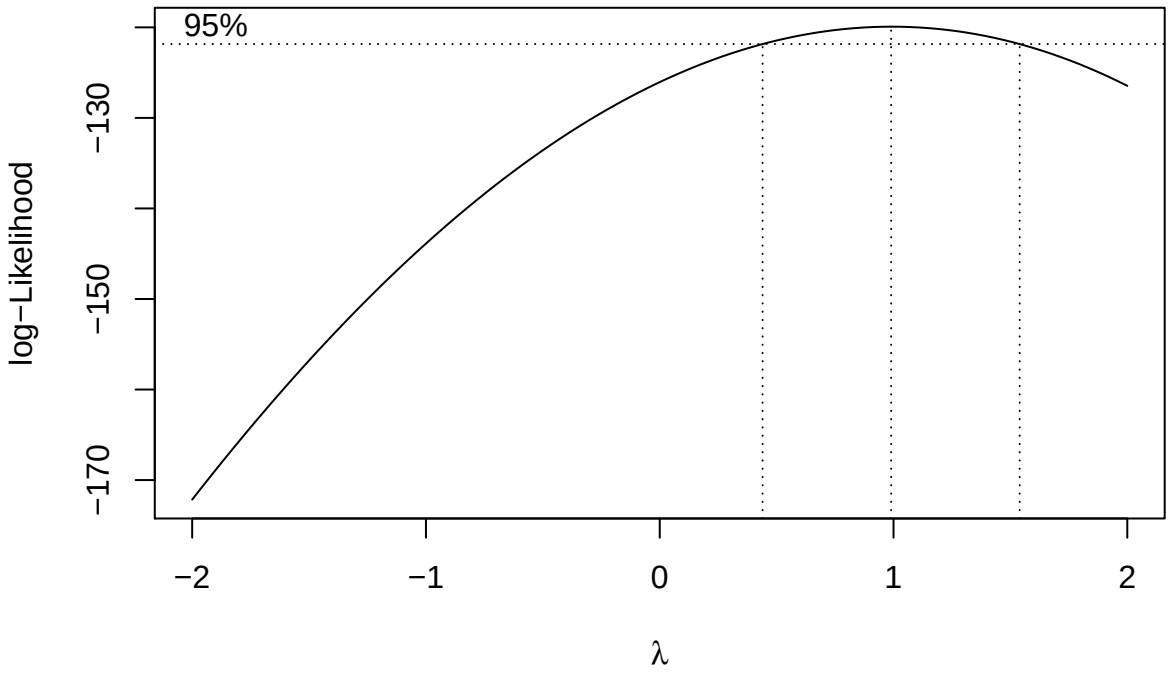
residual plot

```
yhat_full_smoker <- mlr_full_smoker_transform$fitted.values
res_full_smoker <- mlr_full_smoker_transform$residuals
smokers %>%
```

```
ggplot(aes(yhat_full_smoker, res_full_smoker)) +
geom_point() +
theme_bw() +
geom_hline(yintercept = 0, color="red")
```

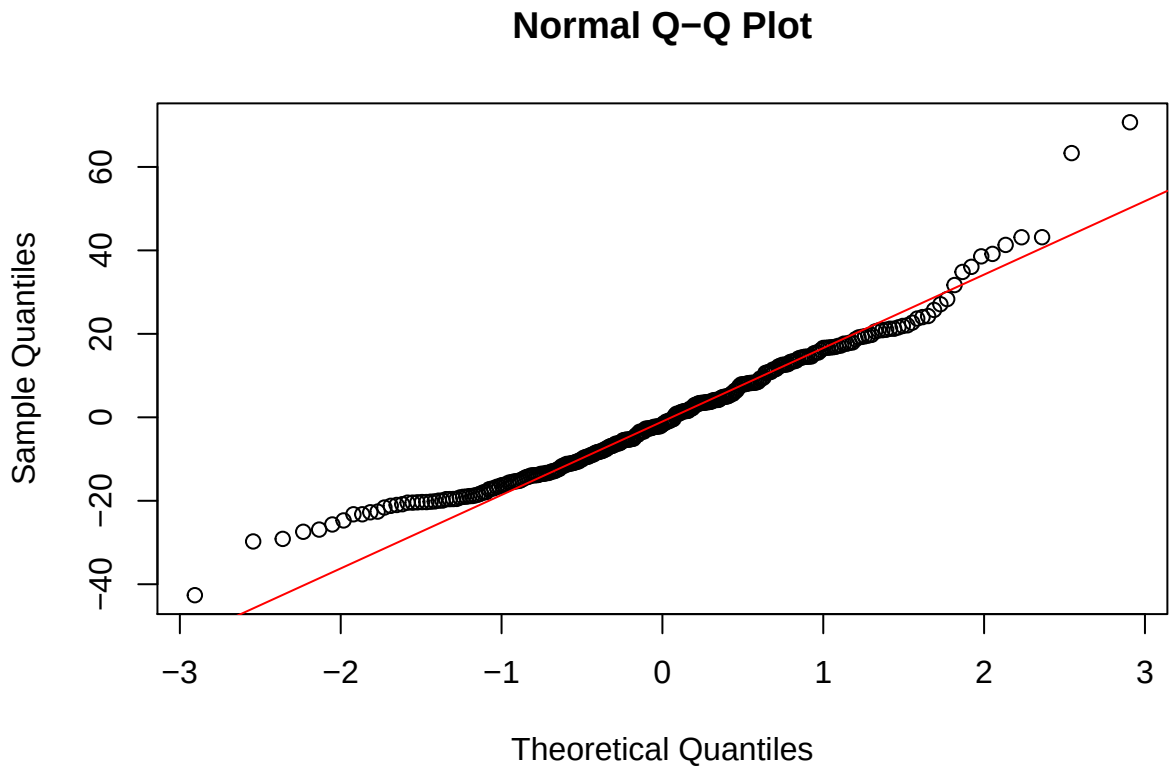



```
boxcox(mlr_full_smoker_transform)
```



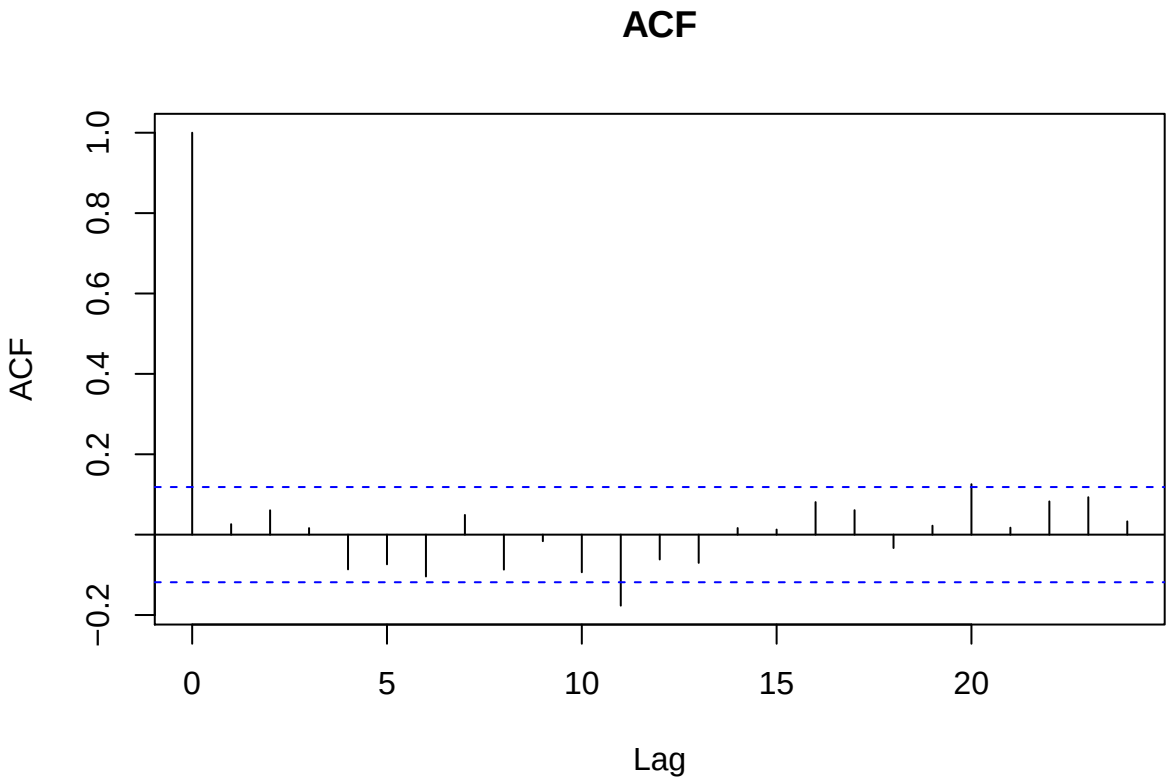
QQPlot

```
{
  qqnorm(mlr_full_smoker_transform$residuals)
  qqline(mlr_full_smoker_transform$residuals, col="red")
}
```



ACF

```r
acf(mlr_full_smoker_transform$residuals, main="ACF")
```



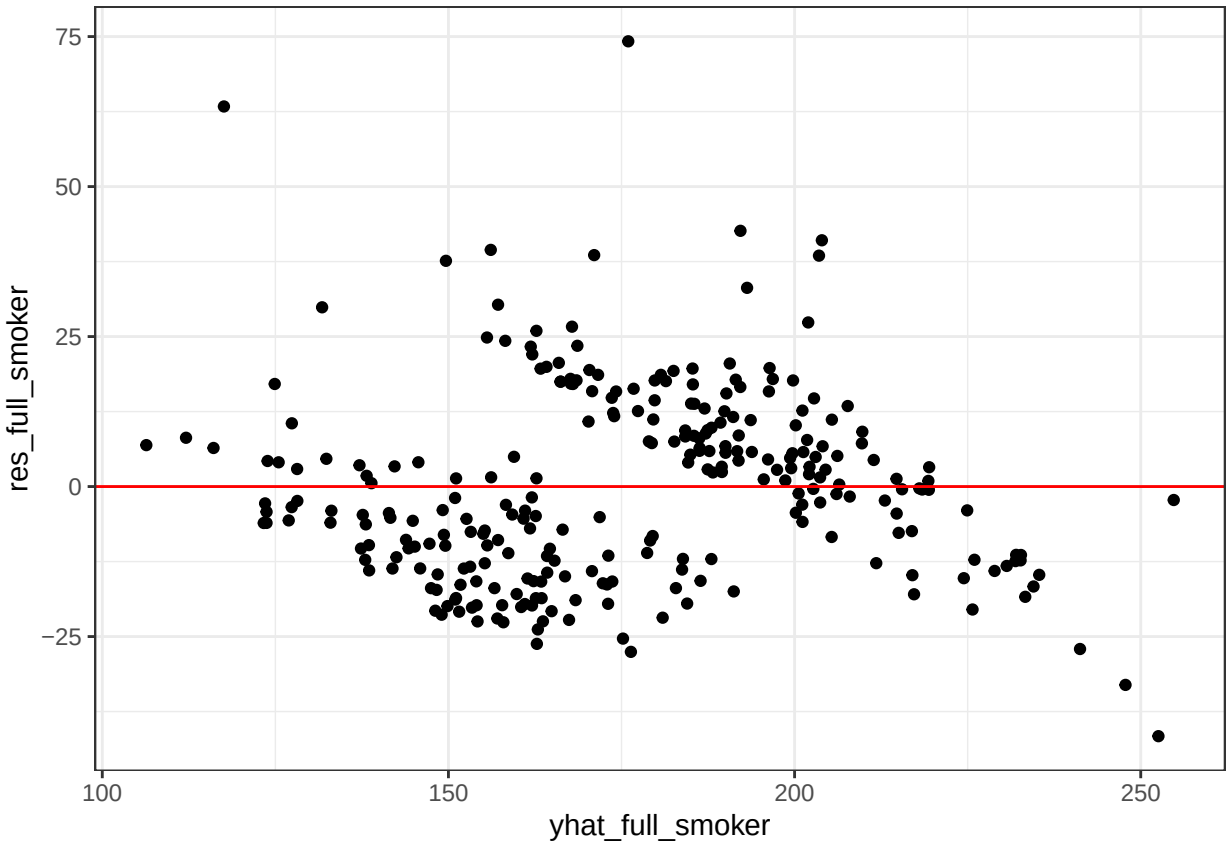
What about Region

```r
mlr_full_smoker_full = lm(charges ~  bmi+age+children+region , data=smokers_transform)
summary(mlr_full_smoker_full)
```

```
##
## Call:
## lm(formula = charges ~ bmi + age + children + region, data = smokers_transform)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -41.603 -12.339  -0.845   9.705  74.211
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     18.58918    5.66726   3.280  0.00118 **
## bmi              4.21653    0.16320  25.837  < 2e-16 ***
## age              0.76698    0.07129  10.759  < 2e-16 ***
## children         0.53142    0.85647   0.620  0.53547
## regionnorthwest -1.53030    2.92036  -0.524  0.60071
## regionsoutheast -5.14260    2.71691  -1.893  0.05946 .
## regionsouthwest -0.93879    2.94778  -0.318  0.75038
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.23 on 267 degrees of freedom
```

```
## Multiple R-squared:  0.763,  Adjusted R-squared:  0.7577
## F-statistic: 143.3 on 6 and 267 DF,  p-value: < 2.2e-16
```

```
yhat_full_smoker <- mlr_full_smoker_full$fitted.values
res_full_smoker <- mlr_full_smoker_full$residuals
smokers %>%
  ggplot(aes(yhat_full_smoker, res_full_smoker)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



Residual Seems Fine, can we drop these predictors?

```
summary(mlr_full_smoker_transform)
```

```
##
## Call:
## lm(formula = charges ~ bmi + age, data = smokers_transform)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -42.622 -12.877  -1.715  10.868  70.699
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.9145     5.4586   3.648 0.000317 ***
## bmi           4.1245     0.1560  26.436  < 2e-16 ***
## age           0.7634     0.0708  10.781  < 2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.26 on 271 degrees of freedom
## Multiple R-squared:  0.7587, Adjusted R-squared:  0.7569
## F-statistic: 426.1 on 2 and 271 DF,  p-value: < 2.2e-16
```

```
anova(mlr_full_smoker_transform,mlr_full_smoker_transform_full )
```

```
## Analysis of Variance Table
##
## Model 1: charges ~ bmi + age
## Model 2: charges ~ bmi + age + sex + region
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1    271 71645
## 2    267 70442  4    1202.7 1.1396 0.3382
```

We can drop these other predictors

Outliers

```
##critical value using Bonferroni procedure
n<-dim(smokers_transform)[1]
p<-3
crit<-qt(1-0.05/(2*n), n-1-p)
##externally studentized residuals
ext.student.res<-rstudent(mlr_full_smoker_transform)
##identify
ext.student.res[abs(ext.student.res)>crit]
```

```
##      129     1301
## 4.041989 4.510779
```

```
##leverages
lev<-lm.influence(mlr_full_smoker_transform)$hat
##identify
round(lev[lev>2*p/n],4)
```

```
##    251    293    413    544    550    665    804    861   1048   1125   1157
## 0.0268 0.0284 0.0226 0.0326 0.0258 0.0224 0.0251 0.0301 0.0547 0.0225 0.0307
```

```
DFFITS<-dffits(mlr_full_smoker_transform)
round(DFFITS[abs(DFFITS)>2*sqrt(p/n)],3)
```

```
##    129    293    550    578    820    861    918   1048   1157   1231   1301
##  0.569 -0.231 -0.297  0.260  0.218 -0.329  0.246 -0.656 -0.225  0.244  0.302
```

```
DFBETAS<-dfbetas(mlr_full_smoker_transform)
tempdfbetas = abs(DFBETAS)>2/sqrt(n)
tempdfbetas[(tempdfbetas[,1] == TRUE | tempdfbetas[,2] == TRUE | tempdfbetas[,3] == TRUE ) ,]
```

```
##      (Intercept)   bmi   age
## 35         FALSE  TRUE FALSE
## 95         FALSE FALSE  TRUE
## 129         TRUE  TRUE FALSE
## 293        FALSE  TRUE FALSE
## 477        FALSE FALSE  TRUE
## 531         TRUE FALSE FALSE
## 550         TRUE  TRUE FALSE
```

```
## 578        FALSE  TRUE FALSE
## 675         TRUE  TRUE FALSE
## 820        FALSE  TRUE FALSE
## 861         TRUE  TRUE FALSE
## 918         TRUE  TRUE FALSE
## 952         TRUE  TRUE FALSE
## 1048        TRUE  TRUE  TRUE
## 1140       FALSE FALSE  TRUE
## 1147       FALSE FALSE  TRUE
## 1157       FALSE  TRUE FALSE
## 1197       FALSE FALSE  TRUE
## 1224        TRUE FALSE  TRUE
## 1231        TRUE FALSE  TRUE
## 1232        TRUE FALSE FALSE
## 1301       FALSE FALSE  TRUE
## 1302       FALSE FALSE  TRUE
```

```
COOKS<-cooks.distance(mlr_full_smoker_transform)
COOKS[COOKS>qf(0.5,p,n-p)]
```

```
## named numeric(0)
```

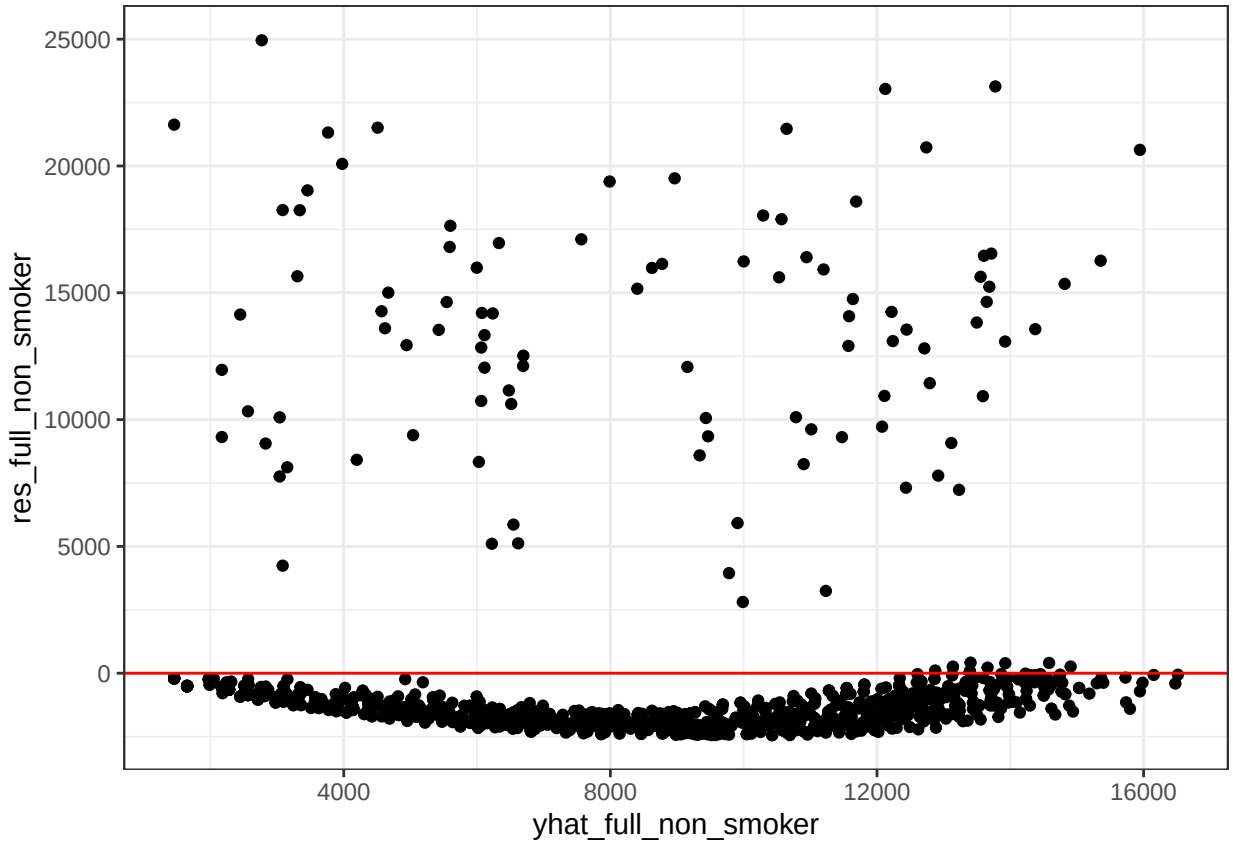
Non Smokers

```
mlr_full_non_smoker = lm(formula = charges ~ age + children + region + sex, data = non_smokers)
summary(mlr_full_non_smoker)
```

```
##
## Call:
## lm(formula = charges ~ age + children + region + sex, data = non_smokers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2441.2 -1870.1 -1380.6  -673.9 24954.7
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1695.87     520.71  -3.257 0.001162 **
## age               265.53      10.01  26.524  < 2e-16 ***
## children          589.06     115.67   5.093 4.18e-07 ***
## regionnorthwest  -550.17     401.17  -1.371 0.170544
## regionsoutheast  -913.18     398.99  -2.289 0.022293 *
## regionsouthwest -1372.97     401.23  -3.422 0.000646 ***
## sexmale          -521.01     281.62  -1.850 0.064585 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4589 on 1057 degrees of freedom
## Multiple R-squared:  0.417,  Adjusted R-squared:  0.4137
## F-statistic:   126 on 6 and 1057 DF,  p-value: < 2.2e-16
```

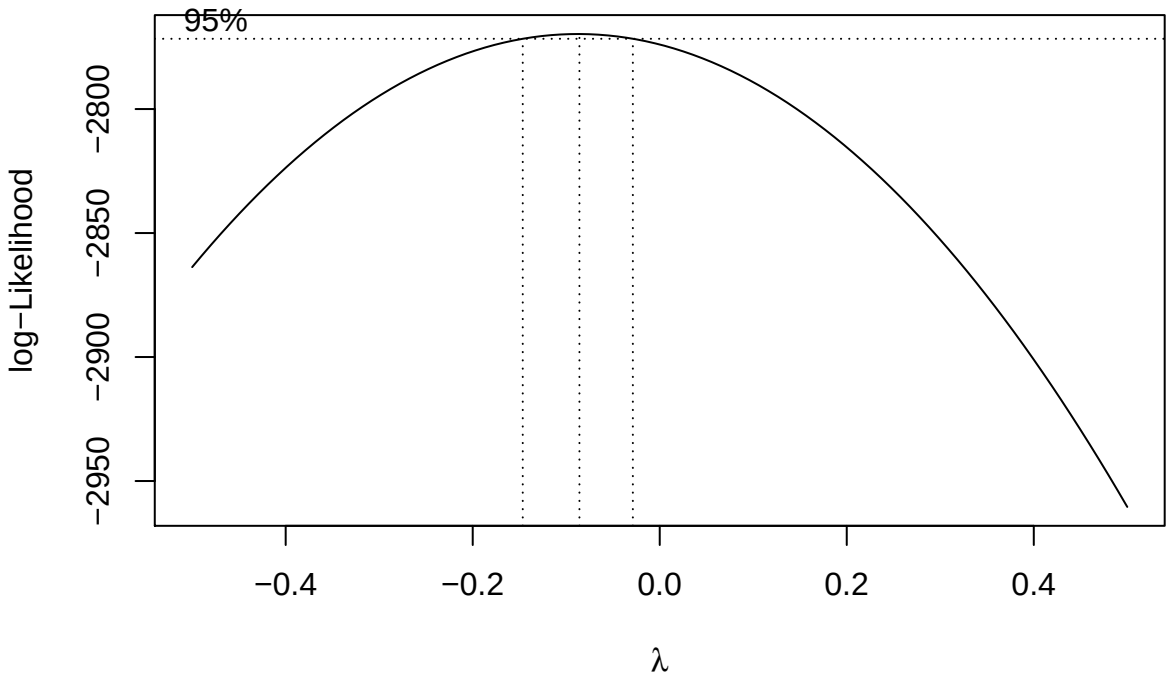
```
yhat_full_non_smoker <- mlr_full_non_smoker$fitted.values
res_full_non_smoker <- mlr_full_non_smoker$residuals
non_smokers %>%
  ggplot(aes(yhat_full_non_smoker, res_full_non_smoker)) +
```

```
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



Transformation

```
boxcox(mlr_full_non_smoker, c(-0.5, 0.5, 0.1))
```

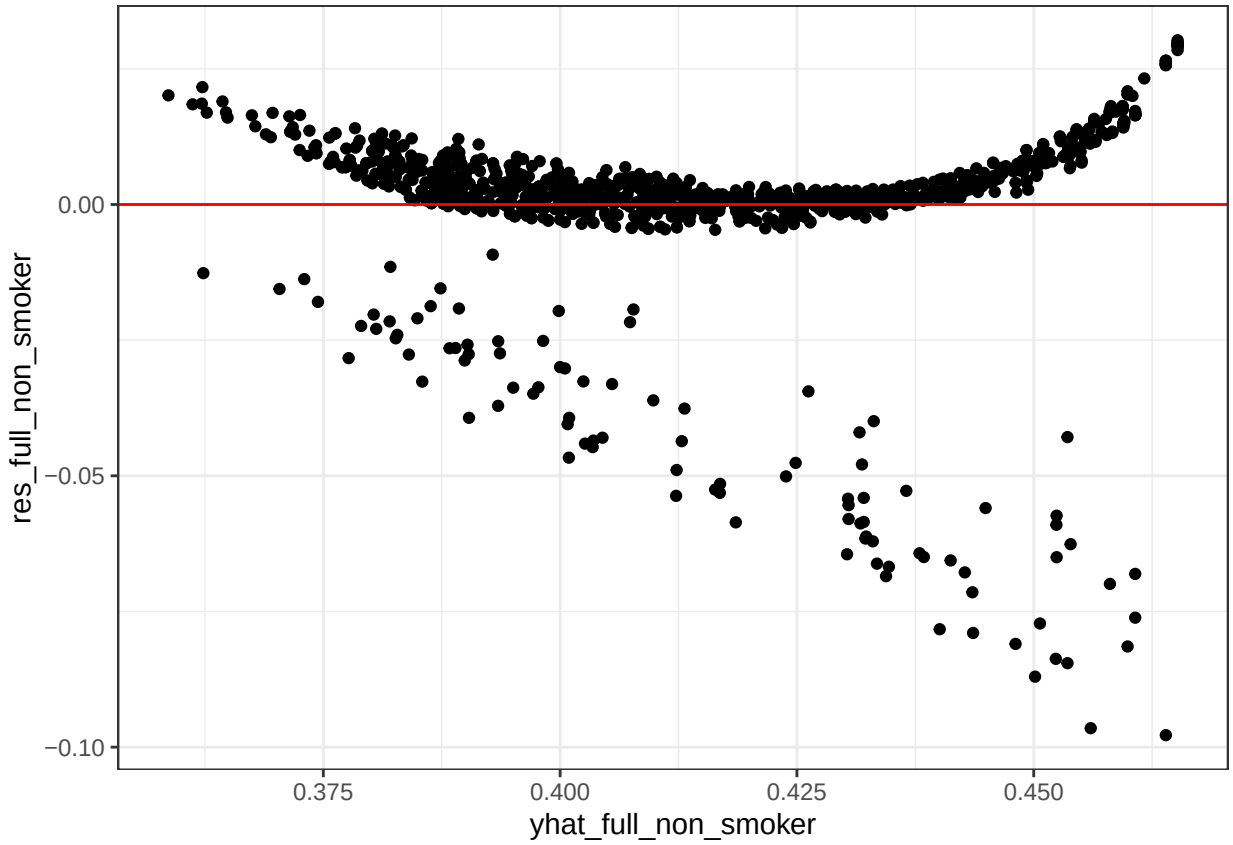



```
non_smoker_transform <- non_smokers
non_smoker_transform$charges <- non_smoker_transform$charges^(-0.1)
mlr_full_non_smoker_transform = lm(formula = charges ~ age + children + region + sex, data = non_smoker_
summary(mlr_full_non_smoker_transform)
```

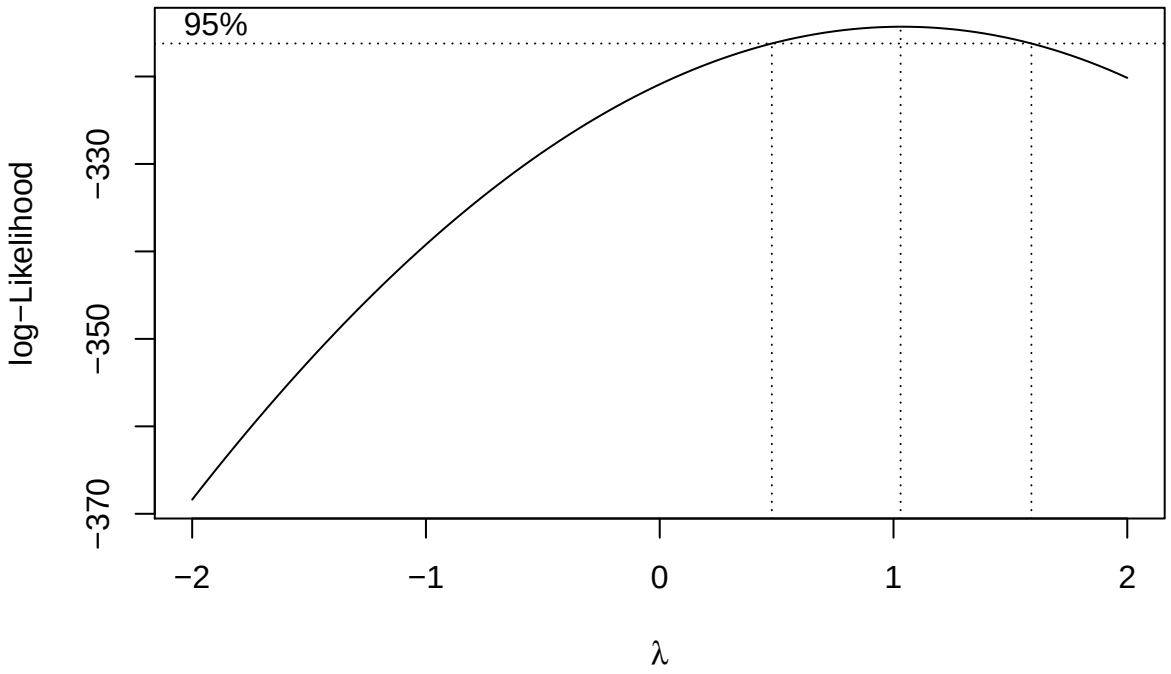
```
##
## Call:
## lm(formula = charges ~ age + children + region + sex, data = non_smoker_transform)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.097790 -0.000251  0.002357  0.006879  0.030263
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4.853e-01  1.947e-03 249.214  < 2e-16 ***
## age             -1.762e-03  3.744e-05 -47.072  < 2e-16 ***
## children        -5.651e-03  4.326e-04 -13.065  < 2e-16 ***
## regionnorthwest  3.100e-03  1.500e-03   2.066   0.0391 *
## regionsoutheast  7.149e-03  1.492e-03   4.792 1.89e-06 ***
## regionsouthwest  7.667e-03  1.500e-03   5.110 3.82e-07 ***
## sexmale          4.482e-03  1.053e-03   4.256 2.27e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01716 on 1057 degrees of freedom
## Multiple R-squared:  0.7023, Adjusted R-squared:  0.7006
## F-statistic: 415.6 on 6 and 1057 DF,  p-value: < 2.2e-16
```

```
yhat_full_non_smoker <- mlr_full_non_smoker_transform$fitted.values
res_full_non_smoker <- mlr_full_non_smoker_transform$residuals
non_smokers %>%
  ggplot(aes(yhat_full_non_smoker, res_full_non_smoker)) +
```

```
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```

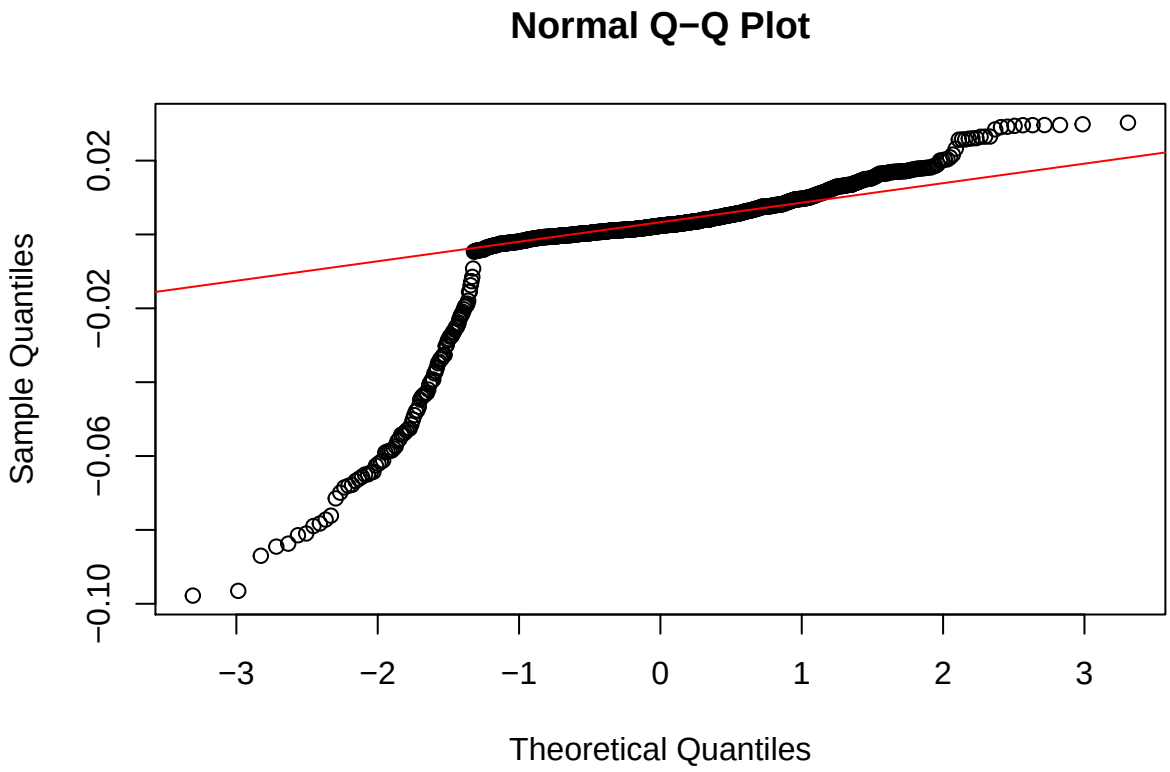



```
boxcox(mlr_full_non_smoker_transform)
```



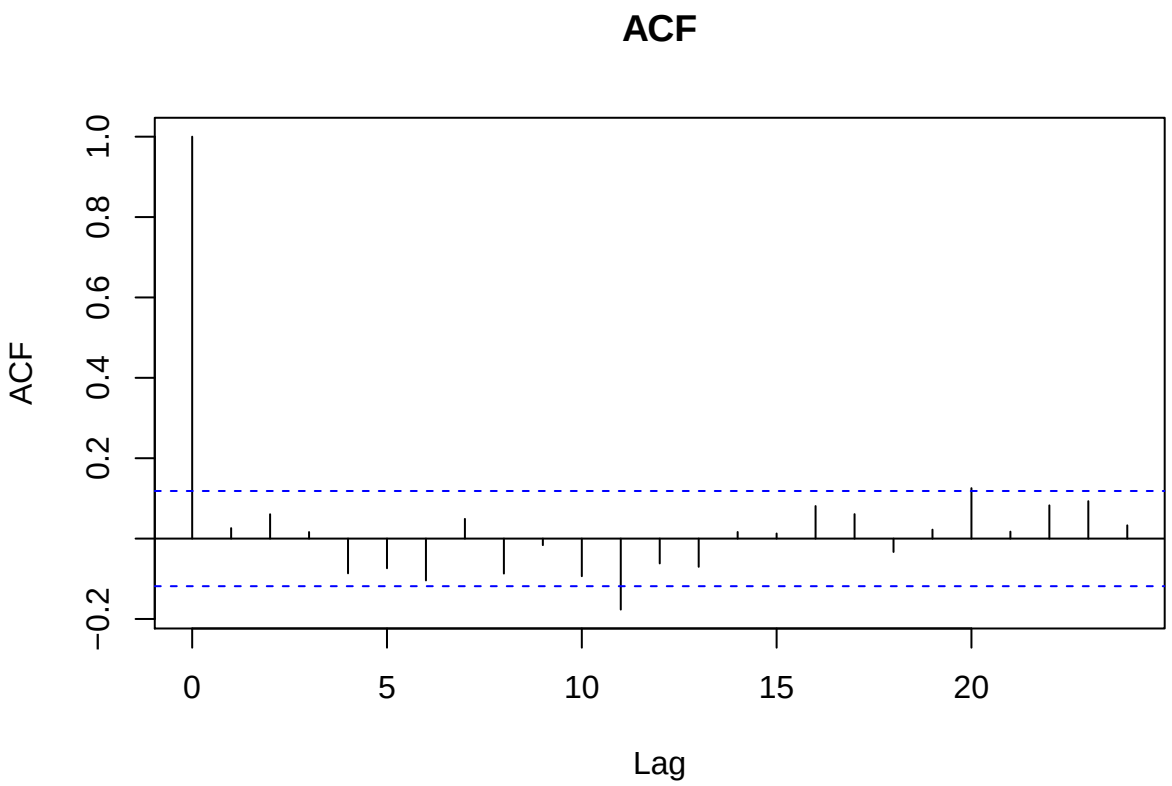
QQPlot

```
{
  qqnorm(mlr_full_non_smoker_transform$residuals)
  qqline(mlr_full_non_smoker_transform$residuals, col="red")
}
```



ACF

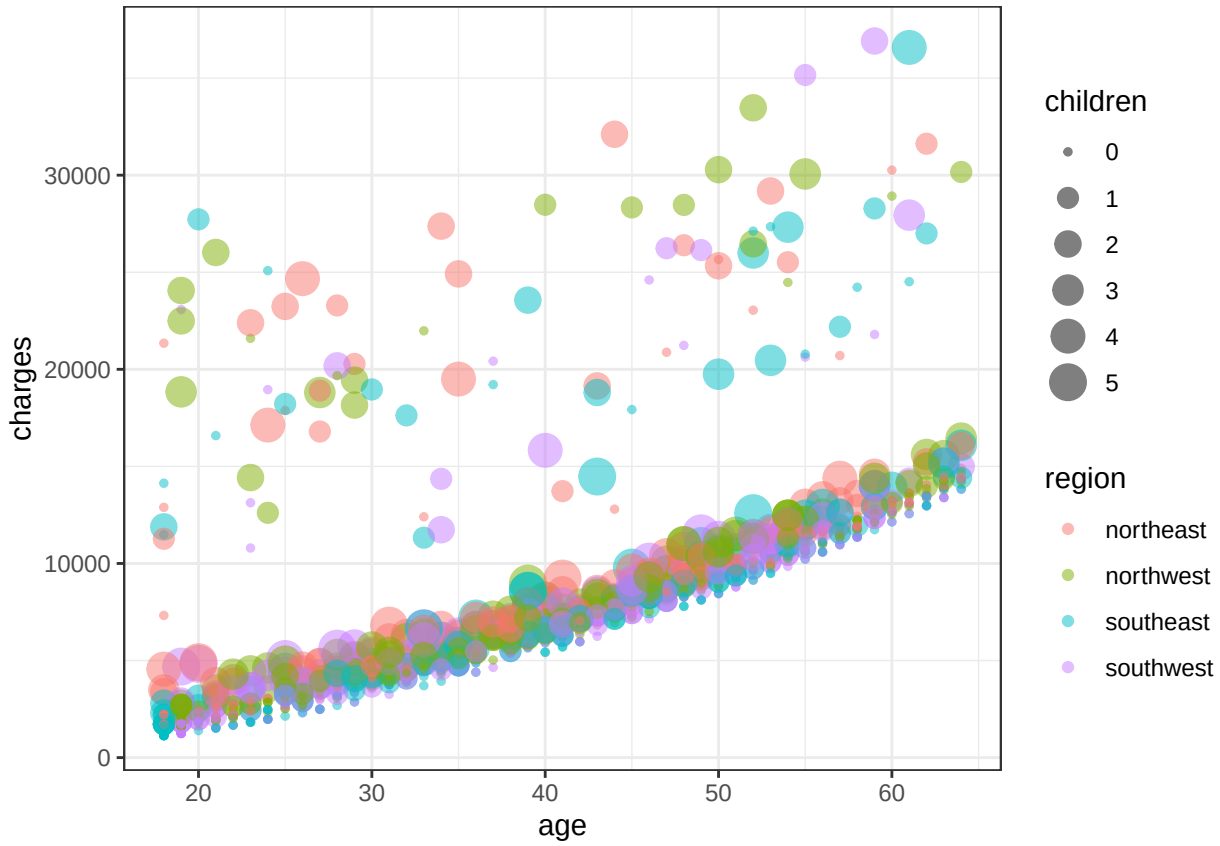
```
acf(mlr_full_smoker_transform$residuals, main="ACF")
```




```
head(non_smokers)
```

```
##   age    sex    bmi children smoker    region   charges significant.charge
## 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
## 7  46 female 33.440        1     no southeast  8240.590              FALSE
```

```
# charges ~ age + children + region + sex
non_smokers %>%
  ggplot(aes(x=age, y=charges, color=region, size=children)) +
  theme_bw() +
  geom_point(alpha=0.5)
```



Lets drop these outliers.

```
drop_outleirs <- non_smokers
DFFITS<-dffits(mlr_full_non_smoker)
want_drop = names(DFFITS[abs(DFFITS)>2*sqrt(5/1064)])
drop_outleirs = non_smokers[setdiff(rownames(non_smokers), want_drop),]
(dim(non_smokers)[1] - dim(drop_outleirs)[1]) / dim(non_smokers)[1]
```

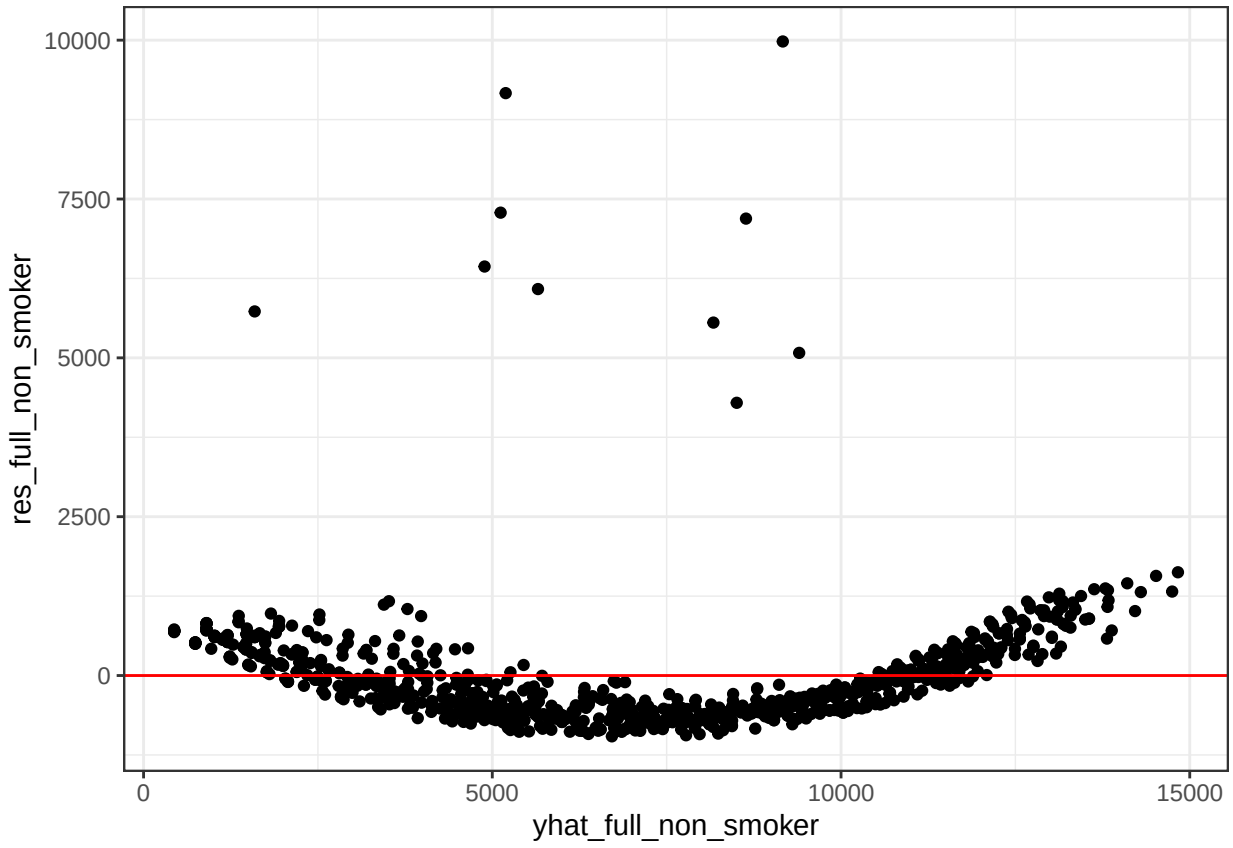
```
## [1] 0.08364662
```

```
mlr_full_non_smoker_transform_drop = lm(formula = charges ~  region + age + children + sex, data = drop_
summary(mlr_full_non_smoker_transform_drop)
```

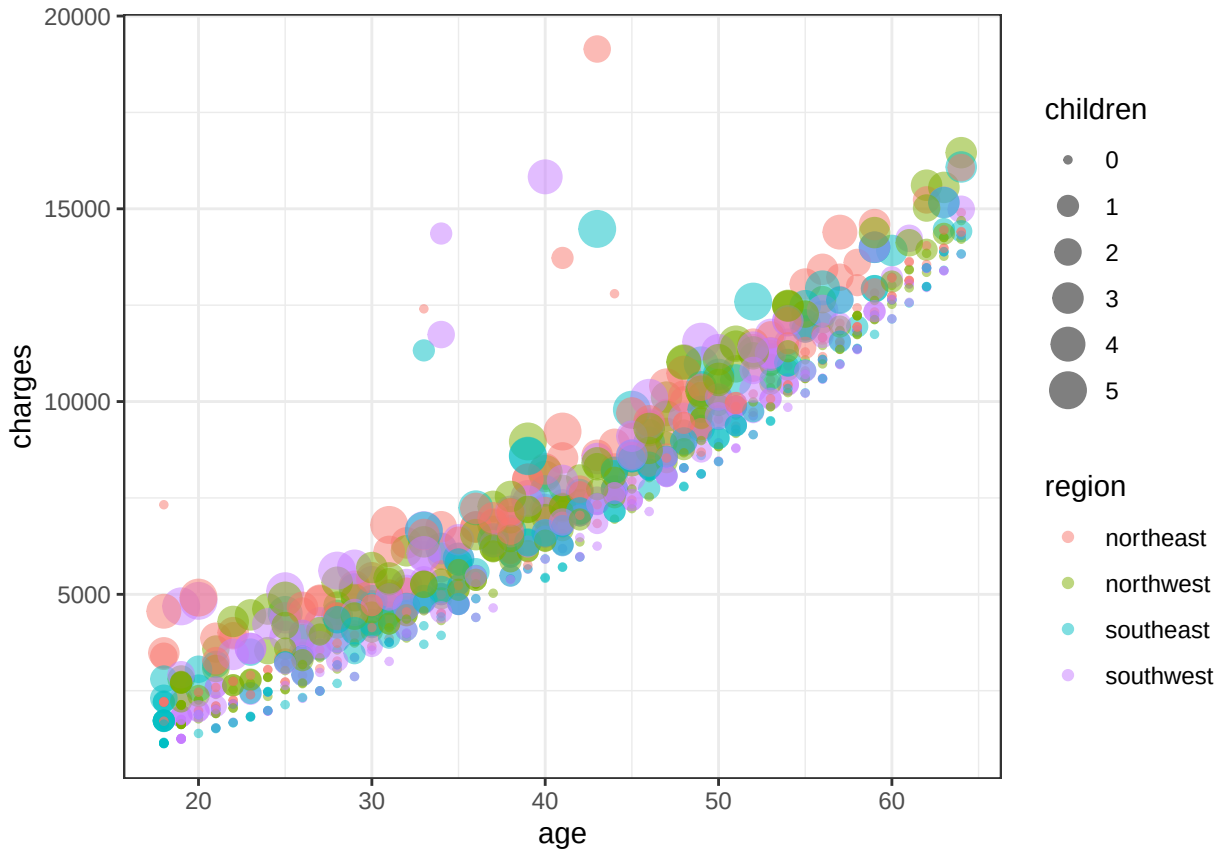
```
##
## Call:
## lm(formula = charges ~ region + age + children + sex, data = drop_outleirs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -957.1 -528.9 -157.0  377.4 9979.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -3191.858    106.474 -29.978  < 2e-16 ***
## regionnorthwest  -380.203     81.624  -4.658 3.64e-06 ***
## regionsoutheast  -693.456     81.505  -8.508  < 2e-16 ***
## regionsouthwest  -656.634     81.119  -8.095 1.71e-15 ***
```

```
## age               265.832      2.041 130.242  < 2e-16 ***
## children          463.159     23.342  19.842  < 2e-16 ***
## sexmale          -461.038     57.043  -8.082 1.88e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 889.8 on 968 degrees of freedom
## Multiple R-squared:  0.9486, Adjusted R-squared:  0.9483
## F-statistic:  2979 on 6 and 968 DF,  p-value: < 2.2e-16
```

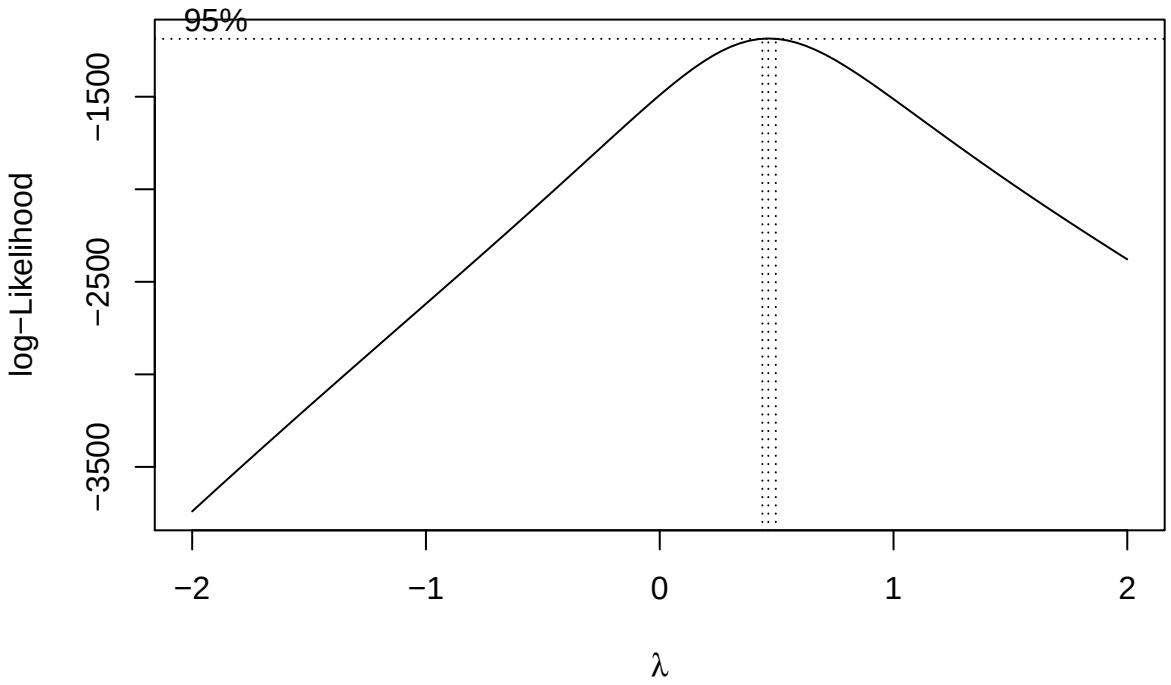
```
yhat_full_non_smoker <- mlr_full_non_smoker_transform_drop$fitted.values
res_full_non_smoker <- mlr_full_non_smoker_transform_drop$residuals
drop_outleirs %>%
  ggplot(aes(yhat_full_non_smoker, res_full_non_smoker)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```




```
drop_outleirs %>%
  ggplot(aes(x=age, y=charges, color=region, size=children)) +
  theme_bw() +
  geom_point(alpha=0.5)
```



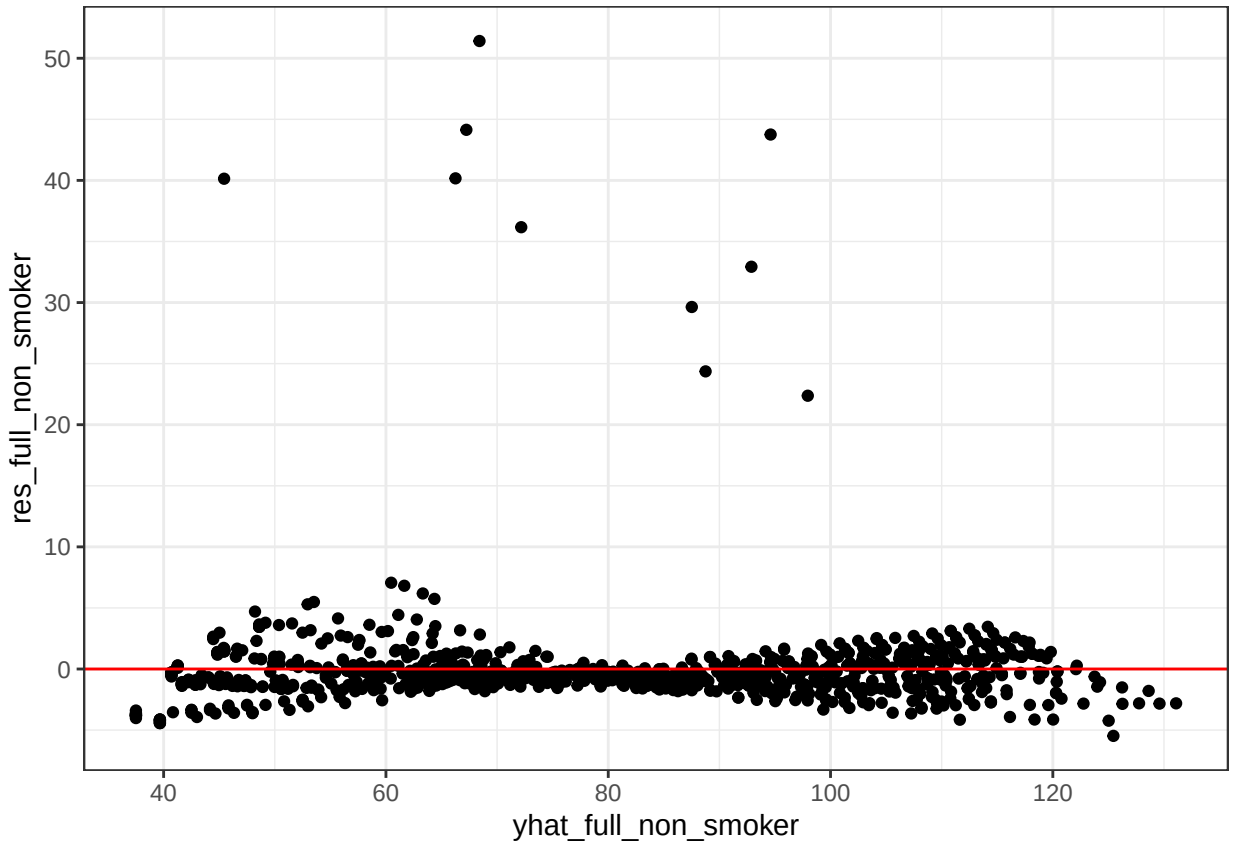

```
boxcox(mlr_full_non_smoker_transform_drop)
```




```
mlr_full_non_smoker_transform_drop2 = lm(formula = (charges)^0.5 ~ age + children + region + sex, data =
summary(mlr_full_non_smoker_transform_drop2)
```

```
##
## Call:
## lm(formula = (charges)^0.5 ~ age + children + region + sex, data = drop_outleirs)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -5.473 -1.294 -0.488  0.603 51.411
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     15.450853   0.497804  31.038  < 2e-16 ***
## age              1.666084   0.009543 174.593  < 2e-16 ***
## children         3.758067   0.109132  34.436  < 2e-16 ***
## regionnorthwest -2.264117   0.381620  -5.933 4.14e-09 ***
## regionsoutheast -4.734523   0.381067 -12.424  < 2e-16 ***
## regionsouthwest -4.244789   0.379261 -11.192  < 2e-16 ***
## sexmale         -3.195678   0.266695 -11.983  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.16 on 968 degrees of freedom
## Multiple R-squared:  0.9713, Adjusted R-squared:  0.9711
## F-statistic:  5462 on 6 and 968 DF,  p-value: < 2.2e-16
```

```
yhat_full_non_smoker <- mlr_full_non_smoker_transform_drop2$fitted.values
res_full_non_smoker <- mlr_full_non_smoker_transform_drop2$residuals
drop_outleirs %>%
  ggplot(aes(yhat_full_non_smoker, res_full_non_smoker)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



Second dropping

```
drop_outleirs2 <- drop_outleirs
DFFITS<-dffits(mlr_full_non_smoker_transform_drop2)
want_drop = names(DFFITS[abs(DFFITS)>2*sqrt(5/975)])
drop_outleirs2 = drop_outleirs[setdiff(rownames(drop_outleirs), want_drop),]
(dim(drop_outleirs)[1] - dim(drop_outleirs2)[1]) / dim(drop_outleirs)[1]
```

```
## [1] 0.01538462
```

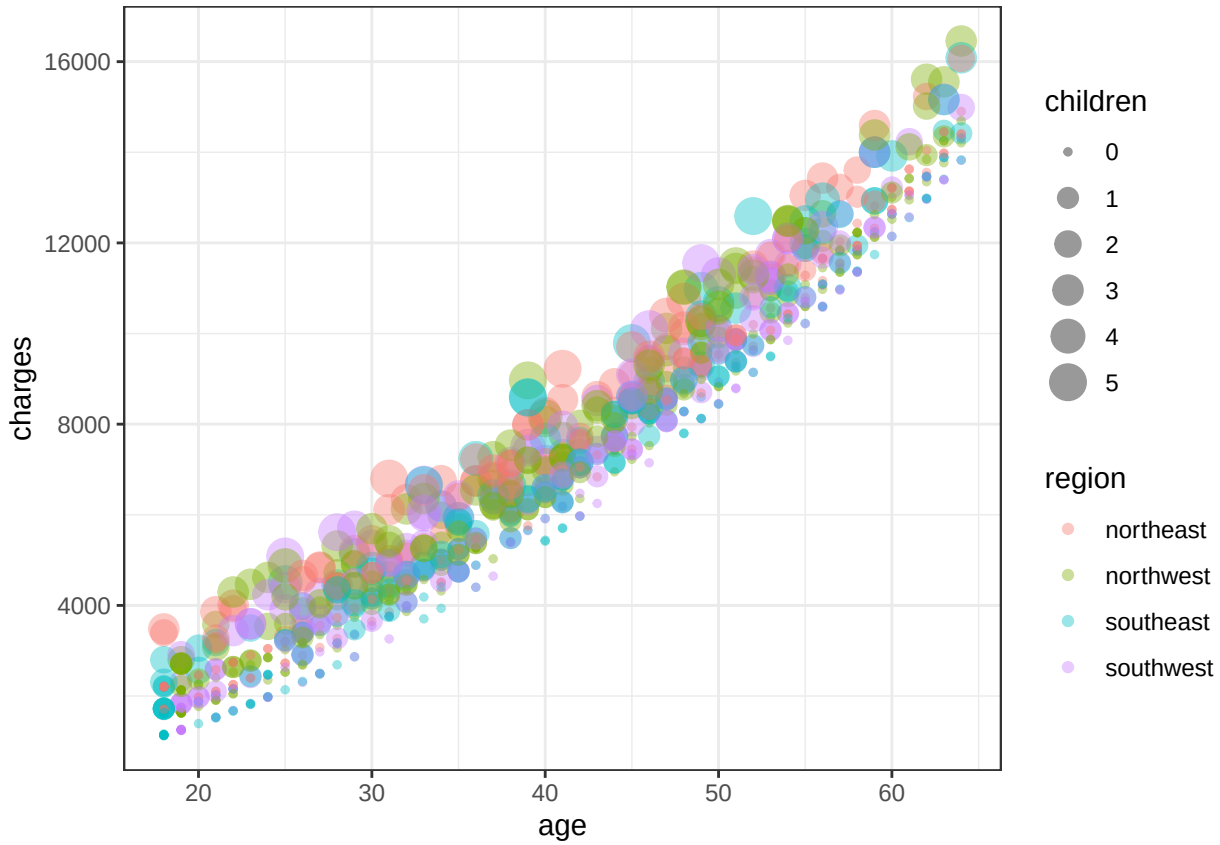
```
(dim(non_smokers)[1] - dim(drop_outleirs2)[1]) / dim(non_smokers)[1]
```

```
## [1] 0.09774436
```

```
head(drop_outleirs2)
```

```
##   age    sex   bmi children smoker    region  charges significant.charge
## 2  18   male 33.77        1     no southeast 1725.552              FALSE
## 3  28   male 33.00        3     no southeast 4449.462              FALSE
## 5  32   male 28.88        0     no northwest 3866.855              FALSE
## 6  31 female 25.74        0     no southeast 3756.622              FALSE
## 7  46 female 33.44        1     no southeast 8240.590              FALSE
## 8  37 female 27.74        3     no northwest 7281.506              FALSE
```

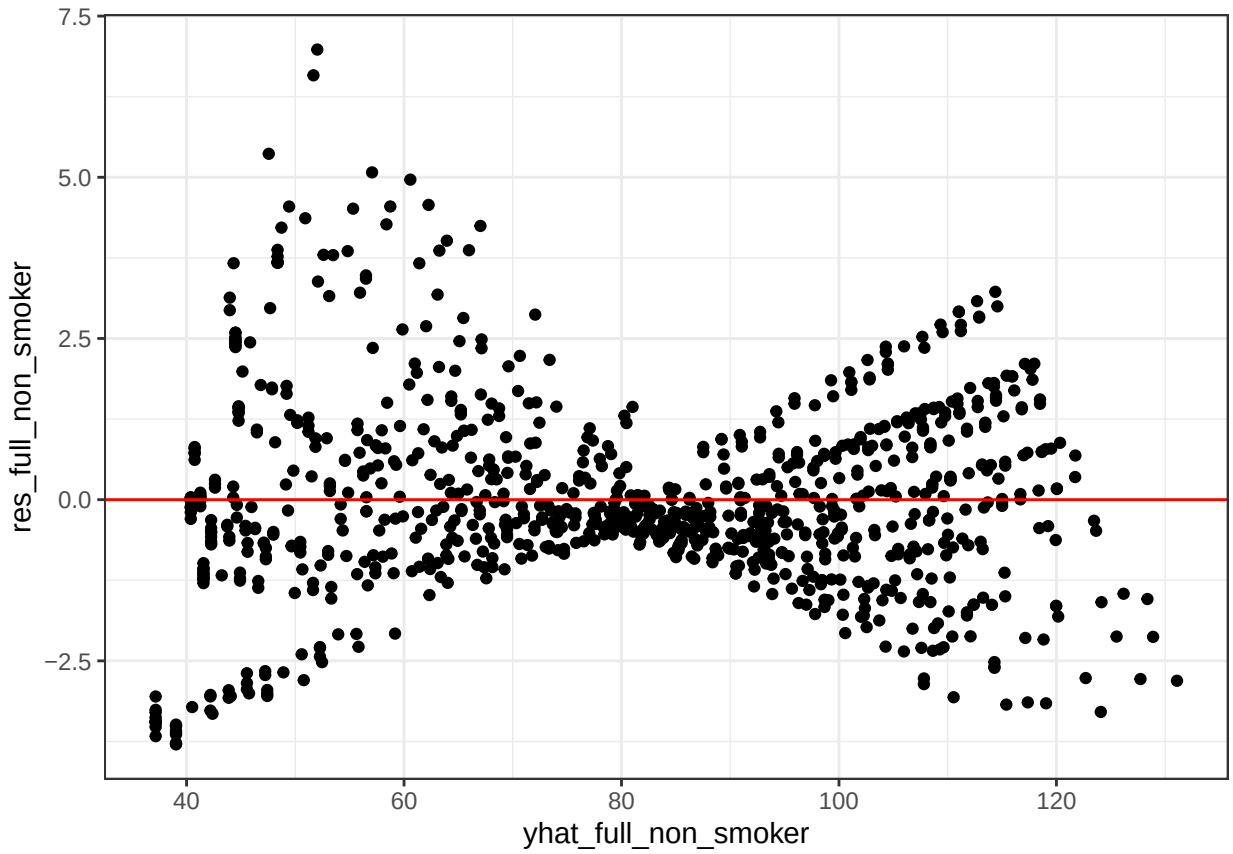
```
drop_outleirs2 %>%
  ggplot(aes(x=age, y=charges, color=region, size=children)) +
  geom_point(alpha = 0.4) + theme_bw()
```



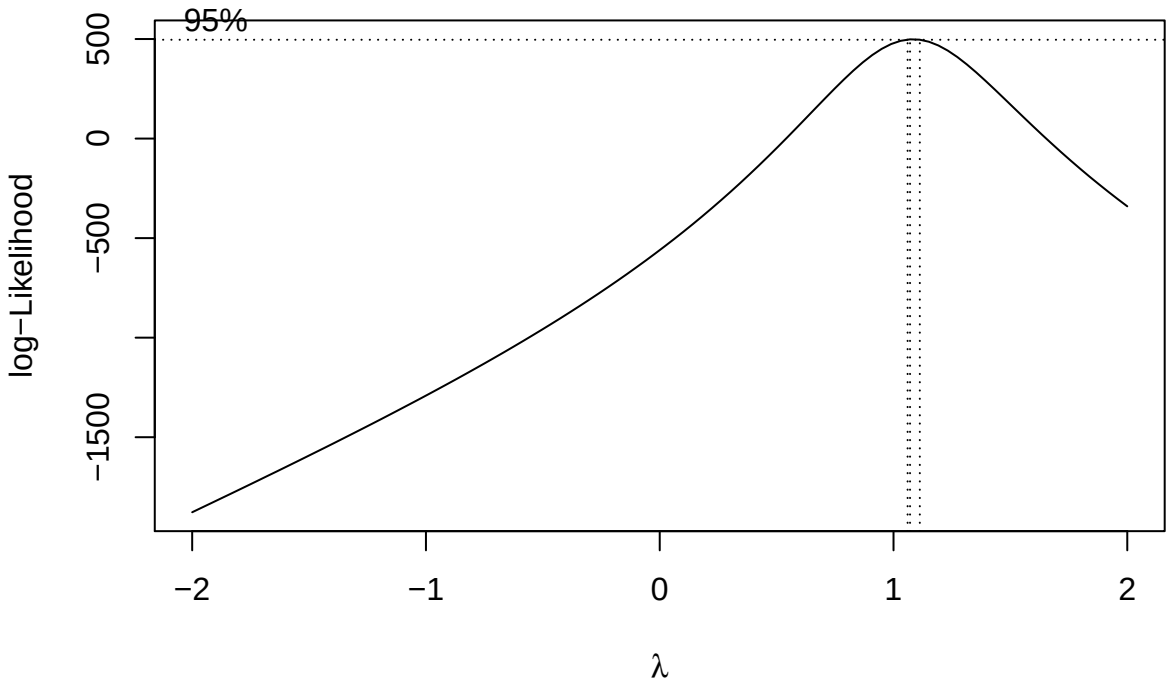

```
mlr_full_non_smoker_transform_drop3 = lm(formula = (charges)^0.5 ~  region + age + children + sex, data
summary(mlr_full_non_smoker_transform_drop3)
```

```
##
## Call:
## lm(formula = (charges)^0.5 ~ region + age + children + sex, data = drop_outleirs2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.7948 -0.7606 -0.1970  0.8076  6.9809
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     14.284035   0.181588   78.66   <2e-16 ***
## regionnorthwest -1.388811   0.138855  -10.00   <2e-16 ***
## regionsoutheast -4.108629   0.138912  -29.58   <2e-16 ***
## regionsouthwest -3.918607   0.138610  -28.27   <2e-16 ***
## age              1.678697   0.003466  484.30   <2e-16 ***
## children         3.585316   0.040673   88.15   <2e-16 ***
## sexmale         -3.230049   0.096866  -33.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.499 on 953 degrees of freedom
## Multiple R-squared:  0.9962, Adjusted R-squared:  0.9962
## F-statistic: 4.199e+04 on 6 and 953 DF,  p-value: < 2.2e-16
```

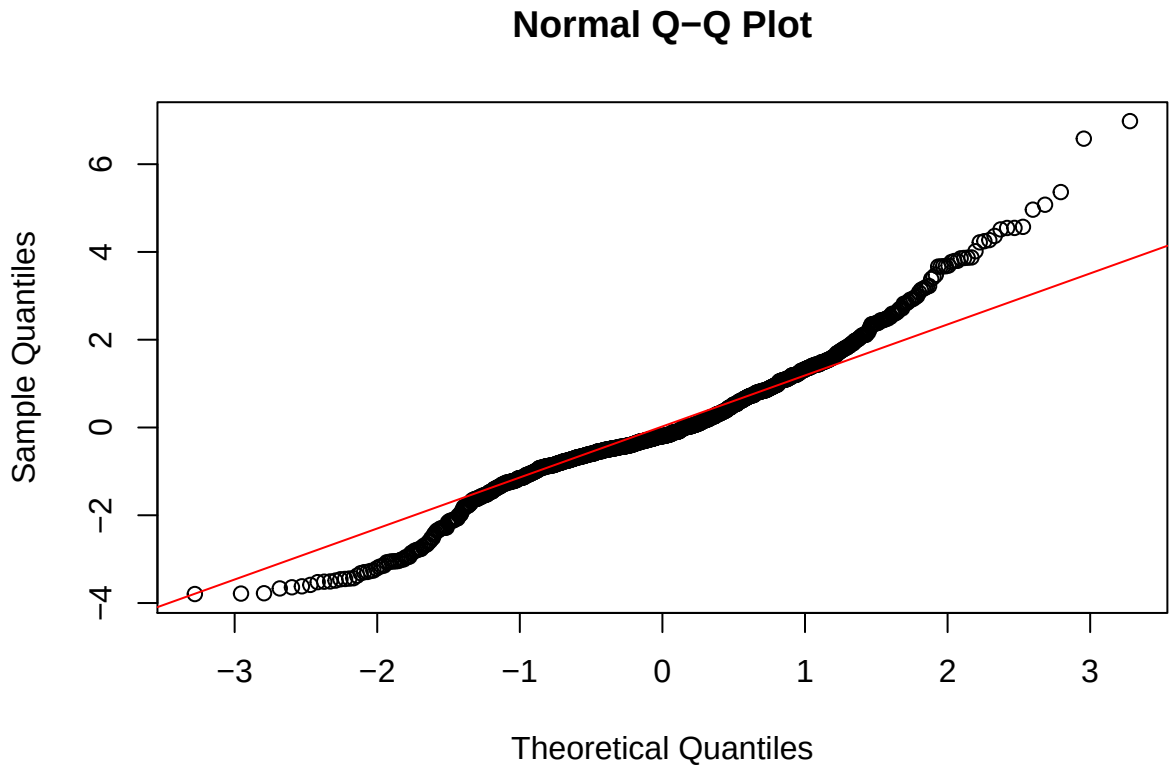
```
yhat_full_non_smoker <- mlr_full_non_smoker_transform_drop3$fitted.values
res_full_non_smoker <- mlr_full_non_smoker_transform_drop3$residuals
drop_outleirs2 %>%
  ggplot(aes(yhat_full_non_smoker, res_full_non_smoker)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```




```
boxcox(mlr_full_non_smoker_transform_drop3)
```



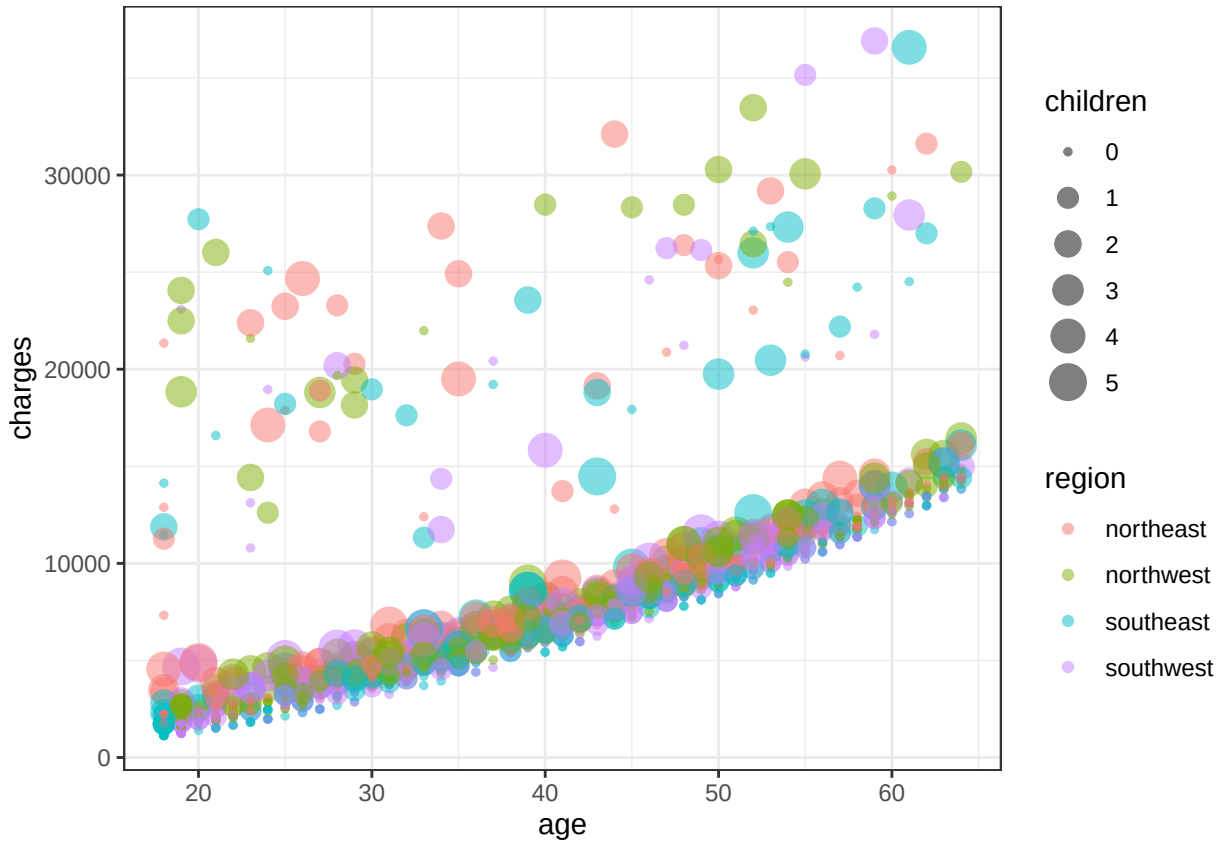

```
{
  qqnorm(mlr_full_non_smoker_transform_drop3$residuals)
  qqline(mlr_full_non_smoker_transform_drop3$residuals, col="red")
}
```



Next Steps

We need to find for non smokers what is causing some data points to be more charges. . . ?

```
non_smokers %>%
  ggplot(aes(x=age, y=charges, size=children, color=region)) +
  theme_bw() +
  geom_point(alpha=0.5)
```



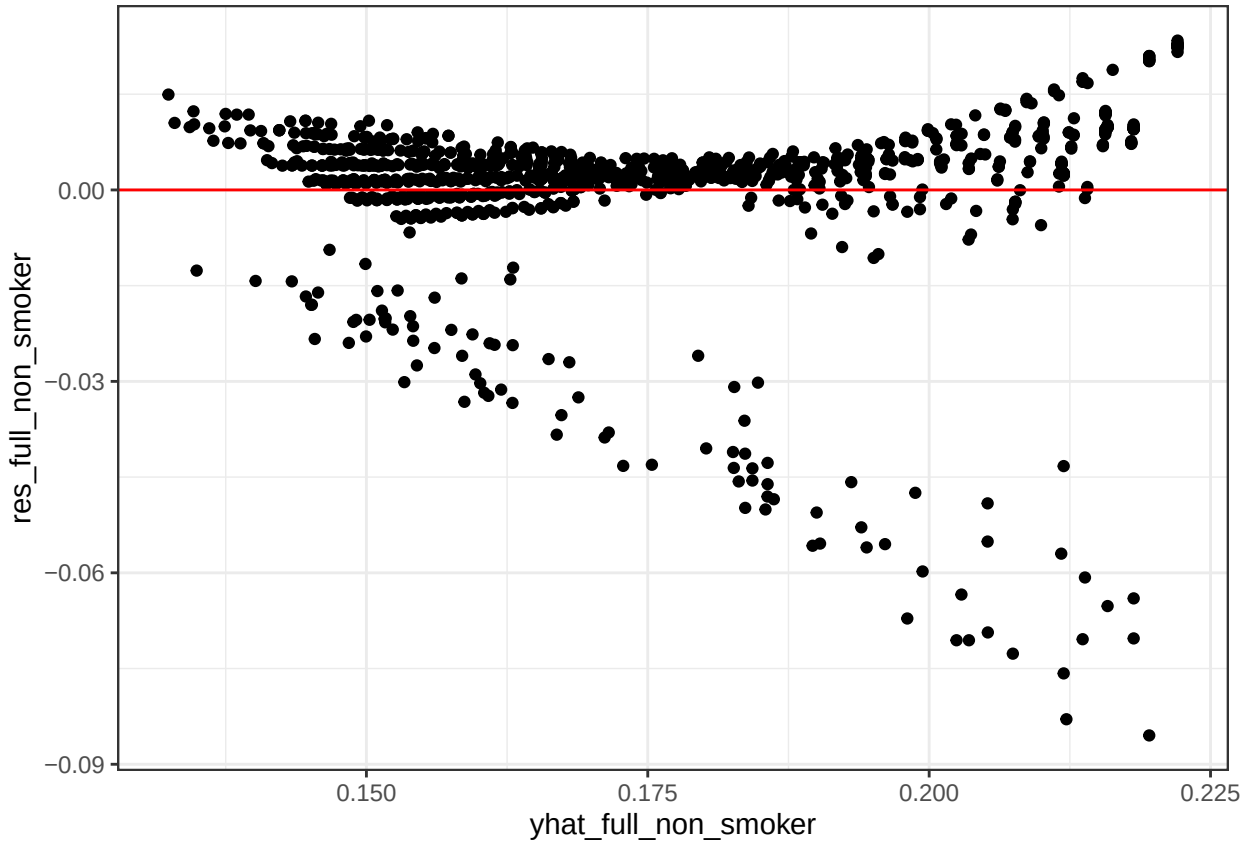
Try checking log of age?

```
mlr_full_non_smoker_log_age = lm(formula = (charges)^(-0.2) ~ log(age) + children + region + sex, data =
summary(mlr_full_non_smoker_log_age)
```

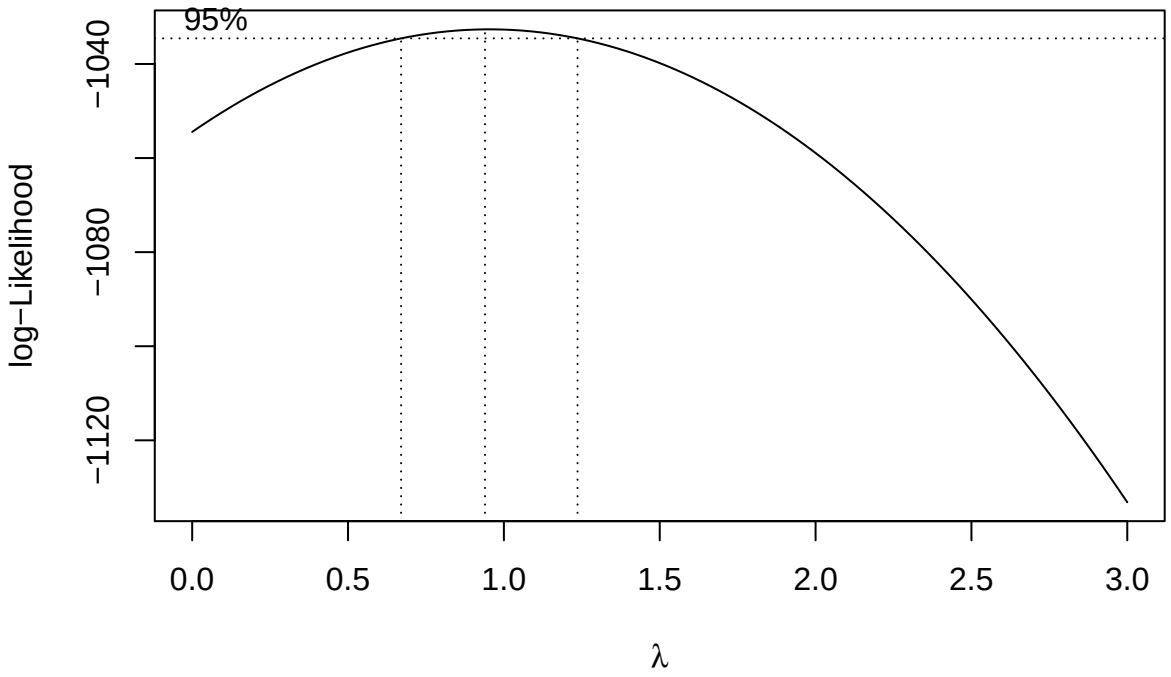
```
##
## Call:
## lm(formula = (charges)^(-0.2) ~ log(age) + children + region +
##     sex, data = non_smokers)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.085480 -0.000036  0.002798  0.005271  0.023402
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.3701217  0.0040794  90.730  < 2e-16 ***
## log(age)         -0.0547231  0.0011007 -49.717  < 2e-16 ***
## children         -0.0041158  0.0003533 -11.650  < 2e-16 ***
## regionnorthwest   0.0027735  0.0012218   2.270   0.0234 *
## regionsoutheast   0.0062215  0.0012152   5.120 3.63e-07 ***
```

```
## regionsouthwest  0.0066586  0.0012221   5.449 6.32e-08 ***
## sexmale          0.0039021  0.0008578   4.549 6.01e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.01398 on 1057 degrees of freedom
## Multiple R-squared:  0.7255, Adjusted R-squared:  0.724
## F-statistic: 465.7 on 6 and 1057 DF,  p-value: < 2.2e-16
```

```
yhat_full_non_smoker <- mlr_full_non_smoker_log_age$fitted.values
res_full_non_smoker <- mlr_full_non_smoker_log_age$residuals
non_smokers %>%
  ggplot(aes(yhat_full_non_smoker, res_full_non_smoker)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```




```
boxcox(mlr_full_non_smoker_log_age, c(0,3,0.1))
```

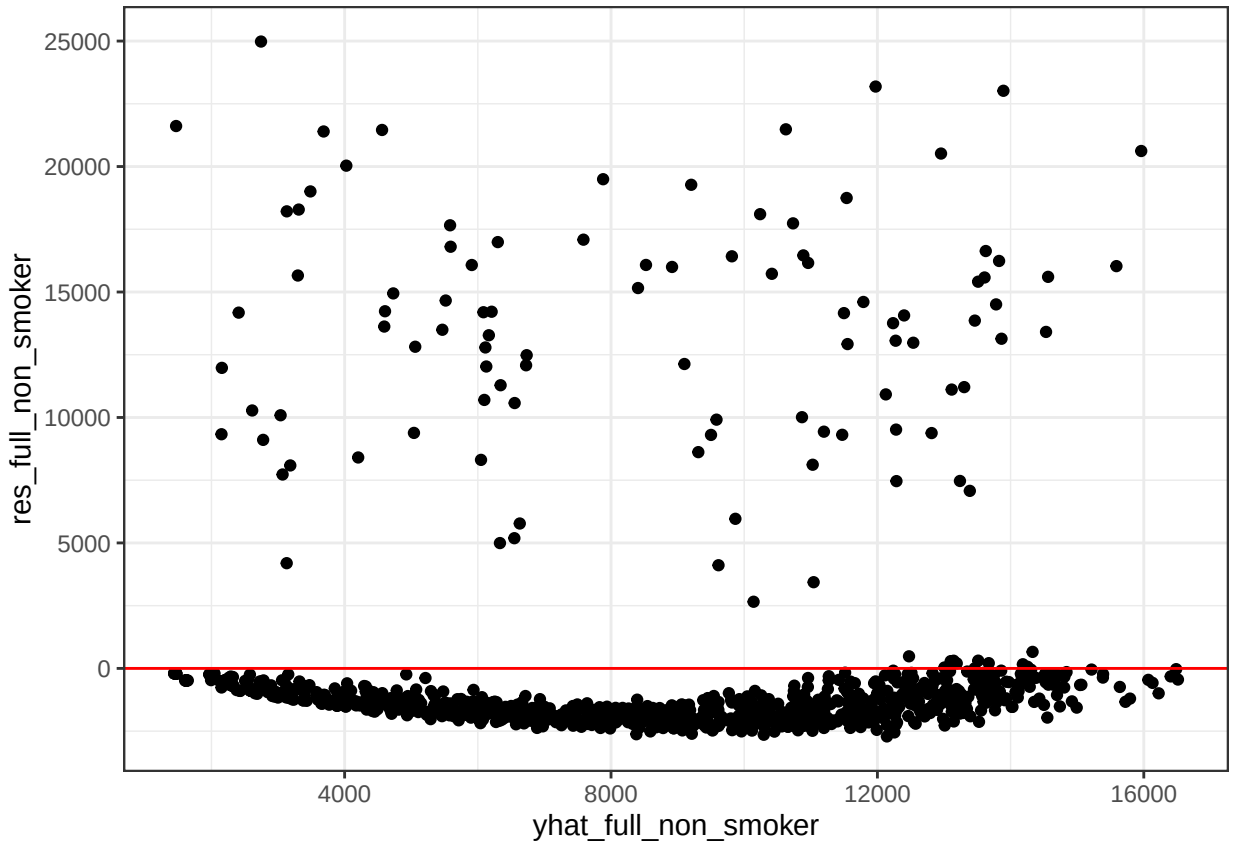


Interaction age and region?

```
mlr_full_non_smoker_region_inter = lm(formula = charges ~ (age)*bmi + region + children + sex, data = no
summary(mlr_full_non_smoker_region_inter)
```

```
##
## Call:
## lm(formula = charges ~ (age) * bmi + region + children + sex,
##     data = non_smokers)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2709.7 -1866.9 -1359.3  -677.6 24979.9
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -1201.7270  2129.9278  -0.564 0.572731
## age               238.4680    51.6271   4.619 4.33e-06 ***
## bmi               -13.8407    68.4291  -0.202 0.839749
## regionnorthwest  -544.8748   401.4481  -1.357 0.174985
## regionsoutheast  -985.7698   412.2071  -2.391 0.016957 *
## regionsouthwest -1400.7248   402.5979  -3.479 0.000523 ***
## children          586.8693   115.7543   5.070 4.70e-07 ***
## sexmale          -531.6770   282.0645  -1.885 0.059711 .
## age:bmi             0.8441     1.6427   0.514 0.607489
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4592 on 1055 degrees of freedom
## Multiple R-squared:  0.4175, Adjusted R-squared:  0.4131
## F-statistic: 94.51 on 8 and 1055 DF,  p-value: < 2.2e-16
```

```
yhat_full_non_smoker <- mlr_full_non_smoker_region_inter$fitted.values
res_full_non_smoker <- mlr_full_non_smoker_region_inter$residuals
non_smokers %>%
  ggplot(aes(yhat_full_non_smoker, res_full_non_smoker)) +
  geom_point() +
  theme_bw() +
  geom_hline(yintercept = 0, color="red")
```



Logistic

```
set.seed(6021) ##for reproducibility
sample<-sample.int(nrow(data), floor(.70*nrow(data)), replace = F)
train<- data[sample, ] ##training data frame
test<-data[-sample, ] ##test data frame
result<-glm(significant.charge ~ age + bmi + children + smoker + region + sex, family="binomial", data=
summary(result)
```

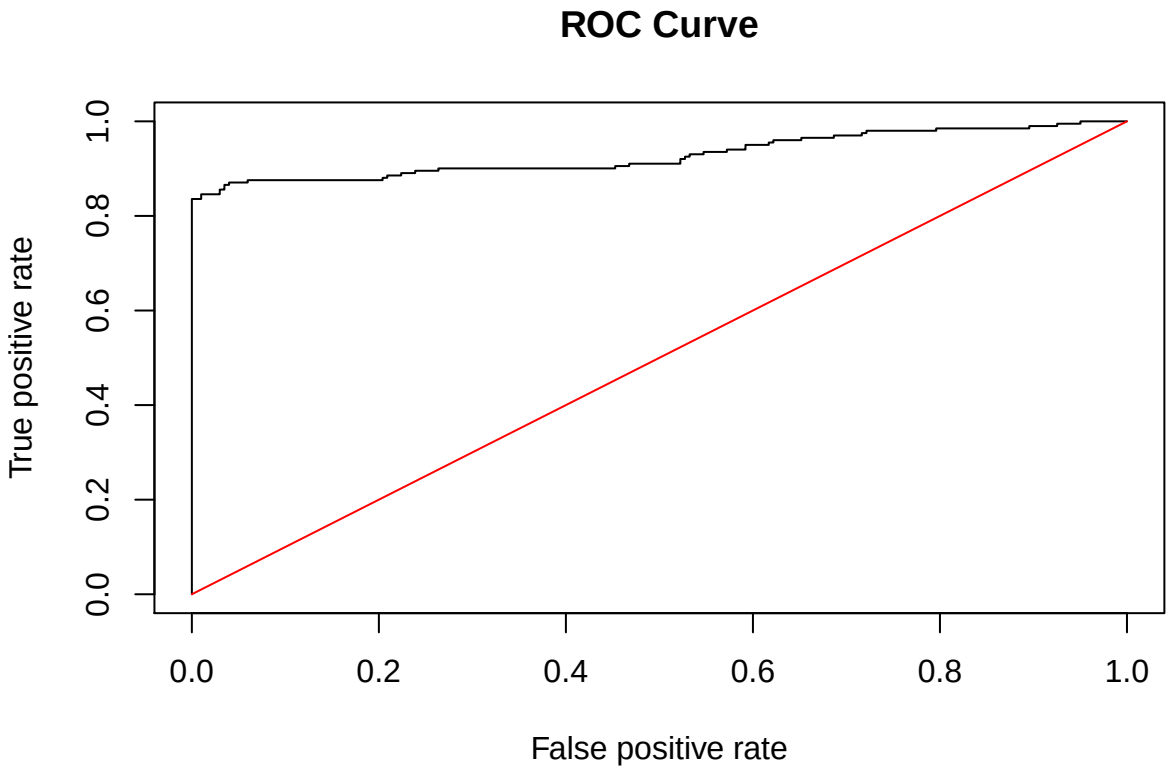
```
##
## Call:
## glm(formula = significant.charge ~ age + bmi + children + smoker +
##     region + sex, family = "binomial", data = train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5205  -0.3475  -0.0284   0.3683   3.4962
##
```

```
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)      -9.26037    0.89751 -10.318   <2e-16 ***
## age               0.18490    0.01342  13.781   <2e-16 ***
## bmi               0.03984    0.01999   1.993   0.0463 *
## children          0.18946    0.09252   2.048   0.0406 *
## smokeryes        22.76270  597.59493   0.038   0.9696
## regionnorthwest  -0.47986    0.32161  -1.492   0.1357
## regionsoutheast  -0.85665    0.33588  -2.551   0.0108 *
## regionsouthwest  -0.50317    0.32335  -1.556   0.1197
## sexmale          -0.56716    0.22860  -2.481   0.0131 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1297.57  on 935  degrees of freedom
## Residual deviance:  500.78  on 927  degrees of freedom
## AIC: 518.78
##
## Number of Fisher Scoring iterations: 18
```

```
library(ROCR)
```

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

```
##predicted survival rate for test data based on training data
preds<-predict(result,newdata=test, type="response")
##transform the input data into a format that is suited for the
##performance() function
rates<-prediction(preds, test$significant.charge)
##store the true positive and false positive rates
roc_result<-performance(rates,measure="tpr", x.measure="fpr")
##plot ROC curve and overlay the diagonal line for random guessing
plot(roc_result, main="ROC Curve")
lines(x = c(0,1), y = c(0,1), col="red")
```




```
##compute the AUC
auc<-performance(rates, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.9288631
```

Matrix

```
table(test$significant.charge, preds>0.5)
```

```
##
##         FALSE TRUE
##   FALSE   182   19
##   TRUE     25  176
```

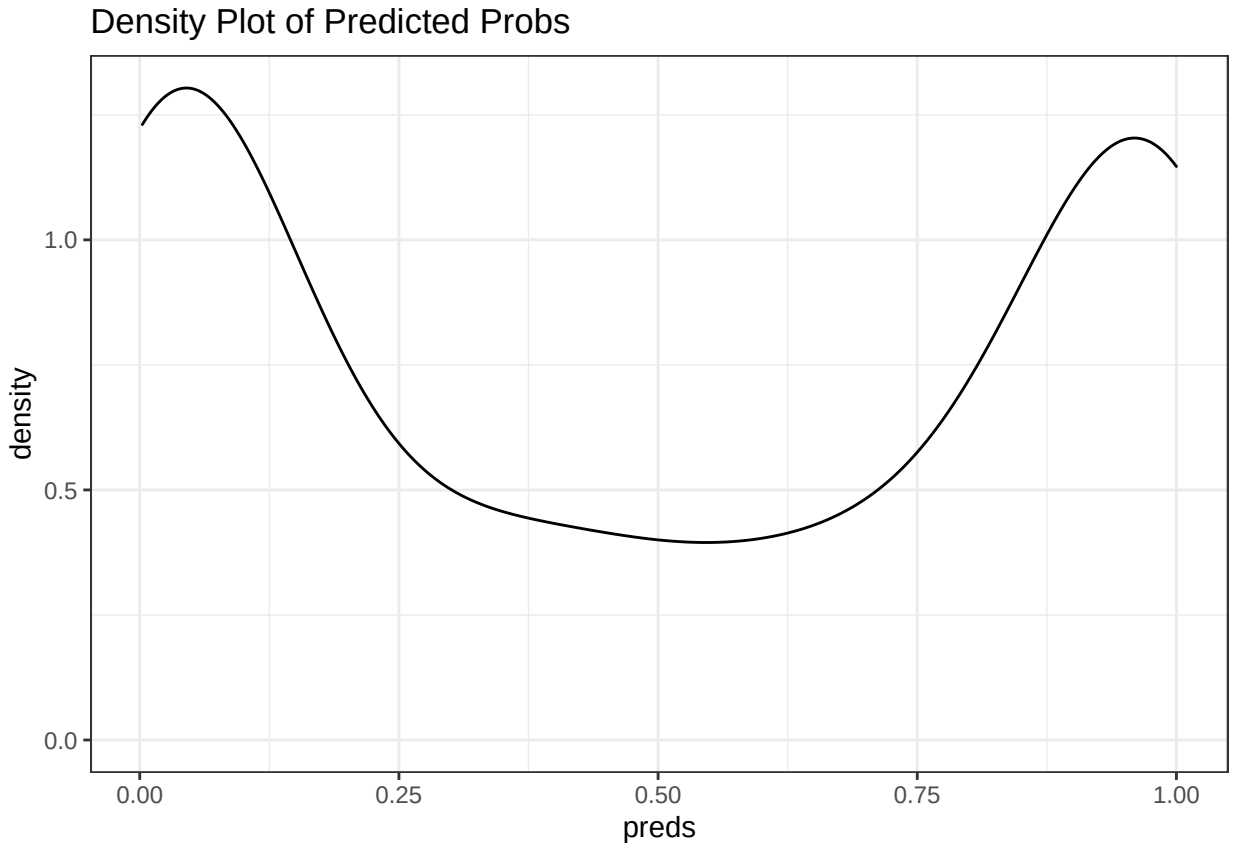
Threshold value manipulation

```
table(test$significant.charge, preds>0.25)
```

```
##
##         FALSE TRUE
##   FALSE   143   58
##   TRUE     20  181
```

Doesn't play a huge role in decreasing the False Positive Rate. We want to make sure that when someone signs up for a plan that they don't get charged significantly given their condition.

```
test<-data.frame(test,preds)
ggplot(test,aes(x=preds))+
geom_density()+
labs(title="Density Plot of Predicted Probs") + theme_bw()
```


Density Plot of Predicted Probs
density
1.0
0.5
0.0
0.00
0.25
0.50
0.75
1.00
preds